Journal of Machine Learning Research 1 (2015) 1-53 Submitted ; Published 0

Learning Instrumental Variables with Structural and Non-Gaussianity Assumptions

Ricardo Silva RICARDO@STATS.UCL.AC.UK
Department of Statistical Science and Centre for Computational Statistics and Machine Learning
University College London, WC1E 6BT, UK
The Alan Turing Institute, 96 Euston Road, London NW1 2DB, UK

Shohei Shimizu SHOHEI-SHIMIZU@BIWAKO.SHIGA-U.AC.JP
The Center for Data Science Education and Research
Shiga University, 1-1-1 Banba Hikone, Shiga 522-8522, Japan
The Institute of Scientific and Industrial Research, Osaka University, Japan
RIKEN Center for Advanced Intelligence Project, Japan

Editor: TBA

Abstract

Learning a causal effect from observational data requires strong assumptions. One possibility is to use instrumental variables, which are typically justified by background knowledge. It is possible, under further assumptions, to discover whether a variable is structurally instrumental to a target causal effect $X \rightarrow Y$. However, the few existing approaches are lacking on how general these assumptions can be, and how to express possible equivalence classes of solutions. We present instrumental variable discovery methods that systematically characterize which set of causal effects can and cannot be discovered under local graphical criteria that define instrumental variables, without reconstructing full causal graphs. We also introduce the first methods to exploit non-Gaussianity assumptions, highlighting identifiability problems and solutions. Due to the difficulty of estimating such models from finite data, we investigate how to strengthen assumptions in order to make the statistical problem more manageable.

1. Contribution

Given observational data for a treatment variable X, an outcome Y, and a set of covariates $\mathbf{V}$ that precede X an Y causally, we present methods to estimate the causal effect of X on Y when hidden common causes between X and Y cannot be blocked by conditioning on observed variables. This complements approaches where hidden common causes can be blocked. We will assume that the model is linear, although this assumption can be relaxed to some extent, as discussed in the conclusion. Much of the contribution is theoretical, and intended to describe what to the best of our knowledge is the first graphical account of the limits of what can be discovered about instrumental variables from constraints in the observational distribution. We also discuss a pragmatic implementation of such ideas and clarify their practical limitations.

Consider a linear graphical causal model (Spirtes et al., 2000; Pearl, 2000) where, given a directed acyclic graph (DAG) $\mathcal{G}$, we define a joint distribution in terms of conditional

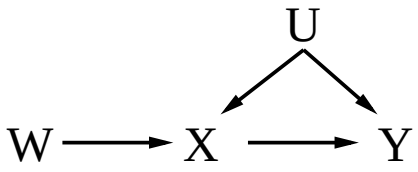


Figure 1: A graph illustrating a possible IV structure. X and Y have an unmeasured confounder U. W is an instrument as it is unconfounded with Y, has no direct effect on it, and causes X. In this paper, variables named "U" will denote hidden variables.

relationships between each variable V_i and its given *parents* in $\mathcal{G}$:

$$V_i = \sum_{V_j \in par_{\mathcal{G}}(i)} \lambda_{ij} V_j + e_i. \tag{1}$$

That is, each random variable V_i corresponds to a vertex in $\mathcal{G}$, where $par_{\mathcal{G}}(i)$ are the parents of V_i in $\mathcal{G}$ and e_i is an independent error term. Equation (1) is called a *structural equation* in the sense that it encodes a relationship that remains stable under a *perfect intervention* on other variables. Following the notation of Pearl (2000), we use the index "$do(V_k = v_k)$" to denote the regime under which some variable V_k is fixed to some level v_k by an external agent. If V_k is a parent of V_i, the *(differential) causal effect* of V_k on V_i is defined as:

$$\frac{\partial E[V_i \mid do(V_k = v_k)]}{\partial v_k} = \lambda_{ik}. \tag{2}$$

Each λ_{ik} will be referred to as a *structural coefficient*. Our goal is to estimate the differential causal effect of some treatment X on some outcome Y from observational data. If the common hidden causes of these two variables can be blocked by other observable variables, a formula such as the back-door adjustment of Pearl (2000) or the Prediction Algorithm of Spirtes et al. (2000) can be used to infer it. In general, unmeasured confounders of X and Y might remain unblocked.

When unmeasured confounding remains, and where it is reasonable to assume the linear structure (1), a possibility is to exploit an *instrumental variable* (or *instrument*, or *IV*) (Morgan and Winship, 2015): some observable variable W that is not an effect of either X and Y, it is unconfounded with Y, and has no direct effect on Y. Figure 1 illustrates one possible DAG containing an instrument.

