Ancestral Instrument Method for Causal Inference without Complete Knowledge*

Debo Cheng[1] , **Jiuyong Li**[1] , **Lin Liu**[1] , **Jiji Zhang**[2] , **Thuc duy Le**[1] and **Jixue Liu**[1]
[1] STEM, University of South Australia, Adelaide, SA, Australia
[2] Department of Religion and Philosophy, Hong Kong Baptist University, Hong Kong, China
{debo.cheng,jiuyong.li,liu.liu,thuc.le,jixue.liu}@unisa.edu.au, zhangjiji@hkbu.edu.hk

Abstract

Unobserved confounding is the main obstacle to causal effect estimation from observational data. Instrumental variables (IVs) are widely used for causal effect estimation when there exist latent confounders. With the standard IV method, when a given IV is valid, unbiased estimation can be obtained, but the validity requirement on a standard IV is strict and untestable. Conditional IVs have been proposed to relax the requirement of standard IVs by conditioning on a set of observed variables (known as a conditioning set for a conditional IV). However, the criterion for finding a conditioning set for a conditional IV needs a directed acyclic graph (DAG) representing the causal relationships of both observed and unobserved variables. This makes it challenging to discover a conditioning set directly from data. In this paper, by leveraging maximal ancestral graphs (MAGs) for causal inference with latent variables, we study the graphical properties of ancestral IVs, a type of conditional IVs using MAGs, and develop the theory to support data-driven discovery of the conditioning set for a given ancestral IV in data under the pretreatment variable assumption. Based on the theory, we develop an algorithm for unbiased causal effect estimation with a given ancestral IV and observational data. Extensive experiments on synthetic and real-world datasets demonstrate the performance of the algorithm in comparison with existing IV methods.

1 Introduction

Inferring the total causal effect of a treatment (a.k.a. exposure, intervention or action) on an outcome of interest is a central problem in scientific discovery and it is essential for decision making in many areas such as epidemiology [Martinussen and others, 2019] and economics [Card, 1993; Verbeek, 2008; Imbens and Rubin, 2015]. With observational data, a major hurdle to causal effect estimation is the bias caused by confounders. Therefore the unconfoundedness assumption is commonly made by causal inference methods [Imbens and Rubin, 2015].

When there are latent or unobserved confounders, the unconfoundedness assumption becomes unreliable. In this case, the instrumental variable (IV) approach [Card, 1993; Martens *et al.*, 2006] is considered a powerful way to achieve unbiased causal effect estimation. The IV approach leverages an IV (denoted as S), a variable known to be a cause of the treatment W, controlling treatment assignment, to deal with unobserved confounding. Given a valid IV, an unbiased estimate of the total causal effect of W on outcome Y can be obtained based on the estimated causal effect of S on W and the estimated causal effect of S on Y.

The requirements for a standard IV are very strong and it is impossible to find a standard IV in many applications. For a variable S to be a valid standard IV, it must be a cause of W and satisfy the *exclusion restriction* (i.e. the causal effect of S on Y must be only through W) and be *exogenous* (i.e. S does not share common causes with Y) [Martens *et al.*, 2006; Imbens, 2014]. These conditions are strict and can only be justified by domain knowledge. In particular, the exogeneity implies that S must be a factor "external" to the system under consideration and connects to the system only through the treatment W, which is impossible to validate in practice.

A conditional IV relaxes the requirements of a standard IV significantly and it is more likely to exist in an application than a standard IV [Pearl, 2009; Brito and Pearl, 2002]. With the concept of a conditional IV, an "internal" variable S can be a valid IV when conditioning on a set of observed variables $\mathbf{Z}$. In this case, S is known as a conditional IV which is instrumentalized by $\mathbf{Z}$, and the key to the success of the conditional IV method (in obtaining unbiased causal effect estimation) is to find a proper conditioning set $\mathbf{Z}$ for a given conditional IV.

However, the criterion for finding $\mathbf{Z}$ is based on complete causal structure knowledge (i.e. a complete causal DAG with observed and unobserved variables), which, if at all possible, can only be obtained from domain knowledge, not observational data. Moreover, recent work [Vander Zander *et al.*, 2015] has shown that the search for $\mathbf{Z}$ in a DAG is NP-hard for a given conditional IV. The authors also proposed the concept of an ancestral IV in a DAG, a restricted version of a conditional IV, to work towards efficient search for $\mathbf{Z}$. Nonetheless, the search for $\mathbf{Z}$ for an ancestral IV in a DAG

*Appendices of the paper are available at https://arxiv.org/abs/2201.03810.

still requires a DAG containing all the observed and unobserved variables. Therefore, the majority of existing methods for finding a conditioning set of a conditional IV need a causal graph which may not be known in many applications.

There are some works which *use* the conditional IV without complete causal knowledge, such as random forest for IV [Athey *et al.*, 2019], an estimator based on the assumption of the existence of some invalid and some valid IVs (sisVIVE) [Kang *et al.*, 2016], and IV.tetrad [Silva and Shimizu, 2017], but they do not identify conditioning sets. We differentiate our work from these works in the Related Work section in more detail and compare our method with them in the Experiments section.

In this paper, we design an algorithm for identifying a conditional set that instrumentalizes a given ancestral IV, a type of conditional IVs, in data directly. In order to achieve this, we study the graphical properties of an ancestral IV using a MAG (maximal ancestral graph [Richardson and Spirtes, 2002; Zhang, 2008a]) and develop the theory for data-driven discovery of a conditioning set for a given ancestral IV. To the best of our knowledge, there is no existing method for finding a conditioning set of a conditional IV directly from data.

The contributions of this work are summarized as follows.

- We study the novel graphical properties of an ancestral IV using MAGs, which enables a data-driven approach to applying the IV method to obtain unbiased causal effect estimation when there are latent confounders.
- We establish graphical criteria for determining a conditioning set of a given ancestral IV via a MAG (PAG).
- Based on the theorems, we propose an effective algorithm for unbiased causal effect estimation from data with latent variables. The experiments on synthetic and real-world datasets demonstrate the performance of the proposed algorithm.

