Efficiently Finding Conditional Instruments for Causal Inference

Benito van der Zander
University of Lübeck
Germany
benito@tcs.uni-luebeck.de

Johannes Textor
Utrecht University
The Netherlands
johannes.textor@gmx.de

Maciej Liśkiewicz
University of Lübeck
Germany
liskiewi@tcs.uni-luebeck.de

Abstract

Instrumental variables (IVs) are widely used to identify causal effects. For this purpose IVs have to be exogenous, i.e., causally unrelated to all variables in the model except the explanatory variable X. It can be hard to find such variables. A generalized IV method has been proposed that only requires exogeneity conditional on a set of covariates. This leads to a wider choice of potential IVs, but is rarely used yet. Here we address two issues with conditional IVs. First, they are conceptually rather distant to standard IVs; even variables that are independent of X could qualify as conditional IVs. We propose a new concept called *ancestral IV*, which interpolates between the two existing notions. Second, so far only exponential-time algorithms are known to find conditional IVs in a given causal diagram. Indeed, we prove that this problem is NP-hard. Nevertheless, we show that whenever a conditional IV exists, so does an ancestral IV, and ancestral IVs can be found in polynomial time. Together this implies a complete and constructive solution to causal effect identification using IVs in linear causal models.

1 Introduction

When studying a system that cannot be manipulated, we can only attempt to infer its cause-effect relationships from observational data. Conclusions drawn from observational studies are confounded when the putative cause and effect variables share common unobserved causes. This threat to causal inference is sometimes called the *endogeneity problem* [Antonakis *et al.*, 2010]. In practice, the endogeneity problem is often addressed by making the parametric assumptions needed to justify a linear regression model. Under these assumptions, it is possible to remove confounding by using an *instrumental variable* (instrument, IV) [Angrist *et al.*, 1996; Imbens, 2014a].

For instance, suppose our system under study can be described by the set of structural equations depicted in Fig. 1A, then (assuming all arrows describe linear functions and all variables have variance 1) we have $\mathrm{Cov}(X, Y) = \gamma + \lambda_1\lambda_2$, which differs from the causal effect γ. Yet, given that

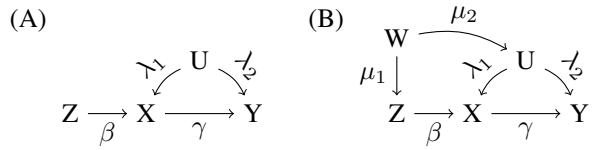


Figure 1: (A) The classic IV model. (B) Z is not an IV, but is a *conditional instrument* (conditional IV) given W.

$\mathrm{Cov}(Y, Z) = \beta\gamma$ and $\mathrm{Cov}(X, Z) = \beta$, we can estimate the causal effect as $\mathrm{Cov}(Y, Z)/\mathrm{Cov}(X, Z)$. This approach is called the *IV method*, which is valid under two conditions called *exogeneity* (Z shares no causes with and is not caused by Y nor U) and *exclusion restriction* (Z only affects Y through X). Both conditions hold in Fig. 1A. The conditions are untestable[1] and have to be justified from domain knowledge. However, the IV method can sometimes also be applied if exogeneity does not hold. For instance, Z is not exogenous in Fig. 1B, but by conditioning on W we get $\mathrm{Cov}(Y, Z \mid W)/\mathrm{Cov}(X, Z \mid W) = \gamma$. In such cases, we call Z a *conditional instrument*, and we say that W *instrumentalizes* Z.

This begs the more general question: When, and how, can a variable be instrumentalized using a covariate set $\mathbf{W}$? A graphical criterion exists to answer this question when a causal diagram (directed acyclic graph, DAG) is given [Pearl, 2009], but it requires exponential time to find $\mathbf{W}$. Indeed, we show here that this is an NP-hard problem. Nevertheless, we prove that the following related question can be answered constructively in polynomial time: Given a DAG, does any (perhaps conditional) IV exist? In other words, *finding* a conditional IV is easier than instrumentalizing a *given* variable. This surprising result implies that we can always efficiently find a conditional IV in a DAG if one exists.

Our paper starts with a brief outline of our notation and graphical formalisms. We divide our results into a conceptual and a computational part: Section 3 introduces a new three-level hierarchy of increasingly general IV definitions, which forms the basis for our results. Section 4 presents algorithms and hardness proofs.

[1] More precisely, instrumentality is testable to some extent if all variables are discrete [Pearl, 2009]. However, most practical methods assume X to be normally distributed and therefore continuous, in which case we cannot test instrumentality [Bonet, 2001].

2 Background

We denote sets in bold upper case ($\mathbf{S}$), and abbreviate singleton sets as $S = \{S\}$. Graphs are written calligraphically ($\mathcal{G}$), and variables in upper case (X). We consider graphs $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ with nodes (variables) $\mathbf{V}$ and directed ($A \to B$) edges $\mathbf{E}$. We assume that the set $\mathbf{V}$ is partitioned disjointly into a set of *latent* variables (that are not measured) and a set $\mathbf{M}$ of *observed* (measured) variables.

Nodes linked by an edge are *adjacent*. A *path* of length n is a node sequence $V_1, \dots, V_{n+1}$, in which no V_i occurs more than once, such that there exists an edge sequence $E_1, E_2, \dots, E_n$ for which every edge E_i connects V_i, V_{i+1}. Then V_1 is called the *start node* and V_{n+1} the *end node* of the path. We use the kinship terms *child*, *parent*, *ancestor* and *descendant* to describe node relationships in graphs in the same way as in [Pearl, 2009]; in this convention, every node is an ancestor (but not a parent) and a descendant (but not a child) of itself. For a node set $\mathbf{Y}$ we denote by $An(\mathbf{Y})$ the set of all ancestors of nodes in $\mathbf{Y}$. Analogously, the descendant set $De(\mathbf{Y})$ is the set of all descendants of any node in $\mathbf{Y}$. Given a graph $\mathcal{G}$ and a node set $\mathbf{Y} \subseteq \mathbf{V}$, the *ancestor graph* $\mathcal{G}_{An(\mathbf{Y})}$ is the subgraph of $\mathcal{G}$ consisting only of the nodes in $An(\mathbf{Y})$ and all edges between them.