Using $\sigma_{ab.\mathbf{s}}$ to represent the (conditional) covariance of two variables A and B (given set $\mathbf{S}$), the parameterization in (1) implies $\sigma_{wx} = \lambda_{xw}\sigma_{ww}$ and $\sigma_{wy} = \lambda_{yx}\lambda_{xw}\sigma_{ww}$. It follows that $\lambda_{yx} = \sigma_{wy}/\sigma_{wx}$. We can estimate σ_{wy} and σ_{wx} from observations, allowing for a consistent estimate of λ_{yx}. Notice that $\sigma_{wx} \neq 0$ is required. W in this case is called an instrumental variable for the causal relationship $X \to Y$.

It is not possible to test whether some observable variable is an IV from its joint distribution with X and Y alone. IV assumptions can nevertheless be falsified by exploiting constraints in the joint distribution of multiple observable variables (Chu et al., 2001; Brito

and Pearl, 2002; Kuroki and Cai, 2005), where such constraints are necessary but not sufficient to identify IVs and the corresponding causal effects. Our contribution are IV discovery algorithms for causal effect estimation. We characterize in which ways such algorithms can find the correct causal effect, and in which sense they will fail. We also introduce variations of the assumptions that are needed for practical reasons, complementing existing methods that rely on other sets of assumptions. Some of the ideas used in our methods are based on principles from causal discovery in linear non-Gaussian models (Shimizu et al., 2006).

The structure of the paper is as follows. In Section 2, we discuss the challenges of inferring causal effects when treatment and outcome are confounded by unobserved variables, and provide an overview of our approach. In Section 3, we discuss the theory behind two classes of testable constraints that can be detected from data. The resulting algorithm has several practical issues, and a more realistic alternative is provided in Section 4, which is then validated experimentally in Section 5. Other related approaches are discussed in Section 6.

2. Outline of Methodology: Learning Under Unmeasured Confounding

We assume that the system of interest is a linear DAG causal model with observable variables $\mathbf{V} \cup \{X, Y\}$. X and Y do not precede any element of $\mathbf{V}$. Y does not precede X. The goal is to estimate the differential causal effect of X on Y.

This task is common in applied sciences, as in many cases we have a particular causal effect $X \to Y$ to be estimated, and a set of covariates preceding X and Y is available. See Morgan and Winship (2015) for several examples. This is in contrast to the more familiar causal structure discovery tasks in the machine learning literature, where an equivalence class of a whole causal system is learned from data, and where some causal queries may or may not be identifiable (Spirtes et al., 2000). The focus here in on quantifying the strength of a particular causal effect λ_{yx}, as opposed to unveiling the directionalities and connections of a causal graph. This allows more focused algorithms that bypass a full graph estimation. This philosophy was exploited by Entner et al. (2012) in the task of finding possible sets of observable variables that can block the effect of any hidden common cause of X and Y.

However, the approach by Entner et al. will not provide a causal effect estimate if such a set does not exist. For instance, in Figure 1, if U is a latent variable, their algorithm will provide no answer concerning the causal effect of X on Y. *Our goal is to cover this scenario, which complements approaches that require unmeasured confounding to be blocked.* The methodological framework to accomplish this task is by discovering candidate instrumental variables without prior knowledge of the causal structure, besides the basic ordering assumptions about $\mathbf{V}, X$ and Y. The challenge is that we cannot guarantee which candidate instrumental variables are actual instruments without further assumptions.

We will make use of structural characterizations of causality using graphical models (Pearl, 2000; Spirtes et al., 2000). Prior exposure to causal graphical models is assumed, with key definitions summarized in Section 2.1. In Section 2.2, we outline the challenges and explain the general concept of *equivalence class of causal effects*, a concept adapted from Maathuis et al. (2009) to the instrumental variable case. Finally, in Section 2.3 we provide a road map of the steps used in our approach.