2 Background

2.1 Graphical Notation and Definitions

A graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ consists of a set of nodes $\mathbf{V} = \{V_1, \ldots, V_p\}$, denoting random variables, and a set of edges $\mathbf{E} \subseteq \mathbf{V} \times \mathbf{V}$, representing the relationships between nodes. Two nodes linked by an edge are *adjacent*. In the paper, an edge in $\mathbf{E}$ can be a directed edge $\rightarrow$, a bi-directed edge $\leftrightarrow$, or a partially directed edge $\circ\!\!\rightarrow$, where the circle at the left end of the edge indicates uncertainty of the orientation.

A path π between V_i and V_j in a graph comprises a sequence of distinct nodes $\langle V_i, \ldots, V_j \rangle$ with every pair of successive nodes being adjacent. V_i and V_j are end nodes of π, and other nodes on π are non-end nodes. A path is a directed or causal path if all edges along it are directed such as $V_i \rightarrow \ldots \rightarrow V_j$. We use '$*$' to indicate an arbitrary edge mark of an edge, i.e. arrow ($>$), tail ($-$) or circle ($\circ$). V_i is a collider on a path if $V_{i-1} *\!\!\rightarrow V_i \leftarrow\!\!* V_{i+1}$ is in $\mathcal{G}$. A *collider path* is a path on which every non-endpoint node is a collider. A path of length one is a *trivial collider path*.

If there is $V_i \leftrightarrow V_j$ in a graph, V_i and V_j are called spouses to each other. We use $Adj(V)$, $Pa(V)$, $Ch(V)$, $An(V_i)$, $De(V_i)$, $Sp(V)$ and $PossAn(V)$ to denote the sets of all adjacent nodes, parents, children, ancestors, descendants, spouses and possible ancestors of V, respectively, in the same way as in [Perković *et al.*, 2018]. The definitions of a node's parents, children, ancestors and descendants are provided in Appendix A [Cheng *et al.*, 2022]. A directed cycle occurs when the first and last nodes on a path are the same node. A *DAG* contains directed edges without cycles. In a DAG with observed and unobserved variables, if there exists $V_i \leftarrow U \rightarrow V_j$ where U is a latent variable, V_i and V_j are often called spouses to each other.

Ancestral graphs are often used to represent the mechanisms of the data generation process that may involve latent variables [Zhang, 2008a]. An ancestral graph is a graph that does not contain directed cycles or almost directed cycles [Richardson and Spirtes, 2002]. An almost directed cycle occurs if $V_i \leftrightarrow V_j$ and $V_j \in An(V_i)$.

To save space, the definitions of Markov property, faithfulness, d-separation (denoted as $\perp\!\!\!\perp_d$), d-connecting (denoted as $\not\!\perp\!\!\!\perp_d$), m-separation (denoted as $\perp\!\!\!\perp_m$), m-connecting (denoted as $\not\!\perp\!\!\!\perp_m$), and the graphical criteria of d-separation and m-separation are introduced in Appendix A.

Definition 1 (MAG)**.** *An ancestral graph $\mathcal{M} = (\mathbf{V}, \mathbf{E})$ is a MAG when every pair of non-adjacent nodes V_i and V_j in $\mathcal{M}$ are m-separated by a set $\mathbf{Z} \subseteq \mathbf{V} \backslash \{V_i, V_j\}$.*

A DAG obviously meets the conditions of a MAG, so syntactically, a DAG is also a MAG without bi-directed edges. It is worth noting that a causal DAG over a set of observed and unobserved variables can be converted to a MAG over the observed variables uniquely according to the construction rules [Zhang, 2008b]. A set of Markov equivalent MAGs can be represented uniquely by a *partial ancestral graph* (PAG) that is defined in Appendix A.

Definition 2 (Visibility [Zhang, 2008a])**.** *Given a MAG $\mathcal{M} = (\mathbf{V}, \mathbf{E})$, a directed edge $V_i \rightarrow V_j$ is visible if there is a node $V_k \notin Adj(V_j)$, such that either there is an edge between V_k and V_i that is into V_i, or there is a collider path between V_k and V_i that is into V_i and every node on this path is a parent of V_j. Otherwise, $V_i \rightarrow V_j$ is said to be invisible.*

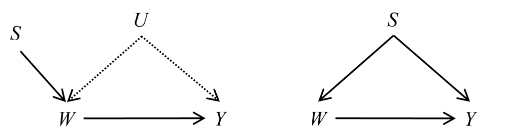
In a given DAG $\mathcal{G}$, if V_i and V_j are not adjacent and $V_i \notin An(V_j)$, then $Pa(V_i)$ blocks all paths between V_i to V_j. In a given MAG $\mathcal{M}$, there is a similar conclusion, but the blocked set is $D\text{-}SEP(V_i, V_j)$ as defined below, instead of $Pa(V_i)$.

Definition 3 ($D\text{-}SEP(V_i, V_j)$ in a MAG $\mathcal{M}$ [Spirtes *et al.*, 2000])**.** *In a MAG $\mathcal{M} = (\mathbf{V}, \mathbf{E})$, assume that V_i and V_j are not adjacent. A node $V_k \in D\text{-}SEP(V_i, V_j)$ if $V_k \neq V_i$, and there is a collider path between V_k to V_i such that every node on this path (including V_k) is in $An(V_i)$ or $An(V_j)$ in $\mathcal{M}$.*

2.2 Instrumental Variables

In this section, we introduce the concepts of standard IVs, conditional IVs and ancestral IVs in a DAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ with $\mathbf{V} = \mathbf{X} \cup \mathbf{U} \cup \{W, Y\}$, where $\mathbf{X}$ is the set of all observed variables and $\mathbf{U}$ is the set of unobserved variables.