A path from a node X to Y is called *causal* or *directed* if it only contains directed edges pointing away from X. A graph is a DAG if it does not contain a directed path from a node to itself of length > 1. A node V on a path π is called a *collider* if two arrowheads of π meet at V, e.g. if π contains $U \to V \leftarrow Q$. There can be no collider if π is shorter than 2. Two nodes Z, Y are called *d-connected* by a set $\mathbf{W}$ if there is a path π between them on which every collider is an ancestor of $\mathbf{W}$ and every non-collider is not in $\mathbf{W}$. Then π is called a *d-connecting* or *active* path. If Z, Y are not d-connected by $\mathbf{W}$, we say that $\mathbf{W}$ *d-separates* them or *blocks* all paths between them. We use the notation $(Z \perp\!\!\!\perp Y \mid \mathbf{W})_{\mathcal{G}}$ to indicate this separation, analogously $(Z \not\perp\!\!\!\perp Y \mid \mathbf{W})_{\mathcal{G}}$ if Z, Y are d-connected by $\mathbf{W}$. If Z, Y are d-connected (d-separated) by the empty set, we simply say they are d-connected (d-separated). For a path π, we denote by $\pi[X \sim Z]$ the subpath of π consisting of the edges between X and Z.

Given a directed graph $\mathcal{G}$, the *moral graph* $\mathcal{G}^m$ [Lauritzen *et al.*, 1990] is the undirected graph obtained by transforming $\mathcal{G}$ as follows: (1) For all pairs of edges of the form $A \to B$, $C \to B$, if A and C are not adjacent in $\mathcal{G}$, add an undirected edge $A - C$ to $\mathcal{G}$. (2) Substitute every directed edge $A \to B$ by an undirected edge $A - B$.

3 A Hierarchy of Conditional Instruments

Because IVs rely on causal assumptions, it is natural to base IV definitions on causal models. Directed acyclic graphs (DAGs) are simple causal models consisting of nodes representing variables and edges representing causal relationships. In this paper, we focus on the acyclic case in which reciprocal causation is not allowed. We assume that a DAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ with measured variables $\mathbf{M} \subseteq \mathbf{V}$ is given. Also we assume that $\mathcal{G}$ contains an edge $X \to Y$ (the effect to be estimated), and we denote by $\mathcal{G}_c$ the graph obtained by removing said edge from $\mathcal{G}$.

As mentioned above, exogeneity and exclusion restriction are statistically untestable – arguably, they cannot even be expressed in statistical language [Pearl, 2009]. Conversely, requirements that *can* be formulated in statistical language have the advantage of being testable. We make this distinction explicit by expressing testable requirements in statistical language (labeled with *), and untestable ones[2] in graphical language. The standard IV definition then reads as follows:

Definition 3.1 (IV). *Z is an* instrumental variable *relative to $X \to Y$, if*

*(a) Z correlates with X,

(b) Z is d-separated from Y in $\mathcal{G}_c$.

Note that (a) implies the existence of a path π from Z to X that is d-connected. Our definition is thus stricter than the one given by [Pearl, 2009], in which (a) only requires the existence of π. However, for actual estimation of causal effects e.g. by two-stage regression, Z needs to be correlated with X, so our definition does not miss any relevant cases.

Definition 3.2 (Conditional Instrument [Pearl, 2009]). *Z is said to be a* conditional instrument *relative to $X \to Y$, if there exists a set $\mathbf{W} \subseteq \mathbf{M}$ such that*

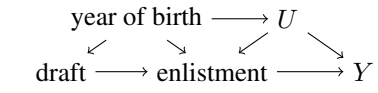
*(a) Z correlates with X conditioned on $\mathbf{W}$,

(b) $\mathbf{W}$ d-separates Z and Y in $\mathcal{G}_c$,

(c) $\mathbf{W}$ consists of non-descendants of Y.

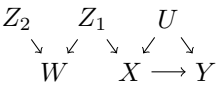
For instance, Z is a conditional instrument in Fig. 1B using $\mathbf{W} = W$. Again, condition (a) is statistically testable, and it implies the existence of a path π from Z to X that is d-connected by $\mathbf{W}$. Conditions (a) and (b) are direct generalizations of the same conditions in the standard IV definition. Restriction (c) is necessary because conditioning on descendants of Y would bias estimation by the reversal paradox.

The idea of applying instruments together with conditioning variables predates its graphical definition. Take the seminal work of Angrist on the labor market impact of voluntary military service [Angrist, 1998; Angrist and Pischke, 2008]: During the "Vietnam draft lottery", randomly chosen men were called to serve in the war. As not every drafted person enlisted, this is a classical example of a randomized trial with imperfect compliance: The IV Z in this case is the draft, and the independent variable X is enlistment. Several dependent variables Y could be studied with this setup. However, because different numbers of men were drafted for each birth cohort, the IV was only exogenous conditioned on the year of birth (this would be our $\mathbf{W}$ in Definition 3.2). A DAG describing this scenario could look like this:



In this typical example, we condition on a set of covariates to render a variable exogenous. However, Definition 3.2 also allows for quite different scenarios that no longer fit the usual IV setting. For example, in the DAG

[2] By this we mean "untestable from observed data". In principle we could test some of these requirements experimentally.