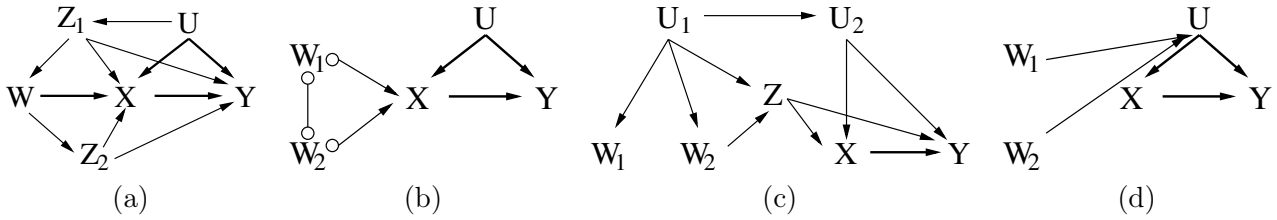


Figure 2: (a) Variable W is an instrument for the relation $X \to Y$ conditioning on $\{Z_1, Z_2\}$. (b) Both W_1 and W_2 are instruments. The circles at the end of the edges indicate that the direction between W_1 and W_2 is irrelevant, as well as the possibility of unmeasured confounding among $\{W_1, W_2, X\}$. (c) The typical covariance constraints ("tetrads") that are implied by instrumental variables also happen in the case where no instruments exist, implying that rank constraints in the covariance matrix are not enough information to discover IVs. (d) A case that is difficult even when considering information from non-Gaussian distributions.

2.1 Graphical Terminology

Directed acyclic graphs (DAGs) encode independencies among vertices, which correspond to independencies among random variables. This follows from the usual one-to-one relationship between vertices in a graph and random variables in a distribution. In a DAG, any two vertices $\{V_i, V_j\}$ can be connected by up to one *directed edge*. Given an edge $V_i \to V_j$, we say that V_i is the *tail* and V_j is the *head* of the edge. We also say V_i and V_j are *endpoints* of edge $V_i \to V_j$ where V_i is a *parent* of V_j and V_j is a *child* or V_i.

A *path* in a graph is a sequence of edges, where any two consecutive edges in the sequence share a common endpoint. For instance, $W \to X \leftarrow U \to Y$ is a path in Figure 1. A *collider* in a path is a vertex that is common to two consecutive edges such that this vertex is the head of both edges. For instance, X is a collider in the path given in the previous example, while U is not. A vertex V is an *endpoint of a path* P if it is one of the endpoints of the first or last edge E in the path, and V is not shared with the (possible) edge next to E in P. For instance, W and Y are the endpoints of $W \to X \leftarrow U \to Y$.

A path P is *between* vertices V_i and V_j if V_i and V_j are the endpoints of P. A *trek* is a path with no colliders. For instance, $W \to X \leftarrow U \to Y$ is not a trek, but $X \leftarrow U \to Y$ and $W \to X \to Y$ are. A trek P must have an unique *source*, a vertex in P that is not the head of any edge in P. A special case of a trek is a *directed path*, which is a trek where the source is one of the endpoints. For instance, $W \to X \to Y$ is a directed path between W and Y where W is the source. We also say this path is *from* the source (W) *into* the other endpoint (Y). The source V_i in a directed path P is an *ancestor* of all elements V_j in P, while V_j is a *descendant* of V_i. It is possible that $V_i = V_j$, so V_i is an ancestor and descendant of itself. A *non-directed path* is a path that is not directed.

A *back-door (path)* between V_i and V_j is a trek that is into V_i and V_j. For instance, $X \leftarrow U \to Y$ is a back-door between X and Y.

A vertex V is *active on a path* with respect to some vertex set $\mathbf{S}$ if it is either (i) a collider in this path and itself or one of its descendants is in $\mathbf{S}$; or (ii) not a collider and not

in **S**. A path is *active* if all of its vertices are active, *blocked* otherwise. The notion of active and blocked paths will be important in the sequel, as activation has implications on which variables can be considered to be an instrument with respect to which conditioning sets.

These definitions lead to the concept of *d-separation* (Pearl, 2000). A vertex V_i is d-separated from a vertex V_j given a set **S** if and only if every path between V_i and V_j is blocked by **S**. The interpretation of this definition is explained at length in the graphical modeling literature and we will assume familiarity with it. We say a probabilistic model $\mathcal{M}$ is *Markov* with respect to a DAG $\mathcal{G}$ if every d-separation in $\mathcal{G}$ corresponds to a conditional independence constraint in $\mathcal{M}$. A model $\mathcal{M}$ is *faithful* to $\mathcal{G}$ if every d-separation in $\mathcal{G}$ corresponds to a conditional independence constraint in $\mathcal{M}$ and vice-versa.

2.2 Scope and Fundamental Challenges

The identification of structural coefficients from given causal graphs is a classical problem in the structural equation modeling literature (Bollen, 1989). Much progress has been achieved on describing increasingly intricate combinations of structural features that lead to the identification of such coefficients (Brito and Pearl, 2002; Foygel et al., 2011; Chen et al., 2014). As these sophisticated criteria also lead to constraints which are hard to detect from data, we focus instead on the class of structures that corresponds to classical accounts of instrumental variables (Angrist and Pischke, 2009) as described by Brito and Pearl (2002).