Definition 4 (Standard IV)**.** *A variable S is said to be an IV w.r.t. $W \rightarrow Y$, if (i) S is a cause of W, (ii) S affects Y only through W (i.e. exclusion restriction), and (iii) S does not share common causes with Y (i.e. S is exogenous).*



(a) A DAG to represent an IV S (b) The corresponding MAG

Figure 1: An example of a standard IV represented in two types of causal graphs: (a) DAG; (b) MAG where $W \to Y$ is invisible.

The variable S in the DAG in Fig. 1 (a) depicts a standard IV w.r.t. $W \to Y$. Given a standard IV S, the causal effect of W on Y, denoted as β_{wy} can be calculated as σ_{sy}/σ_{sw}, where σ_{sy} and σ_{sw} are the estimated causal effect of S on Y and the causal effect of S on W, respectively.

A conditional IV in a DAG (Definition 7.4.1 on Page 248 [Pearl, 2009]) is defined as follows.

Definition 5 (Conditional IV)**.** *Given a DAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ with $\mathbf{V} = \mathbf{X} \cup \mathbf{U} \cup \{W, Y\}$, a variable S is said to be a conditional IV w.r.t. $W \to Y$ if there exists a set of observed variables $\mathbf{Z} \subseteq \mathbf{X}$ such that (i) $S \not\perp\!\!\!\perp_d W \mid \mathbf{Z}$, (ii) $S \perp\!\!\!\perp_d Y \mid \mathbf{Z}$ in $\mathcal{G}_{\underline{W}}$, and (iii) $\forall Z \in \mathbf{Z}$, $Z \notin De(Y)$.*

In the above definition, $\mathcal{G}_{\underline{W}}$ is the DAG obtained by removing $W \to Y$ from $\mathcal{G}$. It is worth noting that $\mathbf{Z}$ is a set of observed variables and $\mathbf{Z} \neq \emptyset$ for a conditional IV S.

Detention 5 allows a conditional IV S such that S is not related to W, but conditioning on $\mathbf{Z}$, is related to W when $\mathbf{Z}$ contains a descendant of S. This might lead to a misleading result [Vander Zander *et al.*, 2015]. The following defined notion mitigates this issue.

Definition 6 (Ancestral IV in DAG [Vander Zander *et al.*, 2015])**.** *Given a DAG $\mathcal{G} = (\mathbf{X} \cup \mathbf{U} \cup \{W, Y\}, \mathbf{E})$, a variable $S \in \mathbf{X}$ is said to be an ancestral IV w.r.t. $W \to Y$, if there exists a set of observed variables $\mathbf{Z} \subseteq \mathbf{X} \setminus \{S\}$ such that (i) $S \not\perp\!\!\!\perp_d W \mid \mathbf{Z}$, (ii) $S \perp\!\!\!\perp_d Y \mid \mathbf{Z}$ in $\mathcal{G}_{\underline{W}}$, and (iii) $\mathbf{Z} \subseteq \{An(Y) \cup An(S)\}$ and $\forall Z \in \mathbf{Z}$, $Z \notin De(\overline{Y})$.*

In a given DAG $\mathcal{G}$, an ancestral IV is a conditional IV, but a conditional IV may not be an ancestral IV. However, the application of a standard IV, conditional IV or ancestral IV requires that a causal DAG $\mathcal{G} = (\mathbf{X} \cup \mathbf{U} \cup \{W, Y\}, \mathbf{E})$ must be completely known. Often, it is impractical to get such complete causal knowledge in real-world applications.

Using the IV approach, one way to estimate β_{wy} from data is to employ the generalized linear model. In this work, we consider the potential outcome model [Imbens and Rubin, 2015] to calculate β_{wy} as introduced in the following.

$$\eta\{\mathbf{E}(y \mid w, s, \mathbf{z})\} - \eta\{\mathbf{E}(y_0 \mid w, s, \mathbf{z})\} = f^T(\mathbf{z}) w \beta_{wy} \quad (1)$$

where y, w, s and $\mathbf{z}$ denote the values of Y, W, S and $\mathbf{Z}$ respectively for a given individual, y_0 is the potential outcome with w set to 0, and η is the identity, log or logit link. The function $f^T(\mathbf{z})$ allows us to measure the interactions between W and $\mathbf{Z}$. As commonly done in literature, we utilize a two-stage estimation to estimate β_{wy}. The estimator requires two regression models. The first stage is to build a regression model $\hat{w} = \hat{\mathbf{E}}(w \mid s, \mathbf{z})$ for each individual from data. The second stage is to fit the outcome by using $\mathbf{Z}$ and $f^T(\mathbf{z})\hat{w}$ as regressors. Hence, the estimated coefficient of $f^T(\mathbf{z})\hat{w}$ is $\hat{\beta}_{wy}$. For more details on the estimator, please refer to the literature [Sjolander and Martinussen, 2019].

3 Finding a Conditioning Set for an Ancestral IV in Data

3.1 Problem Setting

In this work, we assume that an ancestral IV S has been given, and there exists a conditioning set $\mathbf{Z} \subseteq \mathbf{X} \setminus \{S\}$ and $\mathbf{Z} \neq \emptyset$ for S in the underlying DAG $\mathcal{G}$ over $\mathbf{X} \cup \mathbf{U} \cup \{W, Y\}$. We assume that $\mathbf{X}$ contains only pretreatment variables as often assumed in literature [Imbens and Rubin, 2015; Silva and Shimizu, 2017], i.e. for each $X \in \mathbf{X}$, $X \notin De(W)$ and $X \notin De(Y)$ in $\mathcal{G}$. The goal of the work is to provide a practical solution for finding a set of observed variables $\mathbf{Z}$ for a given ancestral IV without knowing the complete causal knowledge. Clause (iii) in Definition 6 is too restrictive for finding $\mathbf{Z}$ in data directly because in a PAG, an ancestor and a spouse of a node may not be distinguished. Hence, we consider a relaxed condition for clause (iii) of Definition 6, i.e. $\mathbf{Z}$ does not contain a collider on a d-connecting path between S and W since this is sufficient to address the original problem with the notion of a conditional IV. Hereinafter, we consider that in a DAG $\mathcal{G}$, if a set $\mathbf{Z} \subseteq \mathbf{X} \setminus \{S\}$ satisfies clauses (i) and (ii) in Definition 6, and does not contain a collider between S and W, then $\mathbf{Z}$ instrumentalizes S.