Z_1 is a canonical IV. But Definition 3.2 also allows Z_2 as an IV after conditioning on W: even though Z_1 and X are initially independent, they (may) become correlated after conditioning. Whilst this is technically fine, starting from a variable that is uncorrelated with X and then essentially exploiting the reversal paradox might lend less credibility to such an analysis. In practice, good candidate IVs are often found when circumstances create a scenario akin to (imperfect) randomization, like in the Vietnam lottery example. We would then use conditioning to remove any residual endogeneity from the instrument, not to create spurious correlations. Natural covariates would be the common causes of Z, X and Y (or proxies thereof). With this motivation in mind, we define the following "restricted" version of a conditional IV.

Definition 3.3 (Ancestral Instrument). *Variable Z is said to be an* ancestral instrument *relative to $X \to Y$, if there exists a set $\mathbf{W} \subseteq \mathbf{M}$ such that*

(a) Z correlates with X conditional on $\mathbf{W}$,

(b) $\mathbf{W}$ d-separates Z and Y in $\mathcal{G}_c$,

(c) $\mathbf{W}$ consists of ancestors of Y or Z or both which are non-descendants of Y.

From this definition, it is obvious that ancestral IVs are special case of conditional IVs. Also, each standard IV is an ancestral IV (using $\mathbf{W} = \emptyset$). There are, however, conditional IVs that are not ancestral (e.g. Z_2 in the above DAG), and this might seem to imply that ancestral IVs are less "powerful" than generic conditional ones. Importantly and perhaps surprisingly, the following theorem shows that this is not the case.

Theorem 3.4. *For a given DAG $\mathcal{G}$ and variables X and Y, a conditional IV Z relative to $X \to Y$ exists if and only if an ancestral IV Z' relative to $X \to Y$ exists.*

Proof. Let Z be a conditional IV relative to $X \to Y$, such that the path π from X to Z has the minimum number of colliders. Let $\mathbf{W}$ be a minimal set of non-descendants of Y opening the path π and separating Y and Z in $\mathcal{G}_c$. If π has no colliders then after removing nodes from $\mathbf{W}$ which are neither ancestors of Y nor of Z the set still d-separates Y and Z [Lauritzen *et al.*, 1990] and does not block π. Thus, Z is an ancestral instrument.

Assume π has at least one collider, say C. Let $W \in \mathbf{W}$ be the first descendant of C in $\mathcal{G}_c$. Note that W can coincide with C. Then all paths τ between Y and W that are active given $\mathbf{W} \setminus W$ in $\mathcal{G}_c$ have the form $\tau = \tau_0 \leftarrow W$, where τ_0 starts with Y, and at least one such path exists: Otherwise, if no open path between Y and W exists, W is a conditional instrument relative to $X \to Y$ whose sequence $\pi[X \sim C] \to \ldots \to W$ given $\mathbf{W} \setminus W$ in $\mathcal{G}_c$ has fewer colliders than π. If the sequence does not visit a node twice, it is an active path; Otherwise, if a node, say D, is visited on the sequence twice, we can remove all nodes between the farthest occurrences of D resulting in an active path $\pi[X \sim D] \to \ldots \to W$. If path τ ends with $\to$, i.e. $\tau = \tau_0 \to W$ and no node occurs twice in the sequence $\tau_0 \to W \leftarrow \ldots \leftarrow \pi[C \sim Z]$ then it is a path between Y and Z in $\mathcal{G}_c$ which is open given $\mathbf{W}$. If a node, say D, occurs twice in the sequence, we remove all nodes between the farthest occurrences of D obtaining a shorter sequence with a single occurrence of D. In the shorter sequence, every collider is open given $\mathbf{W}$ and no non-collider is blocked by $\mathbf{W}$. This fact is obvious for any node $V \neq D$. To see that it is true also for D, we consider two cases. If in the shorter sequence node D is a non-collider, then at least one of its occurrences in the original sequence was a non-blocked non-collider. Thus, $\mathbf{W}$ does not block D in the shorter sequence either. If D becomes a collider but in the original sequence both occurrences were non-colliders, then in the original sequence the subsequence between both occurrences has the form $\to D \to \ldots \leftarrow D \leftarrow$. This means that a descendant of D is in $\mathbf{W}$.

If one of these first descendants $W \in \mathbf{W}$ is not an ancestor of Y then there exists a path τ of the form $\tau_1 \to V \leftarrow \ldots \leftarrow W$, for some subpath τ_1 starting in Y that is active given $\mathbf{W} \setminus W$ in $\mathcal{G}_c$. Node V has a descendant in $\mathbf{W} \setminus W$.

Since $\mathbf{W}$ is minimal no node $W \in \mathbf{W}$ that does not open the path between X and Z can be removed without opening a path π' between Y and Z in $\mathcal{G}_c$. Since such a path π' has to contain a collider, W is an ancestor of this collider and a node $W' \in \mathbf{W}$ that is opening the collider.

So, every node in $\mathbf{W}$ is either an ancestor of Y or Z, or is an ancestor of some other node in $\mathbf{W}$. Since $\mathcal{G}_c$ does not contain cycles, every node in $\mathbf{W}$ is an ancestor of Y or Z. □

The result that we lose nothing by restricting ourselves to ancestral IVs has an interesting connection to a similar result for covariate adjustment: If it is possible to identify a causal effect by adjustment, then it is always possible to do so by only adjusting for ancestors of the variables of interest [van der Zander *et al.*, 2014].

The rest of the paper will show that ancestral instruments are algorithmically appealing: Unlike non-ancestral instruments, they can be found efficiently in a given DAG.