Brito and Pearl's criteria are as follows. Given the causal graph $\mathcal{G}$ of a system that includes an edge $X \to Y$, a vertex W is a (conditional) instrument variable for $X \to Y$ given **Z** if and only if:

1. **Z** does not d-separate W from X in $\mathcal{G}$;
2. **Z** d-separates W from Y in the graph obtained by removing the edge $X \to Y$ from $\mathcal{G}$;
3. **Z** are non-descendants of X and Y in $\mathcal{G}$.

For the rest of the paper, we will call the above conditions the *graphical criteria for instrumental variable validity*, or simply "Graphical Criteria." Notice that the validity of a vertex as an IV is dependent on which set **Z** we condition on. That is, if in the corresponding causal graph we find some set **Z** that block all and only the paths relevant to the Graphical Criteria, then we can identify λ_{yx} as $\sigma_{wy.\mathbf{z}}/\sigma_{wx.\mathbf{z}}$. Figure 2(a) illustrates a case.

Unless strong background knowledge is available, the relevant structure needs to be learned from the data. The lack of an edge in Figure 1 is not testable (Chu et al., 2001), but in a situation such as Figure 2(b), the *simultaneous* lack of edges $W_1 \to Y$ and $W_2 \to Y$ has a testable implication, as in both cases we have $\lambda_{yx} = \sigma_{w_1y}/\sigma_{w_1x}$ and $\lambda_{yx} = \sigma_{w_2y}/\sigma_{w_2x}$. This leads to a *tetrad constraint*,

$$\sigma_{w_1y}\sigma_{w_2x} - \sigma_{w_1x}\sigma_{w_2y} = 0, \tag{3}$$

which can be tested using observable data. Unfortunately, the tetrad constraint is necessary, but not sufficient, to establish that both elements in this pair of variables are instrumental. As an example, consider Figure 2(c). It is not hard to show that $\sigma_{w_1y.z}\sigma_{w_2x.z} - \sigma_{w_1x.z}\sigma_{w_2y.z} = 0$. However, the Graphical Criteria for IVs is not satisfied as W_1 is not d-separated from Y given Z if we remove edge $X \to Y$. This is because path $W_1 \leftarrow U_1 \to U_2 \to Y$ is active.

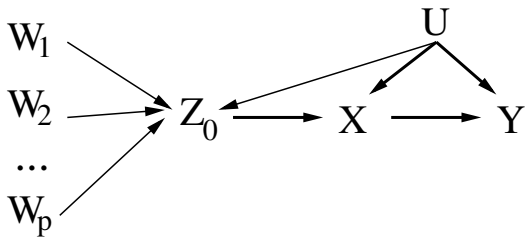


Figure 3: In this model, variables $W_1, \dots W_p$ are instrumental variables conditioning on the empty set. However, conditioning on Z_0 will introduce an active path from each W_i to Y via U, destroying their validity. This is particularly an issue for algorithms such as sisVIVE (Kang et al., 2015), where each variable is either deemed an IV or a conditioning variable.

Indeed, in this case λ_{yx} can be much different from $\sigma_{w_1 y.z}/\sigma_{w_1 x.z}$. A major component of our contribution is to characterize graphically in which ways the solution is not unique.

2.3 Roadmap

In general, the best we can do is to provide a *set* of candidate causal effects, one of which will be correct if at least two instrumental variables (under the same conditioning set) are present in the true graph. In this case, the set can be used, for instance, to provide lower and upper bounds on the causal effect. We will discuss our method in the context of other methods which require different assumptions about the existence of instruments, particularly about the *number* of instruments that exist in the true unknown graph.

That is, we propose algorithms that in theory return *equivalence classes of causal effects*. Such a class of algorithms will be sound in this more relaxed sense of returning a set of candidate effects that includes the true effect *if* there are instrumental variables, but will be incorrect otherwise.

As in (Cooper, 1997; Mani et al., 2006; Entner et al., 2012), we will not need to discover full graphs in order to identify the causal effect, but we will not also find all causal effects that are identifiable from faithfulness assumptions. In particular, we will consider in which sense our algorithms are *complete*: that is, if there are instrumental variables satisfying the Graphical Criteria, we will characterize under which conditions we will find them. Computational considerations are relevant here.

In the Graphical Criteria, the challenging condition is the second, as the first is easily testable by faithfulness and the third is given by assumption. Another way of phrasing condition 2 is:

2a. there is no active (with respect to $\mathbf{Z}$) non-directed path between W and Y that does not include X (that is, no active back-door path nor any active path that includes a collider);

2b. there is no active directed path from W to Y that does not include X.