Furthermore, we consider the case that an ancestral IV S in a DAG is a cause or spouse of W (i.e. a node in $\{Pa(W) \cup Sp(W)\}$) because it is easy to know a cause of W or a spouse of W that is not a direct cause or spouse of Y. In the graphic term, S is adjacent to W but not to Y. For example, when estimating causal effect of *Smoking* on *Lung Cancer*, *Income* is a direct cause of $Smoking$, but not a direct cause of *Lung Cancer* [Spirtes *et al.*, 2000]. Hence, *Income* can be used as an IV. It is feasible for users to find an S similar to the case described above. When we infer causal effect from data, we follow the convention in causal inference, that is, the causal DAG $\mathcal{G}$ satisfies Markov property, the causal DAG $\mathcal{G}$ and the data are faithful to each other [Spirtes *et al.*, 2000; Pearl, 2009]. All proofs in this section are provided in Appendix B.

3.2 Representing an Ancestral IV in MAG

An advantage of MAGs is their ability in representing causal relationships between observed variables without involving latent variables that exist in the system [Spirtes *et al.*, 2000]. A PAG that represents the Markov equivalence class of MAGs can be learned from data with latent variables. The goal of our work is to study the graphical properties of an ancestral IV in a mapped MAG (or equivalently in a PAG) and establish the corresponding theorems for supporting a practical algorithm to estimate β_{wy} from data using ancestral IVs.

When we use a MAG $\mathcal{M}$ over $\mathbf{X} \cup \{W, Y\}$ to represent the data generation mechanism involving latent variables $\mathbf{U}$, an IV in the underlying DAG over $\mathbf{X} \cup \mathbf{U} \cup \{W, Y\}$ can be mapped to $\mathcal{M}$. As all types of IVs (standard, conditional or ancestral IVs) have spurious associations with Y because of the latent confounder between W and Y, we develop a lemma for properly mapping an IV in a DAG to a MAG.

Lemma 1. *Given a DAG $\mathcal{G} = (\mathbf{X} \cup \mathbf{U} \cup \{W, Y\}, \mathbf{E}')$ with the edges $W \to Y$ and $W \leftarrow U \to Y$ in $\mathbf{E}'$, and $U \in \mathbf{U}$. Let $\mathcal{M} = (\mathbf{X} \cup \{W, Y\}, \mathbf{E})$ be the MAG mapped from $\mathcal{G}$ based on the construction rules [Zhang, 2008b]. Suppose that there exists an ancestral IV S conditioning on a set $\mathbf{Z} \subseteq \mathbf{X} \setminus \{S\}$ in $\mathcal{G}$. In the mapped MAG $\mathcal{M}$, the edge $W \to Y$ is invisible and there is an edge $S \to Y$ or $S \leftrightarrow Y$.*

We take the standard IV S in the DAG of Fig. 1 (a) as an example to explain the lemma. In this, S is a standard IV w.r.t. $W \to Y$ and $S \in An(Y)$, so the IV S in the mapped MAG $\mathcal{M}$ has a directed edge $S \to Y$ as shown in Fig. 1 (b) and the edge $W \to Y$ is invisible.

3.3 The Property of an Ancestral IV in MAG

First of all, we introduce the manipulated MAG $\mathcal{M}_{\underline{W}\widetilde{S}}$ that is obtained by replacing $W \to Y$ with $W \leftrightarrow Y$ in $\mathcal{M}$ and removing the edge between S and Y. We have the following lemma to present the property of an ancestral IV S in the MAG $\mathcal{M}$ mapped from a DAG $\mathcal{G}$.

Lemma 2 (The property of an ancestral IV in the mapped MAG)**.** *Given a DAG $\mathcal{G} = (\mathbf{X} \cup \mathbf{U} \cup \{W, Y\}, \mathbf{E}')$ with the edges $W \to Y$ and $W \leftarrow U \to Y$ in $\mathbf{E}'$, and $U \in \mathbf{U}$, and let $\mathcal{M} = (\mathbf{X} \cup \{W, Y\}, \mathbf{E})$ be the MAG mapped from $\mathcal{G}$. Suppose that there exists an ancestral IV S conditioning on a set $\mathbf{Z} \subseteq \mathbf{X} \setminus \{S\}$ in $\mathcal{G}$. In the mapped MAG $\mathcal{M}$, if a set $\mathbf{Z} \subseteq \mathbf{X} \setminus \{S\}$ satisfies the conditions that (i) S and W are not m-separated given $\mathbf{Z}$ in $\mathcal{M}$, and (ii) S and Y are m-separated by $\mathbf{Z}$ in $\mathcal{M}_{\underline{W}\widetilde{S}}$, then $\mathbf{Z}$ instrumentalizes S in the DAG $\mathcal{G}$.*

3.4 Determining a Conditioning Set Using a MAG

The following lemma from the work in [Maathuis *et al.*, 2015] is useful for our purpose.

Lemma 3. *Let X and Y be two non-adjacent nodes in a MAG $\mathcal{M}$, then $X \perp\!\!\!\perp_m Y \mid D\text{-}SEP(X, Y)$.*

Therefore, we have the following corollary for finding a set $\mathbf{Z} \subseteq \mathbf{X} \setminus \{S\}$ in a MAG $\mathcal{M}$ that instrumentalizes S in the underlying DAG $\mathcal{G}$.