4 Finding Conditional Instruments

In this section we are concerned with the following problem: Given a DAG $\mathcal{G}$ and variables X, Y, can we find a variable Z and a set $\mathbf{W} \subseteq \mathbf{M}$ that renders Z into a conditional instrument with respect to $X \to Y$? If this is possible, we say that Z can be *instrumentalized* using $\mathbf{W}$. If many such $\mathbf{W}$ exist, we give a preference to "simple" sets – e.g., we prefer to instrumentalize using the empty set if possible. Our main workhorse to achieve this is the notion of a *nearest separator*, which we introduce now.

4.1 Nearest Separators

Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a graph and let $\mathbf{M} \subseteq \mathbf{V}$ denote the measured nodes and let Y and Z be nodes in $\mathbf{V}$. We say that Y and Z are *separable* in $\mathcal{G}$ if there exists $\mathbf{W} \subseteq \mathbf{M}$ such that $(Z \perp\!\!\!\perp Y \mid \mathbf{W})_{\mathcal{G}}$. For given nodes Y and Z in $\mathbf{V}$ we call a subset $\mathbf{W} \subseteq \mathbf{M}$ a *nearest separator* according to (Y, Z) if and only if (i) $(Z \perp\!\!\!\perp Y \mid \mathbf{W})_{\mathcal{G}}$ and (ii) for all $X \in An(Y \cup Z) \setminus \{Y, Z\}$ and any path π in the moral graph $(\mathcal{G}_{An(Y \cup Z)})^m$ connecting X and Z, if there exists $\mathbf{W}' \subseteq \mathbf{M}$

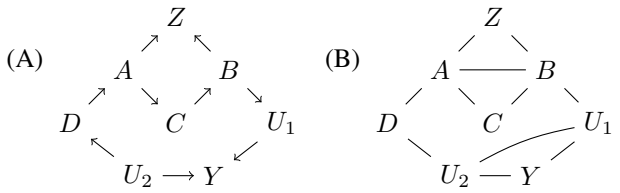


Figure 2: (A) DAG $\mathcal{G}$ with unmeasurable U_1, U_2 and (B) the corresponding moral graph. $\{D, B\}$ is a nearest separator according to (Y, Z) but $\{A, B\}$ not since it does not satisfy (ii).

such that $(Z \perp\!\!\!\perp Y \mid \mathbf{W}')_{\mathcal{G}}$ and $\mathbf{W}'$ does not block π then $\mathbf{W}$ does not block π either. For an example nearest separator see Fig. 2. The following algorithm computes such separators.

function NEAREST-SEPARATOR($\mathcal{G}, Y, Z$)
  Let $\mathbf{M}$ be the set of measured variables in $\mathcal{G}$
  Construct the moralized graph $\mathcal{M} = (\mathcal{G}_{An(Y \cup Z)})^m$
  $\mathbf{W} := \emptyset$
  while $\exists\, \pi = Y, V_1, \ldots, V_k, Z$ - a path from Y to Z in $\mathcal{M}$ s.t. $k \geq 1$, π is not blocked by $\mathbf{W}$, and $\{V_1, \ldots, V_k\} \cap \mathbf{M} \neq \emptyset$ **do**
    $\mathbf{W} := \mathbf{W} \cup \{\text{first measurable node } V_i \text{ of } \pi\}$
  if $(Z \perp\!\!\!\perp Y \mid \mathbf{W})$ **then return W else return** $\perp$

Lemma 4.1. *Algorithm* NEAREST-SEPARATOR *finds a nearest separator $\mathbf{W} \subseteq An(Y \cup Z)$ if Y and Z are separable in $\mathcal{G}$; otherwise it returns $\perp$. Moreover, if Y and Z can be separated in $\mathcal{G}$ by a set that does not contain a descendant of Y, then $\mathbf{W} \subseteq An(Y \cup Z) \setminus De(Y)$. The runtime of the algorithm is $\mathcal{O}(nm)$, where $n = |\mathbf{V}|$ and $m = |\mathbf{E}|$.*

Proof. The algorithm finds $\mathbf{W}$ satisfying $(Z \perp\!\!\!\perp Y \mid \mathbf{W})_{\mathcal{G}}$ if such a set exists, because in this case it exists in the moral graph of ancestors [Lauritzen *et al.*, 1990] and in an undirected graph a separating set can be found greedily.

This $\mathbf{W}$ contains only ancestors of Z and Y, so if $\mathbf{W}$ contains a descendant V of Y then V is an ancestor of Z. Hence there is a causal path from Y to V and a causal path from V to Z. Thus, there exists a causal path from Y to Z that can only be blocked by a descendant of Y, and every separating set must contain a descendant of Y. This proves the condition that the algorithm returns a set of non-descendants of Y if such a set exists.

Next, we show that the set $\mathbf{W}$ returned by the algorithm is a nearest separator according to (Y, Z) in $\mathcal{G}$. So, let X be an arbitrary node in $An(Y \cup Z)$ other than Y and Z and let π be a path in the moral graph $(\mathcal{G}_{An(Y \cup Z)})^m$ connecting X and Z such that π is blocked by $\mathbf{W}$. Assume there exists a set $\mathbf{W}' \subseteq \mathbf{M}$ which separates Y and Z in $\mathcal{G}$ but does not block the path π in the moral graph. Then π is blocked by a node $W \in \mathbf{W} \setminus \mathbf{W}'$ which has to block a path π' from Y to Z in the moral graph. This follows from the construction of the algorithm. The subpath of π' between Y and W is not blocked by $\mathbf{W}'$, because the algorithm only chooses W, if no node closer to Y can block the path π'. Hence every path from W to Z in $(\mathcal{G}_{An(Y \cup Z)})^m$ is blocked by $\mathbf{W}'$. Particularly, a subpath of π between W and Z is blocked by $\mathbf{W}'$, too. But this contradicts the assumption that $\mathbf{W}'$ does not block π.