Algorithm 1 IV-BY-MAJORITY$_\infty$

Input: set of random variables $\mathbf{V} \cup \{X, Y\}$
Output: the causal effect of X on Y, or a value (NA) indicating lack of knowledge
for each $W_i \in \mathbf{V}$ **do**
  $\mathbf{Z}_i \leftarrow \mathbf{V} \backslash \{W_i\}$
  $\beta_i \leftarrow \sigma_{w_i y.\mathbf{z}_i} / \sigma_{w_i x.\mathbf{z}_i}$
end for
if more than half of set $\{\beta_i\}$ is equal to the same value β **then**
  return β
end if
return NA

In the next Section, we introduce algorithms that return an equivalence class of causal effects using tetrad constraints, which are complemented by non-Gaussianity assumptions. Motivated by simplicity of presentation, all approaches assume the distribution of the population is known and that computational resources are unbounded. We do not claim these algorithms are practical – the goal is to use them as a theoretical basis to choose and justify stronger assumptions that achieve practical learning. In Section 4, we discuss practical methods for learning from data and the computational and identification compromises we adopt.

3. From Structural Constraints to Instruments and Causal Effects

Consider Algorithm 1 as a method for learning causal effects given the distribution of the population (hence, the "∞" symbol in the name of the algorithm, indicating that this is equivalent to having infinite sample sizes). If W_i is an IV conditioned on $\mathbf{Z}_i \equiv \mathbf{V} \backslash W_i$, then $\beta_i \equiv \sigma_{w_i y.\mathbf{z}_i} / \sigma_{w_i x.\mathbf{z}_i} = \lambda_{yx}$, the true causal effect. Without knowing whether W_i is a conditional IV with respect to $\mathbf{Z}_i$, we cannot make claims about the causal effect.

However, if more than half of elements $W_i \in \mathbf{V}$ are conditional IVs given the respective $\mathbf{Z}_i$, then it follows that more than half of the elements in set $\{\beta_i\}$ will be equal, and equal to λ_{yx}. This is the core assumption introduced by Kang et al. (2015). It sidesteps the problems introduced by models such as the one in Figure 2(c) by assuming that at least half of $\mathbf{V}$ are "valid" IVs. That is, we can partition $\mathbf{V}$ into two sets, $\mathbf{V} = \mathbf{W} \cup \mathbf{Z}$, such that each $W \in \mathbf{W}$ is a conditional IV given $\mathbf{Z} \cup \mathbf{W} \backslash \{W\}$. This is done without knowledge of which variables are valid and which are not. As discussed by Kang et al. (2015), there are situations where this assumption is plausible, or at least weaker than in standard approaches, as in some genetic studies where $\mathbf{V}$ are genetic features of a cell and X, Y are phenotypes. The resulting algorithm (sisVIVE, "some invalid, some valid IV estimator") is very different from Algorithm 1 (as it has to deal with estimates of the covariance matrix in a statistically efficient way, and it never returns NA as it assumes there is always a majority), but still very elegant and computationally efficient.

However, this assumption can be false even when nearly all of $\mathbf{V}$ are possibly valid instruments. Consider Figure 3 where we have an arbitrary number of IVs $W_1, \ldots, W_p$ that are valid by conditioning on the empty set. *None* of them are valid by conditioning on Z_0.

Algorithm 2 IV-TETRAD$_\infty$

1: **Input:** set of random variables $\mathbf{V} \cup \{X, Y\}$
2: **Output:** $\mathcal{C}$, a set of candidate differential causal effects of X on Y
3: Initialize $\mathcal{C} \leftarrow \emptyset$
4: **for** each pair $\{W_i, W_j\} \subseteq \mathbf{V}$ **do**
5: **for** every set $\mathbf{Z} \subseteq \mathbf{V} \backslash \{W_i, W_j\}$ **do**
6: **if** $\sigma_{w_i x.\mathbf{z}} = 0$ **or** $\sigma_{w_j x.\mathbf{z}} = 0$ **then**
7: **next**
8: **end if**
9: **if** $\sigma_{w_i x.\mathbf{z}} \sigma_{w_j y.\mathbf{z}} \neq \sigma_{w_i y.\mathbf{z}} \sigma_{w_j x.\mathbf{z}}$ **then**
10: **next**
11: **end if**
12: $\mathcal{C} \leftarrow \mathcal{C} \cup \{\sigma_{w_i y.\mathbf{z}} / \sigma_{w_i x.\mathbf{z}}\}$
13: **end for**
14: **end for**
15: **return** $\mathcal{C}$

In this situation, IV-BY-MAJORITY$_\infty$ will return NA and sisVIVE may perform badly. We will characterize how this happens by using faithfulness.