Corollary 1. *Given a DAG $\mathcal{G} = (\mathbf{X} \cup \mathbf{U} \cup \{W, Y\}, \mathbf{E}')$ with the edges $W \to Y$ and $W \leftarrow U \to Y$ in $\mathbf{E}'$, and $U \in \mathbf{U}$, and let $\mathcal{M} = (\mathbf{X} \cup \{W, Y\}, \mathbf{E})$ be the MAG mapped from $\mathcal{G}$. For a given ancestral IV S, $D\text{-}SEP(S, Y)$ in the mapped MAG $\mathcal{M}$ is a set that instrumentalizes S in the DAG $\mathcal{G}$.*

Corollary 1 provides a theoretical solution for determining a conditioning set that instrumentalizes a given ancestral IV S in the underlying DAG. Taking a data-driven approach, we can learn a PAG from data with latent variables, but for each of the Markov equivalent MAGs represented by the PAG, there is a corresponding $D\text{-}SEP(S, Y)$ for S. We do not know which MAG is the ground-truth MAG that is mapped from the underlying DAG, and hence we do not know which $D\text{-}SEP(S, Y)$ is the true conditioning set for S. To provide a precise causal effect estimation, in the next section, we propose a theorem to determine a conditioning set $\mathbf{Z}$ from a PAG, in which non-ancestral nodes of S or Y may be contained in $\mathbf{Z}$, but do not result in bias.

Algorithm 1 <u>A</u>ncestral <u>IV</u> estimator <u>i</u>n <u>P</u>AG (*AIViP*)

Input: Dataset $\mathbf{D}$ with the treatment W, the outcome Y, the set of pretreatment variables $\mathbf{X}$ and ancestral IV S

Output: $\hat{\beta}_{wy}$

1: Call the causal structure learning method, rFCI, to learn a PAG $\mathcal{P}$ from $\mathbf{D}$
2: Obtain the manipulated PAG $\mathcal{P}_{\underline{W}\widetilde{S}}$
3: Obtain the set $PossAn(S \cup Y) \setminus \{W, S\}$ in $\mathcal{P}_{\underline{W}\widetilde{S}}$
4: $\mathbf{Z} = PossAn(S \cup Y) \setminus \{W, S\}$
5: fit $\hat{w} = \hat{\mathbf{E}}(w \mid s, \mathbf{z})$
6: fit $\hat{y} = \hat{\mathbf{E}}(y \mid f^T(\mathbf{z})\hat{w}, \mathbf{z})$
7: Let $\hat{\beta}_{wy}$ be the coefficient of $f^T(\mathbf{z})\hat{w}$
8: **return** $\hat{\beta}_{wy}$

3.5 Determining a Conditioning Set Z Using a PAG

For a given ancestral IV S, we have the following theorem for determining a set $\mathbf{Z}$ in a PAG $\mathcal{P}$ that instrumentalizes S.

Theorem 1. *Given a DAG $\mathcal{G} = (\mathbf{X} \cup \mathbf{U} \cup \{W, Y\}, \mathbf{E}')$ with the edges $W \to Y$ and $W \leftarrow U \to Y$ in $\mathbf{E}'$, and $U \in \mathbf{U}$, and let $\mathcal{M} = (\mathbf{X} \cup \{W, Y\}, \mathbf{E})$ be the MAG mapped from $\mathcal{G}$. From data, the mapped MAG $\mathcal{M}$ is represented by a PAG $\mathcal{P} = (\mathbf{X} \cup \{W, Y\}, \mathbf{E}'')$. For a given ancestral IV S which is a cause or spouse of W, the set $PossAn(S \cup Y) \setminus \{W, S\}$ in the learned $\mathcal{P}$ is a set that instumentalizes S in the DAG $\mathcal{G}$.*

Note that $PossAn(S \cup Y) \setminus \{W, S\}$ is a superset of $D\text{-}SEP(S, Y)$ since the former may contain non-ancestral nodes of S or Y, and do not result in bias. Theorem 1 allows us to discover the conditioning set as $PossAn(S \cup Y) \setminus \{W, S\}$ from the manipulated PAG $\mathcal{P}_{\underline{W}\widetilde{S}}$ for the given ancestral IV S without complete causal knowledge.

3.6 The Proposed Ancestral IV Based Estimator

In this section, we develop a data-driven estimator, <u>A</u>ncestral <u>IV</u> estimator <u>i</u>n <u>P</u>AG (*AIViP*) as shown in Algorithm 1, for unbiased causal effect estimation with a given ancestral IV and data with latent confounders.

As presented in Algorithm 1, *AIViP* (Line 1) employs a global causal structure learning method to discover a PAG from data with latent variables. In this work, we employ rFCI [Colombo *et al.*, 2012] and use the function *rfci* in the $\mathbf{R}$ package *pcalg* [Kalisch *et al.*, 2012] to implement rFCI. Lines 2 and 3 construct the manipulated PAG $\mathcal{P}_{\underline{W}\widetilde{S}}$ and get $PossAn(S \cup Y) \setminus \{W, S\}$ from the PAG $\mathcal{P}_{\underline{W}\widetilde{S}}$. Lines 4 to 7 estimate $\hat{\beta}_{wy}$ by the two-stage regression in Eq.(1). We use the function *glm* in the $\mathbf{R}$ package *stats* to fit $\hat{w}$ and the function *ivglm* in the $\mathbf{R}$ package *ivtools* [Sjolander and Martinussen, 2019] to fit $\hat{y}$.

4 Experiments

The goal of the experiments is to evaluate the performance of *AIViP* in obtaining the causal effect estimate $\hat{\beta}_{wy}$, especially, when there is a latent confounder between W and Y. Five benchmark causal effect estimators are used in the comparison experiments, including the IV.tetrad in [Silva and Shimizu, 2017]; some invalid some valid IV estimator (sisVIVE) [Kang *et al.*, 2016]; two-stage least squares for standard IV (TSLS) [Angrist and Imbens, 1995], the most popular estimator; An extension of TSLS for a conditional IV by conditioning on the set of all variables $\mathbf{X} \setminus \{S\}$ (TSLSCIV) [Imbens, 2014]; the causal random forest for IV regression (FIVR), with a given conditional IV [Athey *et al.*, 2019].