The runtime is $\mathcal{O}(nm)$ since a path in the moralized graph can be found in $\mathcal{O}(m)$ steps and at most $\mathcal{O}(n)$ nodes are needed to block all paths. The times needed to construct the moralized graph ($\mathcal{O}(n)$) and to check the separation at the end ($\mathcal{O}(m)$) do not change the asymptotic runtime. □

The following property of nearest separators computed by the algorithm above will turn out to be useful.

Corollary 4.2. *Let X, Y, Z be different variables in $\mathbf{V}$ and let $\mathbf{W} \neq \perp$ be a set returned by* NEAREST-SEPARATOR *for $(\mathcal{G}_c, Y, Z)$. If $X \in \mathbf{W}$, then variable Z is not an ancestral instrument relative to $X \to Y$ in $\mathcal{G}$.*

4.2 Ancestral Instruments

Having introduced the concept of nearest separators and having shown how to find them, we are now ready to present an algorithm to instrumentalize ancestral IVs.

function ANCESTRAL-INSTRUMENT($\mathcal{G}, X, Y, Z$)
  $\mathcal{G}_c := \mathcal{G}$ with edge $X \to Y$ removed
  $\mathbf{W} :=$ NEAREST-SEPARATOR($\mathcal{G}_c, Y, Z$)
  if $(\mathbf{W} = \perp) \vee (\mathbf{W} \cap De(Y) \neq \emptyset) \vee (X \in \mathbf{W})$ **then return** $\perp$
  if $(Z \not\perp\!\!\!\perp X \mid \mathbf{W})$ in $\mathcal{G}_c$ **then return W else return** $\perp$

Theorem 4.3. *For given X, Y and Z in a DAG $\mathcal{G}$ algorithm* ANCESTRAL-INSTRUMENT *returns a set of variables $\mathbf{W}$ that satisfies the properties of ancestral conditional instruments relative to $X \to Y$, if such a set exists; Otherwise it returns $\perp$. The running time of the algorithm is $\mathcal{O}(nm)$.*

Proof. Let $\mathbf{A} = \mathbf{M} \cap An(Y, Z) \setminus De(Y)$. We prove first that if the algorithm returns a set $\mathbf{W} \neq \perp$, it has found $\mathbf{W}$ satisfying the conditions of Definition 3.3.

From Lemma 4.1 we know that NEAREST-SEPARATOR returns a nearest separator $\mathbf{W}$ according to (Y, Z) and that $\mathbf{W} \subseteq \mathbf{M} \cap An(Y, Z)$. Due to the test for descendants of Y we have that $\mathbf{W} \subseteq \mathbf{A}$. From Corollary 4.2 we know that Z cannot be an ancestral instrument if $X \in \mathbf{W}$. Finally, a set $\mathbf{W}$ is returned if it does not separate X and Z in $\mathcal{G}_c$.

Next, assume there exists a set $\mathbf{W}_0 \subseteq \mathbf{A}$, such that $(Y \perp\!\!\!\perp Z \mid \mathbf{W}_0)_{\mathcal{G}_c}$ and a path π_0 between X and Z which is open in $\mathcal{G}_c$ given $\mathbf{W}_0$. We show that the algorithm finds a set $\mathbf{W}$ satisfying the conditions of ancestral instruments.

From the assumption and due to Lemma 4.1, we get that NEAREST-SEPARATOR returns a set $\mathbf{W} \subseteq \mathbf{A}$ which is a nearest separator according to (Y, Z). If π_0 has colliders, then all of them are ancestors of nodes in $\mathbf{W}_0$ that, recall, is a subset of $An(Y, Z)$. So, π_0 has the form $X \leftarrow \ldots \leftarrow Z$, or $X \leftarrow \ldots \leftarrow V \rightarrow \pi_1$, or $X \rightarrow \pi_2$ such that V and all nodes on π_1 and π_2 are ancestors of Y or Z. Paths $V \rightarrow \pi_1$ and $X \rightarrow \pi_2$ correspond to paths in the moralized graph $\mathcal{M} = (\mathcal{G}_{c, An(Y,Z)})^m$, which are not blocked by $\mathbf{W}_0$ and thus not blocked by $\mathbf{W}$ in $\mathcal{M}$. Relying on the properties of the moralized graph, we can prove that there exist paths $V \rightarrow \pi_1'$ or $X \rightarrow \pi_2'$ to Z in $\mathcal{G}_c$ that are not blocked by $\mathbf{W}$ and $V \notin \mathbf{W}$ due to Corollary 4.2. The paths $X \leftarrow \ldots \leftarrow Z$ and $X \leftarrow \ldots \leftarrow V$ are not blocked by $\mathbf{W}$ either, since if the nodes belong to the moral graph, they are not blocked due to Lemma 4.1, otherwise $\mathbf{W}$ cannot block them since it only

contains nodes of the moral graph. Similarly $V \notin \mathbf{W}$, or there would be a path between Y and Z that is not blocked by $\mathbf{W}_0$. Thus, replacing π_1, π_2 by π_1', π_2', resp., in π_0 leads to a path π_0' that is not blocked by $\mathbf{W}$ and the condition $(Z \not\perp\!\!\!\perp X \mid \mathbf{W})_{\mathcal{G}_c}$ is true. The runtime is dominated by NEAREST-SEPARATOR. □