3.1 Structural Signatures of Tetrad Constraints and the Graphical Criteria

Consider Algorithm 2 as a method for learning causal effects, disregarding for now its computational complexity. The idea is to find triplets $(W_i, W_j, \mathbf{Z})$ such that $(W_i, \mathbf{Z})$ and $(W_j, \mathbf{Z})$ both satisfy the Graphical Criteria. Under the assumption of *linear faithfulness* (Spirtes et al., 2000), Line 6 of the algorithm is equivalent to item 1 of the Graphical Criteria. To characterize what Algorithm 2 can say about item 2 of the Graphical Criteria, we will need some more advanced graphical definitions.

3.1.1 t-Separation and Rank Constraints

In what follows, we will use the notion of *t-separation* (Sullivant et al., 2010). Recall the basic definition of a trek from Section 2.1. Another way of describing a trek T is by ordering its endpoints (such that a trek is "from" V_i "to" V_j), implying an ordered pair of (possibly empty) directed paths $(P_1; P_2)$ where: P_1 has *sink* (vertex without children in T) V_i; P_2 has sink V_j; and P_1, P_2 have the same *source* (vertex in T without parents in T).

Definition (t-separation) The ordered pair of vertex sets $(\mathbf{C}_I; \mathbf{C}_J)$ *t-separates* vertex set $\mathbf{V_I}$ from vertex set $\mathbf{V_J}$ if, for every trek $(P_1; P_2)$ from a vertex in $\mathbf{V_I}$ to a vertex in $\mathbf{V_J}$, either P_1 contains a vertex in $\mathbf{C}_I$ or P_2 contains a vertex in $\mathbf{C}_J$.

See Spirtes (2013) and Sullivant et al. (2010) for a generalization of this notion and further examples. As the definition is somewhat complex, in the next Section we provide some basic examples of this concept.

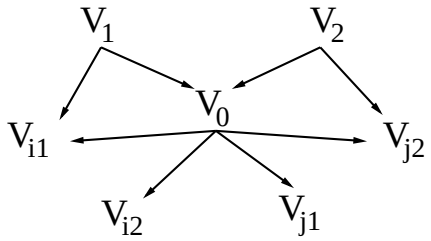


Figure 4: This structure implies testable (rank) constraints on a subset of its variables, meaning some variables can be unobserved.

Like d-separation, t-separation is relevant if it implies testable implications in the joint distribution of the observable variables that under some assumptions will decrease the set of possible graphical structures that could have generated the data (Spirtes et al., 2000). While we will see that t-separation can also imply independence constraints among observed variables, this would be of limited interest, as d-separation can also imply such constraints. The importance of t-separation is the possibility of implying testable consequence of d-separations *given unobserved variables*. This is done under the assumption of linearity, which will lead to rank constraints.

A *rank constraint* in a matrix M is any constraint of the type $rank(M) \leq r$, where r is some constant (Spirtes, 2013). If M is the cross-covariance submatrix given by variables $\{V_i, V_j\}$ indexing the rows, and by $\{V_k, V_l\}$ indexing the columns, then the rank constraint $rank(M) \leq 1$ implies $\sigma_{ik}\sigma_{jl} - \sigma_{il}\sigma_{jk} = 0$, the determinant of M.

Let $\Sigma_{\mathbf{AB}}$ be the cross-covariance matrix of set $\mathbf{A}$ (rows) and set $\mathbf{B}$ (columns). The DAG Trek Separation Theorem of Sullivant et al. (2010) says:

Theorem 1 ***(Trek Separation for DAGs)*** *Let $\mathcal{G}$ be a DAG with vertex set $\mathbf{V}$. Let $\mathbf{A}$ and $\mathbf{B}$ be subsets of $\mathbf{V}$. We have $rank(\Sigma_{\mathbf{AB}}) \leq r$ in all linear structural equation models with graph $\mathcal{G}$ if and only if there exist subsets $\mathbf{C_A}$ and $\mathbf{C_B}$ of $\mathbf{V}$ with $|\mathbf{C_A}| + |\mathbf{C_B}| \leq r$ such that $(\mathbf{C_A}; \mathbf{C_B})$ t-separates $\mathbf{A}$ from $\mathbf{B}$.*

To jump from (testable) rank constraints to (unobservable) structural constraints in $\mathcal{G}$, we assume our model distribution P is *linearly rank-faithful* to a DAG $\mathcal{G}$ (Spirtes, 2013): that is, every rank-constraint holding on a covariance (sub)matrix derived from P is entailed by every linear structural model Markov with respect to $\mathcal{G}$. *Linear faithfulness*, the assumption that vanishing partial correlations hold in the distribution if and only if a corresponding d-separation also holds in $\mathcal{G}$ (Spirtes et al., 2000), is a special case of rank faithfulness, as t-separation implies d-separation (Sullivant et al., 2010).