It is worth noting that since IV.tetrad is the only other data-driven conditional IV method, we also include the data-driven standard IV method (sisVIVE), the most popular standard IV method (TSLS) and its extension to condition IVs (TSLSCIV), and the popular random forest based estimator with a given conditional IV (FIVR).

Implementation and Parameter Setting. The implementation of IV.tetrad is retrieved from the authors' site[1]. The parameters of *num_ivs* and *num_boot* are set to 3 (1 for VitD) and 500, respectively. We report the average result of 500 bootstrapping as the final estimated $\hat{\beta}_{wy}$ for IV.tetrad. The implementation of TSLSCIV is based on the functions *glm* and *ivglm* in the $\mathbf{R}$ packages *stats* and *ivtools*, respectively. TSLS is implemented by using the function *ivreg* in the $\mathbf{R}$ package *AER* [Greene, 2003]. FIVR is implemented by using the function *instrumental_forest* in the $\mathbf{R}$ package *grf* [Athey *et al.*, 2019]. The implementation of sisVIVE is based on the function *sisVIVE* in the $\mathbf{R}$ package *sisVIVE*. The significance level is set to 0.05 for rFCI used by *AIViP*.

Evaluation Metrics. For the synthetic dataset with the true β_{wy}, we report the estimation bias, $|(\hat{\beta}_{wy} - \beta_{wy})/\beta_{wy}| * 100$ (%). For the real-world datasets, we empirically evaluate the performance of all estimators with the results reported in the corresponding references since the true β_{wy} is not available, and we provide the corresponding 95% confidence interval (C.I.) of $\hat{\beta}_{wy}$ for all estimators.

4.1 Simulation Study

We conduct simulation studies to evaluate the performance of *AIViP* when W and Y share a latent confounder U. We generate two groups of synthetic datasets with a range of sample sizes: 2k (i.e. 2,000), 3k, 4k, 5k, 6k, 8k, 10k, 12k, 15k, 18k, and 20k. The set of observed variables $\mathbf{X}$ is $\{X_1, X_2, \ldots, X_{23}, S\}$. We add two and three latent variables for Group I and Group II datasets respectively. The generated synthetic datasets satisfy the three conditions of ancestral IV in Definition 6. The details of the data generating process are provided in Appendix C. To make the results reliable, each reported result is the average of 20 repeated simulations. The estimation biases of all estimators on both groups of synthetic datasets are reported in Table 1.

Results. From Table 1, we have the following observations: (1) the large estimation biases of TSLS show that the confounding bias caused by the latent confounders between S and Y is not controlled by TSLS at all. (2) TSLSCIV has the largest estimation biases on both groups of synthetic datasets, which shows that conditioning on all variables is inappropriate since the data contains collider bias. (3) The estimation biases of *AIViP* on both groups of datasets show that *AIViP* outperforms FIVR and sisVIVE. This is because both methods fail to detect either colliders or confounding bias in the data. (4) *AIViP* slightly outperforms IV.tetrad in Group I datasets and the two methods have similar performance in Group II datasets. Note that IV.tetrad performs well with synthetic datasets, but not with real-world datasets since its data distribution assumption may not be satisfied in real-world datasets.

Group I						
n	*AIViP*	TSLS	TSLSCIV	FIVR	sisVIVE	IV.tetrad
2k	**15.0**	145.2	342.2	79.2	27.3	32.4
3k	**6.6**	143.9	340.4	94.6	184.3	28.0
4k	**27.5**	143.6	343.0	104.4	53.4	30.6
5k	**20.9**	143.9	347.7	114.5	27.7	24.1
6k	**3.9**	142.2	344.0	117.5	55.8	32.1
8k	**11.8**	144.7	340.6	119.9	15.8	31.9
10k	**21.0**	141.8	342.6	130.7	320.6	30.7
12k	**0.2**	144.2	340.4	132.7	23.0	29.1
15k	29.8	145.2	344.8	141.2	142.4	**25.0**
18k	36.7	144.4	342.4	142.3	312.7	**30.6**
20k	**15.2**	144.6	341.4	144.8	187.7	30.9
Group II						
2k	63.8	284.4	884.7	534.4	199.4	**35.8**
3k	54.6	281.5	840.3	538.9	364.5	**40.2**
4k	47.0	286.1	813.9	529.5	327.8	**33.5**
5k	**18.2**	289.7	838.5	571.4	396.4	35.7
6k	**31.5**	283.1	837.1	581.7	353.2	39.5
8k	**26.7**	285.4	836.7	593.0	584.6	40.1
10k	41.5	280.5	807.6	572.5	653.1	**37.0**
12k	**28.6**	286.3	818.6	588.7	696.0	35.3
15k	40.1	285.4	824.5	604.4	652.8	**35.0**
18k	**2.6**	284.1	829.8	612.0	823.6	38.8
20k	**16.4**	291.0	821.9	608.8	634.8	39.8

Table 1: Summary of the estimation bias (%) on both groups of synthetic datasets. The smallest estimation bias on each group is boldfaced. *AIViP* consistently obtains good performance on all datasets.

4.2 Experiments on Real-World Datasets

In our experiments, we need to choose some datasets for which the empirical estimates are widely acceptable since there are no ground truths for the real-world datasets. Hence, we evaluate the performance of *AIViP* on three real-world datasets, Vitamin D data (VitD) [Martinussen and others, 2019], Schoolingreturn [Card, 1993] and 401(k) data [Verbeek, 2008]. These datasets are widely utilized in the assessment of IV methods. Each of the three datasets has a nominated conditional IV for estimating the causal effects, but there is not enough knowledge to determine the conditioning sets for the nominated conditional IVs. The details of the three datasets are introduced in Appendix C.