It is easy to see that if Z is d-separated from Y in $\mathcal{G}_c$, i.e. if $(Z \perp\!\!\!\perp Y \mid \emptyset)_{\mathcal{G}_c}$, then for given $\mathcal{G}_c, Y$, and Z, algorithm NEAREST-SEPARATOR, returns $\mathbf{W} = \emptyset$. Thus, we get:

Corollary 4.4. *For given X, Y and Z in a DAG $\mathcal{G}$ algorithm* ANCESTRAL-INSTRUMENT *returns the empty set $\mathbf{W} = \emptyset$ if and only if Z is an instrumental variable relative to $X \to Y$.*

Further, we can use algorithm ANCESTRAL-INSTRUMENT to find a conditional instrumental variable relative to $X \to Y$: we search exhaustively in $\mathbf{M} \setminus (X \cup De(Y))$ for a variable Z for which the algorithm returns $\mathbf{W} \neq \perp$. The soundness of the algorithm and its time complexity $\mathcal{O}(n^2 m)$ follows from Theorem 4.3. The completeness is a consequence of Theorem 3.4. We obtain the following result.

Corollary 4.5. *There exists an algorithm which, given X and Y, returns a node Z and a node set $\mathbf{W}$ in time $\mathcal{O}(n^2 m)$ such that $\mathbf{W}$ instrumentalizes Z, if such $\mathbf{W}$ and Z exist. Otherwise, it returns $\perp$.*

This corollary is *complete* for effect identification using conditional IVs in the same sense as the do-calculus is complete for causal effect identification in general [Huang and Valtorta, 2006; Shpitser and Pearl, 2006]: if it is possible to estimate a causal effect in a DAG using a conditional IV, then we can find such an IV using our algorithm.

4.3 Instrumentalization is NP-hard in general

We have now solved the problem posed in the previous section: find a variable Z and a set $\mathbf{W}$ such that $\mathbf{W}$ instrumentalizes Z. Now it is natural to wonder about a slightly different problem: given Z, find a set $\mathbf{W}$ that instrumentalizes Z. We refer to this as the *instrumentalization problem*. Intuitively, this new problem might seem to be easier than finding an IV because Z is already fixed. Perhaps surprisingly, the opposite turns out to be true: Instrumentalization is computationally *harder* than finding an IV.

Theorem 4.6. *Determining if, for given $X, Y, Z \in \mathbf{V}$, node Z is a conditional instrument relative to $X \to Y$ is an NP-complete problem.*

Proof. Obviously, the conditions of Definition 3.2 can be verified in polynomial time after guessing $\mathbf{W} \subseteq \mathbf{M}$. Thus, the problem is in NP. To prove the NP-hardness we show a reduction from the 3SAT problem, which is the canonical complete problem for NP [Garey and Johnson, 1979].

Assume $\varphi = \bigwedge_{j=1}^{m} \mathcal{C}_j$ is an instance of 3SAT, which is a Boolean formula in conjunctive normal form over n variables $x_1, \ldots, x_n$ where each clause $\mathcal{C}_j$ is limited to exactly three literals from $x_1, \overline{x}_1, \ldots, x_n, \overline{x}_n$. We construct the DAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ for φ as follows.

For every clause $\mathcal{C}_j$ we define two nodes C_j and C_j' in $\mathbf{V}$. Next, let o_i, resp. $\overline{o}_i$, be the number of occurrences of positive literal x_i, resp. negative literal $\overline{x}_i$, in φ. For each literal x_i we define o_i nodes $V_i^1, \ldots, V_i^{o_i}$ in $\mathbf{V}$, resp., for $\overline{x}_i$ we define nodes $\overline{V}_i^1, \ldots, \overline{V}_i^{\overline{o}_i}$. Additionally, for every x_i two nodes V_i^0, V_i^{-1} are defined and for every $\overline{x}_i$ the node $\overline{V}_i^0$ is included in $\mathbf{V}$. We complete the construction of $\mathbf{V}$ by adding the sets of nodes P_i, F_i, N_i, for all $1 \leq i \leq n$ and nodes X, Z, Y, Y', Y''. Thus,

$$\begin{aligned}\mathbf{V} = &\{X, Z, Y, Y', Y''\} \\ &\cup \{V_i^k \mid 1 \leq i \leq n \wedge -1 \leq k \leq o_i\} \\ &\cup \{\overline{V}_i^k \mid 1 \leq i \leq n \wedge 0 \leq k \leq \overline{o}_i\} \\ &\cup \{P_i, F_i, N_i \mid 1 \leq i \leq n\} \cup \{C_i, C_i' \mid 1 \leq i \leq n\}.\end{aligned}$$

We define the edges of $\mathcal{G}$ such that C_i', C_i form the path $X \to C_1' \to C_1 \leftarrow C_2' \to C_2 \leftarrow \ldots \leftarrow C_m' \to C_m \leftarrow Z$ between X and Z (see the top of Fig. 3), and the nodes P_i, F_i, N_i form the n paths $Y' \to P_i \leftarrow F_i \to N_i \leftarrow Y''$ between Y' and Y'' (see the bottom of Fig. 3). The complete set of edges $\mathbf{E}$ is defined as follows:

$$\begin{aligned}\mathbf{E} = &\{X \to Y, X \to C_1', Y' \to C_1', C_1' \to C_1, C_m \leftarrow Z\} \\ &\cup \{C_{j-1} \leftarrow C_j' \to C_j \mid 1 < j \leq m\} \\ &\cup \{C_j \to V_i^k \mid k\text{-th occurrence of } x_i \text{ in } \varphi \text{ is in } \mathcal{C}_j\} \\ &\cup \{C_j \to \overline{V}_i^k \mid k\text{-th occurrence of } \overline{x}_i \text{ in } \varphi \text{ is in } \mathcal{C}_j\} \\ &\cup \{V_i^{-1} \to V_i^0, V_i^0 \to V_i^k \mid 1 \leq i \leq n \wedge 1 \leq k \leq o_i\} \\ &\cup \{\overline{V}_i^0 \to \overline{V}_i^k \mid 1 \leq i \leq n \wedge 1 \leq k \leq \overline{o}_i\} \\ &\cup \{Y' \to P_i \leftarrow F_i \to N_i \leftarrow Y'' \mid 1 \leq i \leq n\} \\ &\cup \{P_i \to V_i^{-1}, N_i \to \overline{V}_i^0 \mid 1 \leq i \leq n\} \\ &\cup \{Y'' \to V_i^0 \mid 1 \leq i \leq n\} \cup \{Y'' \to Y\}.\end{aligned}$$

The set $\mathbf{M}$ of observed nodes is defined as

$$\mathbf{M} = \{V_i^j, \overline{V}_i^k \mid 1 \leq i \leq n \wedge -1 \leq j \leq o_i \wedge 0 \leq k \leq \overline{o}_i\}.$$

$\mathcal{G}$ can be constructed in polynomial time with respect to the length of the instance formula φ. The construction ensures that sets $\mathbf{W}$ that instrumentalizes Z w.r.t. X, Y bijectively map to satisfying assignments as follows:

$$x_i = \begin{cases} \text{true} & \text{if } \exists k : V_i^k \in \mathbf{W} \\ \text{false} & \text{otherwise.} \end{cases}$$

Our construction ensures the validity of this mapping in the following manner. First, the only possible open path from X to Z is the path via the C_i, C_i' nodes; if $\mathbf{W}$ opens any other path, then it also opens a path from Z to Y. Second, to connect Z and X, set $\mathbf{W}$ needs to contain at least one of $V_i^j, \overline{V}_i^k$ for $i \geq 1$. But it is not possible to pick both V_i^j and $\overline{V}_i^k$ for any $i \geq 1, j, k$ without opening a path from Z to Y. This ensures that no variable is assigned both "true" and "false". If a path has been found that is d-connected by $\mathbf{W}$, then it therefore has to contain an assignment for at least one variable in every clause, and this assignment fulfills said clause. Therefore, we altogether obtain a satisfying assignment. If no such assignment exists, then it is also not possible simultaneously to d-connect Z and X and d-separate Z and Y in the graph. □

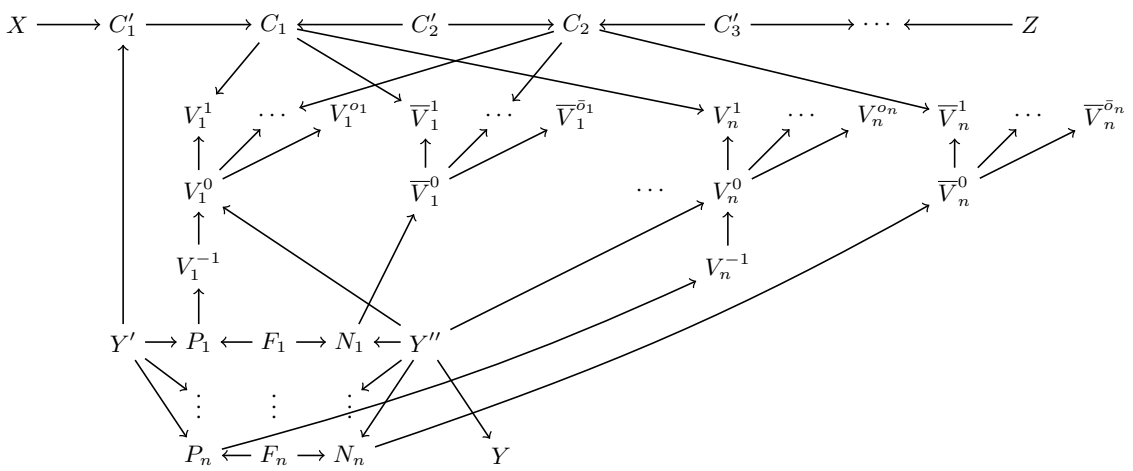


Figure 3: Reduction of 3SAT to the instrumentalization problem (the edge $X \to Y$ is omitted). Each variable C_i stands for a clause in the input formula.

5 Discussion

We proposed a three-level hierarchy that links traditional IVs [Angrist *et al.*, 1996] to conditional IVs [Pearl, 2009]. For the first two stages, classic and ancestral instruments, we gave efficient algorithms for finding IVs and for instrumentalizing given variables. At the same time, we showed that instrumentalization is in general not efficiently solvable unless P=NP. Yet, whenever a conditional IV exists, an ancestral IV also exists. In this sense, our solution is complete – it covers all cases where a causal effect can be identified using the conditional IV method. To our knowledge, no attempt has been made so far to find conditional IVs efficiently or to prove the hardness of this problem, and perhaps due to these algorithmic issues, the existing software that we know of is limited to finding unconditional IVs [Kyono, 2010].

We focused here on estimation of causal effects of the form $X \to Y$, disregarding scenarios where the effect of X on Y is mediated by other observed variables. However, our results easily generalize to IVs with respect to the total effect. Details will appear in an extended version of this paper. On the other hand, IVs are also often used to disentangle reciprocal causation in cyclic models such as typical supply-and-demand systems [Angrist and Pischke, 2008], and that important case remains to be addressed in future research.