3.1.2 Examples

Unlike d-separation, t-separation is defined by a pair of conditioning sets: when we say that V_i is t-separated from V_j given some ordered pair $(\mathbf{C}_I; \mathbf{C}_J)$, the order of the sets in the conditioning pair matters. Moreover, these two sets do not need to be disjoint.

Consider Figure 4. Vertex V_0 does not d-separate V_{i1} from V_{j2}. However, *pair* $(V_0; V_0)$ t-separates V_{i1} from V_{j2}. To see that, let us list all three treks between V_{i1} and V_{j2}:

- $(P_1^{[1]}; P_2^{[1]}) \equiv (V_{i1} \leftarrow V_1; V_1 \rightarrow V_0 \rightarrow V_{j2})$
- $(P_1^{[2]}; P_2^{[2]}) \equiv (V_{i1} \leftarrow V_0; V_0 \rightarrow V_{j2})$
- $(P_1^{[3]}; P_2^{[3]}) \equiv (V_{i1} \leftarrow V_0 \leftarrow V_2; V_2 \rightarrow V_{j2})$

In the first trek, V_0 is contained in $P_2^{[1]}$; in the second case, both paths in the trek satisfy the criterion of containing V_0; in the final trek, $P_1^{[3]}$ plays this role. Notice that $(V_0; \emptyset)$ does not t-separate V_{i1} from V_{j2} (first trek remains "unblocked") and neither does $(\emptyset; V_0)$ (third trek remains "unblocked").

That V_{i1} is t-separated from V_{j2} given $(V_0; V_0)$, however, brings out no useful implication concerning the cross-covariance matrix $\Sigma_{\mathbf{AB}}$ of $(\mathbf{A} \equiv \{V_{i1}\}, \mathbf{B} \equiv \{V_{j2}\})$: all it says is that $rank(\Sigma_{\{V_{i1}\}\{V_{j2}\}}) \leq 2$, which is a vacuous claim. We may attempt to introduce V_0 in the two sets $\mathbf{A}$ and $\mathbf{B}$, assuming V_0 is observable, as this is compatible with the notion that V_0 can t-separate itself from itself. Unfortunately, there is nothing to be gained, as $rank(\Sigma_{\{V_{i1}V_0\}\{V_{j2}V_0\}}) \leq 2$, which is again a vacuous claim.

Pair $(\emptyset; V_0)$ does t-separate V_{i1} and V_{i2} with a testable implication $rank(\Sigma_{\{V_{i1}V_0\}\{V_{i2}V_0\}}) \leq 1$, but this does not provide anything useful to a structure learning algorithm, as d-separation of V_{i1} and V_{i2} given V_0 can be tested directly instead.

Consider now $\mathbf{A} \equiv \{V_{i1}, V_{i2}\}$, $\mathbf{B} \equiv \{V_{j1}, V_{j2}\}$. To be useful, we need a conditioning pair $(\mathbf{C_A}; \mathbf{C_B})$ where $|\mathbf{C_A}| + |\mathbf{C_B}| \leq 1$. Following the previous examples, it should be clear that no such a pair will exist out of the given variables. However, if V_1 is observable we could de-activate the path $V_{i1} \leftarrow V_1 \rightarrow V_0$ and consider the *partial* cross-covariance matrix of $\{V_{i1}, V_{i2}\}$ against $\{V_{j1}, V_{j2}\}$ given V_1. The Trek Separation Theorem for DAGs, however, says nothing explicit about conditional cross-covariances. The independence model given by conditioning on a variable is sometimes a DAG itself, but this is not true in general (Richardson and Spirtes, 2002). Sullivant et al. (2010) present versions of the theorem for some classes of DAGs under conditioning, but not in a completely general way.

Fortunately, for our purposes this can be easily dealt with: simply introduce the conditioning variables in both sets $\mathbf{A}$ and $\mathbf{B}$. In this example, define $\mathbf{A} \equiv \{V_{i1}, V_{i2}, V_1\}$ and $\mathbf{B} \equiv \{V_{j1}, V_{j2}, V_1\}$. Then it is not hard to verify that $(V_0; V_1)$ t-separates $\mathbf{A}$ from $\mathbf{B}$ (recall that V_1 t-separates itself from itself). Cross-covariance $\Sigma_{\mathbf{AB}}$ is rank-deficient, which means its determinant is zero. Because its determinant can be written as $\sigma_{11}(\sigma_{i1j1.1}\sigma_{i2j2.1} - \sigma_{i1j2.1}\sigma_{i2j1.1})$ and we assume $\sigma_{11} \neq 0$, the (conditional) tetrad constraint will hold.