VitD contains 2,571 individuals and 5 variables: age, filaggrin (an instrument), vitd (the treatment variable), time (follow-up time), and death (the outcome variable) [Sjolander and Martinussen, 2019]. We take the estimated $\hat{\beta}_{wy} = 2.01$ with 95% C.I. (0.96, 4.26) from the work [Martinussen and others, 2019] as the reference causal effect.

Schoolingreturn contains 3,010 individuals and 19 variables [Card, 1993]. The treatment is the education of employees. The outcome is raw wages in 1976 (in cents per hour). A goal of the study is to investigate the causal effect of education on earnings. Card [Card, 1993] uses geographical

[1]http://www.homepages.ucl.ac.uk/~ucgtrbd/code/iv_discovery

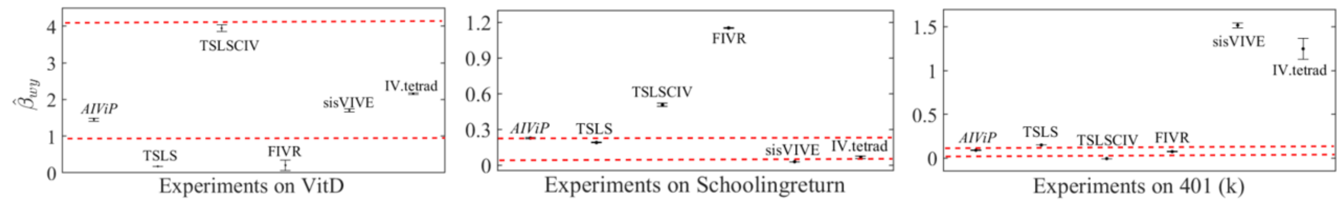


Figure 2: The estimated $\hat{\beta}_{wy}$ on the three real-world datasets. The two dotted lines indicate empirical 95% confidence interval of the references. Note that the performance of *AIViP* is consistent with the empirical values of the causal effects on all three real-world datasets.

proximity to a college, i.e. *nearcollege* as an instrument variable. We take $\hat{\beta}_{wy} = 13.29\%$ with 95% C.I. (0.0484, 0.2175) from [Verbeek, 2008] as the reference causal effect.

401(k) contains 9,275 individuals from the survey of income and program participation (SIPP) conducted in 1991 [Verbeek, 2008; Abadie, 2003]. The program participation is about the most popular tax-deferred programs, i.e. individual retirement accounts (IRAs) and 401 (k) plans. There are 11 variables about the eligibility for participating in the 401 (k) plan. The treatment is *p401k* (an indicator of participation in 401 (k)), and *pira* (a binary variable, $pira$ = 1 denotes participation in IRA) is the outcome of interest. *e401k* is used as an instrument for *p401k* (an indicator of eligibility for 401 (k)). We take $\hat{\beta}_{wy} = 7.12\%$ with 95% C.I. (0.047, 0.095) [Verbeek, 2008] as the reference causal effect.

Results. All results on the three datasets are visualized in Fig. 2. From Fig. 2, we have the following observations: (1) *AIViP* obtains results consistent with the reference causal effect values since the estimated causal effects are either in or close to the empirical 95% C.I. of the reference values on all three datasets. (2) The results of each comparison method are consistent with the reference values in at most two datasets. We note that IV.tetrad performs badly and this may attribute to the fact that its strong assumption on data distribution may not be satisfied. (3) *AIViP* has consistent performance across the three datasets, but all other methods' performances are not consistent across the three datasets and this may attribute to their failure in using the correct conditioning sets to reduce biases. The observations show the advantage of *AIViP* since it identifies the conditioning sets for reducing biases and does not have a strong assumption on data distributions.

5 Related Work

The IV method is a powerful tool in causal inference when the treatment and outcome are confounded by latent variables [Angrist and Imbens, 1995; Hernán and Robins, 2006]. It is impossible to test whether a variable is a valid standard IV from observational data alone. Assuming that all variables have discrete values, Pearl proposed *the instrumental inequality* to verify whether a variable is a valid IV [Pearl, 1995]. Kuroki and Cai proposed a criterion to find variables that satisfy the conditions of a standard IV in the linear structural model [Kuroki and Cai, 2005]. They provided a tighter condition than Pearl's [Pearl, 1995], and the developed method can be applied to data with continuous or discrete variables. Chu et al. [Chu *et al.*, 2001] proposed the concept of a semi-instrumental variable for a continuous variable. An IV is a semi-instrument, but the converse does not hold. Under the linearity assumption, Zhang et al. [Zhang *et al.*, 2020] proposed a symbiotic approach to causal discovery and identification by using a quasi-instrumental set. The four works reviewed above are either theoretical solutions or on a dataset with several variables (less than 5).

Kang et al. proposed a data-driven IV estimator, sisVIVE [Kang *et al.*, 2016]. sisVIVE requires that a set of candidate IVs and a set of observed variables are known and less than 50% of the candidate IVs are invalid. Hartford et al. proposed a deep learning based estimator to estimate β_{wy} from data [Hartford *et al.*, 2021]. This method also requires that less than 50% candidate IVs are invalid. Our work is different from these data-driven methods, as our work is about ancestral IVs and how to find a conditioning set from data.

The most relevant work to ours is the IV.tetrad method [Silva and Shimizu, 2017]. IV.tetrad aims to find a pair of valid conditional IVs $\{S_i, S_j\}$ from data by using the TETRAD constraint with the strong assumption of linear non-Gaussian causal models. In IV.tetrad, all observed variables in $\mathbf{X}$ excluding S_i and S_j are included in the conditional set $\mathbf{Z}$ that instrumentalizes S_i and S_j simultaneously. This assumption does not always satisfied and this limits the usefulness of IV.tetrad (as shown in our experiments). Different from IV.tetrad, we focus on finding a conditioning set $\mathbf{Z}$ that instrumentalizes a given ancestral IV S, to enable the practical use of conditional IVs.