From a computational complexity perspective, the result that instrumentalization is hard whereas finding a conditional IV is easy is rather intriguing. This can be explained by noting that the solution space of the IV problem decomposes into some instances that are easy to find (ancestral IVs) and others that are hard to find (non-ancestral IVs like the ones used in our reduction). Our hardness proof heavily uses long paths on which every observed variable is a collider. In DAGs with only observed variables (Markovian DAGs), such long collider paths are impossible. Thus, the question whether non-ancestral IVs are easier to find in Markovian DAGs remains open. However, this question is not of immediate practical relevance since in Markovian DAGs, we can identify causal effects using hierarchical regression or using adjustment, which have superior statistical properties.

We hope our findings will benefit both users of DAGs and users of standard IV methods, for partly different reasons. With DAGs we have the do-calculus at our disposal, which always finds a formula to estimate a causal effect if one exists [Huang and Valtorta, 2006; Shpitser and Pearl, 2006]. However, these formulas can grow large, which complicates estimation [VanderWeele, 2009; Glynn and Kashin, 2013]. IV methods are statistically well-understood, and can identify causal effects in many settings where standard approaches like covariate adjustment fail, e.g., when an unobserved confounder affects X and Y. Yet, IV methods are rarely used so far in fields where DAGs are popular, like Epidemiology [Greenland, 2000]. Phrasing generalized IV methods in the DAG framework and providing efficient algorithms for IV construction may make them more palatable for DAG users, and embracing IV estimation procedures for causal effects might benefit those fields.

Econometrics is perhaps the field where IV methods are most commonly used [Imbens, 2014a]. Econometricians have thus far largely refrained from using graphical causal models [Pearl, 2013; Imbens, 2014b], and might not be comfortable with Pearl's fully generalized IV definition that includes IVs which are uncorrelated with X. Therefore, we hope that interested Econometricians might find some merit in our notion of ancestral IVs, which reflects the combined use of IVs and covariates for adjustment. In this context, it is interesting to note that the linearity requirement of the IV method does not necessarily apply to the covariates: an estimation approach has been developed into which the covariates can enter non-parametrically [Kasy, 2009].

In summary, we have presented a hierarchy of generalized IV definitions and gave efficient algorithms to find generalized IVs for all levels of this hierarchy. We hope our results will help DAG users to adopt the IV method, and users of the IV method to adopt DAGs. Our algorithms are implemented in the open-source software DAGitty [Textor *et al.*, 2011].

References

[Angrist and Pischke, 2008] Joshua D. Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2008.

[Angrist *et al.*, 1996] Joshua D. Angrist, Guido W. Imbens, and Donald B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996.

[Angrist, 1998] Joshua D. Angrist. Estimating the labor market impact of voluntary military service using social security data on military applicants. *Econometrica*, 66(2):249–288, 1998.

[Antonakis *et al.*, 2010] John Antonakis, Samuel Bendahan, Philippe Jacquart, and Rafael Lalive. On making causal claims: A review and recommendations. *The Leadership Quarterly*, pages 1086–1120, 2010.

[Bonet, 2001] Blai Bonet. Instrumentality tests revisited. In *Proceedings of UAI*, pages 48–55, 2001.

[Garey and Johnson, 1979] Michael Garey and David Johnson. *Computers and intractability: a guide to the theory of NP-completeness*. W. H. Freeman & Co, 1979.

[Glynn and Kashin, 2013] Adam Glynn and Konstantin Kashin. Front-door versus back-door adjustment with unmeasured confounding: Bias formulas for front-door and hybrid adjustments. Technical report, Harvard University, 2013.

[Greenland, 2000] Sander Greenland. An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology*, 29(4):722–729, Aug 2000.

[Huang and Valtorta, 2006] Yimin Huang and Marco Valtorta. Pearl's calculus of intervention is complete. In *UAI '06, Proceedings of the 22nd Conference in Uncertainty in Artificial Intelligence, Cambridge, MA, USA, July 13-16, 2006*, 2006.

[Imbens, 2014a] Guido Imbens. Instrumental variables: An econometricians perspective. *Statistical Science*, 29(3):323–358, 2014.

[Imbens, 2014b] Guido Imbens. Rejoinder of "instrumental variables: An econometrician's perspective". *Statistical Science*, 29(3):375–379, 2014.

[Kasy, 2009] Maximilian Kasy. Semiparametrically efficient estimation of conditional instrumental variables parameters. *The International Journal of Biostatistics*, 5:Article 22, 2009.

[Kyono, 2010] Trent Mamoru Kyono. Commentator: A front-end user-interface module for graphical and structural equation modeling. Technical Report R-364, University of California, Los Angeles, 2010.

[Lauritzen *et al.*, 1990] Steffen L. Lauritzen, A. Philip Dawid, Birgitte N. Larsen, and Hanns-Georg Leimer. Independence properties of directed markov fields. *Networks*, 20(5):491–505, 1990.

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

[Pearl, 2013] Judea Pearl. Reflections on heckman and pinto's 'Causal analysis after Haavelmo'. Technical Report R-420, University of California, Los Angeles, 2013.

[Shpitser and Pearl, 2006] Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semi-markovian causal models. In *AAAI*, pages 1219–1226, 2006.

[Textor *et al.*, 2011] Johannes Textor, Juliane Hardt, and Sven Knüppel. DAGitty: A graphical tool for analyzing causal diagrams, 2011.

[van der Zander *et al.*, 2014] Benito van der Zander, Maciej Liśkiewicz, and Johannes Textor. Constructing separators and adjustment sets in ancestral graphs. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence (UAI 2014)*, pages 907–916. AUAI Press, 2014.

[VanderWeele, 2009] Tyler J. VanderWeele. On the relative nature of overadjustment and unnecessary adjustment. *Epidemiology*, 20(4):496–499, Jul 2009.