6 Conclusion

One of the major challenges for the real-world application of causal effect estimation is the latent variables in a system, especially when the treatment and outcome share latent confounders. In this work, we study the graphical properties of an ancestral IV using a MAG to estimate causal effect from data with latent variables, including latent confounders. We have proposed the theory for supporting the search for a set of observed variables (a conditioning set) that instrumentalizes a given ancestral IV in a mapped MAG, as well as in a PAG for data-driven discovery of a conditioning set of a given ancestral IV. Based on the theory, we propose an algorithm, *AIViP* to achieve unbiased causal effect estimation from data with latent variables. The extensive experiments on synthetic and real-world datasets demonstrate that *AIViP* is very capable of handling data with latent confounders, even when the data contains collider bias, and *AIViP* outperforms the state-of-the-art estimators.

Acknowledgements

We wish to acknowledge the support from the Australian Research Council under DP200101210. JZ's research was supported in part by the RGC of Hong Kong under GRF13602720 and a start-up fund from HKBU.

References

[Abadie, 2003] Alberto Abadie. Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, 113(2):231–263, 2003.

[Angrist and Imbens, 1995] Joshua D Angrist and Guido W Imbens. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association*, 90(430):431–442, 1995.

[Athey *et al.*, 2019] Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.

[Brito and Pearl, 2002] Carlos Brito and Judea Pearl. Generalized instrumental variables. In *UAI*, pages 85–93, 2002.

[Card, 1993] David Card. Using geographic variation in college proximity to estimate the return to schooling. *Econometrica*, 69(9):1127–1160, 1993.

[Cheng *et al.*, 2022] Debo Cheng, Jiuyong Li, et al. Ancestral instrument method for causal inference without complete ksnowledge. *arXiv preprint arXiv:2201.03810*, 2022.

[Chu *et al.*, 2001] Tianjiao Chu, Richard Scheines, and Peter Spirtes. Semi-instrumental variables: a test for instrument admissibility. In *UAI*, pages 83–90, 2001.

[Colombo *et al.*, 2012] Diego Colombo, Marloes H Maathuis, et al. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, 40(1):294–321, 2012.

[Greene, 2003] William H Greene. *Econometric Analysis*. Pearson Education India, 2003.

[Hartford *et al.*, 2021] Jason S Hartford, Victor Veitch, et al. Valid causal inference with (some) invalid instruments. In *ICML*, pages 4096–4106. PMLR, 2021.

[Hernán and Robins, 2006] Miguel A Hernán and James M Robins. Instruments for causal inference: an epidemiologist's dream? *Epidemiology*, 17(4):360–372, 2006.

[Imbens and Rubin, 2015] Guido W Imbens and Donald B Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

[Imbens, 2014] Guido W Imbens. Instrumental variables: An econometrician's perspective. *Statistical Science*, 29(3):323–358, 2014.

[Kalisch *et al.*, 2012] Markus Kalisch, Martin Mächler, et al. Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, 47(11):1–26, 2012.

[Kang *et al.*, 2016] Hyunseung Kang, Anru Zhang, et al. Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. *Journal of the American Statistical Association*, 111(513):132–144, 2016.

[Kuroki and Cai, 2005] Manabu Kuroki and Zhihong Cai. Instrumental variable tests for directed acyclic graph models. In *AISTATS*, pages 190–197, 2005.

[Maathuis *et al.*, 2015] Marloes H Maathuis, Diego Colombo, et al. A generalized back-door criterion. *The Annals of Statistics*, 43(3):1060–1088, 2015.

[Martens *et al.*, 2006] Edwin P Martens, Wiebe R Pestman, et al. Instrumental variables: application and limitations. *Epidemiology*, 17(3):260–267, 2006.

[Martinussen and others, 2019] Torben Martinussen et al. Instrumental variables estimation under a structural cox model. *Biostatistics*, 20(1):65–79, 2019.

[Pearl, 1995] Judea Pearl. On the testability of causal models with latent and instrumental variables. In *UAI*, pages 435–443, 1995.

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

[Perković *et al.*, 2018] Emilija Perković, Johannes Textor, and Markus Kalisch. Complete graphical characterization and construction of adjustment sets in markov equivalence classes of ancestral graphs. *Journal of Machine Learning Research*, 18:1–62, 2018.

[Richardson and Spirtes, 2002] Thomas Richardson and Peter Spirtes. Ancestral graph markov models. *The Annals of Statistics*, 30(4):962–1030, 2002.

[Silva and Shimizu, 2017] Ricardo Silva and Shohei Shimizu. Learning instrumental variables with structural and non-gaussianity assumptions. *Journal of Machine Learning Research*, 18(120):1–49, 2017.

[Sjolander and Martinussen, 2019] Arvid Sjolander and Torben Martinussen. Instrumental variable estimation with the R package ivtools. *Epidemiologic Methods*, 8(1), 2019.

[Spirtes *et al.*, 2000] Peter Spirtes, Clark N Glymour, et al. *Causation, Prediction, and Search*. MIT Press, 2000.

[Vander Zander *et al.*, 2015] B. Vander Zander, Johannes Textor, et al. Efficiently finding conditional instruments for causal inference. In *IJCAI*, pages 3243–3249, 2015.

[Verbeek, 2008] Marno Verbeek. *A Guide to Modern Econometrics*. John Wiley & Sons, 2008.

[Zhang *et al.*, 2020] Chi Zhang, Bryant Chen, et al. A simultaneous discover-identify approach to causal inference in linear models. In *AAAI*, pages 10318–10325, 2020.

[Zhang, 2008a] Jiji Zhang. Causal reasoning with ancestral graphs. *Journal of Machine Learning Research*, 9(7):1437–1474, 2008.

[Zhang, 2008b] Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873–1896, 2008.