This means that without observing V_0 we can make claims about the conditional structure between $\{V_{i1}, V_{i2}\}$ and $\{V_{j1}, V_{j2}\}$. In general, given some $\mathbf{Z}$, and by assuming rank-faithfulness (and that $\Sigma_{\mathbf{ZZ}}$ is full rank), we can claim that there is a pair $(\mathbf{C_A}; \mathbf{C_B})$, $|\mathbf{C_A}| + |\mathbf{C_B}| \leq |\mathbf{Z}| + 1$, which t-separates $\{V_{i1}, V_{i2}\}$ from $\{V_{j1}, V_{j2}\}$.

Another example of t-separation can be obtained from Figure 2(b). Here, $\mathbf{C}_I = \emptyset$ and $\mathbf{C}_J = \{X\}$; $\mathbf{V_I} = \{W_1, W_2\}, \mathbf{V_J} = \{X, Y\}$. In Figure 2(c), $\mathbf{C}_I = \emptyset$ and $\mathbf{C}_J = \{U_1, Z\}$; $\mathbf{V_I} = \{Z, W_1, W_2\}, \mathbf{V_J} = \{Z, X, Y\}$.

3.2 Analysis of Algorithm 2

The algorithm applies rank constraints to quartets $\{W_i, W_j\} \times \{X, Y\}$ where X is the treatment and Y is the outcome. We need to characterize which structures are compatible with the Trek Separation Theorem and, among those, which satisfy the Graphical Criteria. In Line 9 of Algorithm 2, the conditional tetrad constraint is equivalent to $rank(\Sigma_{\mathbf{AB}}) \leq |\mathbf{Z}| + 1$ for $\mathbf{A} = \{W_i, W_j\} \cup \mathbf{Z}$ and $\mathbf{B} = \{X, Y\} \cup \mathbf{Z}$. If the constraint holds, it follows that there is some pair $(\mathbf{C_A}; \mathbf{C_B})$ that t-separates $\{W_i, W_j\} \cup \mathbf{Z}$ from $\{X, Y\} \cup \mathbf{Z}$. It is required that $\mathbf{Z} \subseteq \mathbf{C_A} \cup \mathbf{C_B}$, since $\mathbf{Z}$ is contained in both $\mathbf{A}$ and $\mathbf{B}$.

For simplicity, we assume that there is an edge $X \to Y$ corresponding to a non-zero coefficient λ_{yx}. Otherwise, if W_i satisfies the Graphical Criteria for (X, Y) given $\mathbf{Z}$, then under unmeasured confounding between treatment and outcome we have that the lack of an edge $X \to Y$ implies $W_i \perp\!\!\!\perp Y \mid \mathbf{Z}$ and $W_i \not\perp\!\!\!\perp Y \mid \mathbf{Z} \cup \{X\}$. This corresponds to the collider orientation rule of the FCI algorithm (Spirtes et al., 2000) that implies $\lambda_{yx} = 0$. If on top of the lack of edge $X \to Y$ we have that there is no unblocked unmeasured confounding between X and Y, then $X \perp\!\!\!\perp Y \mid \mathbf{Z}$, which again by faithfulness will allow us to infer $\lambda_{yx} = 0$ (Spirtes et al., 2000). Algorithm 2 is assumed to be invoked only if the FCI algorithm and the method of Entner et al. (2012) do not provide any results.

If we assume there is indeed an edge $X \to Y$ in the true graph, and since $\mathbf{Z}$ does not d-separate $\{W_i, W_j\}$ from X (faithfulness and Line 6 of Algorithm 2), then $\mathbf{Z}$ does not d-separate $\{W_i, W_j\}$ from $\{X, Y\}$. This means that if we assume that X is a direct cause of Y, then $\mathbf{C_A} \cup \mathbf{C_B} \neq \mathbf{Z}$, or otherwise $\mathbf{Z}$ would d-separate $\{W_i, W_j\}$ from $\{X, Y\}$. Since by Theorem 1 we have $|\mathbf{C_A}| + |\mathbf{C_B}| \leq |\mathbf{Z}| + 1$, then there is exactly one element left in $\mathbf{C_A} \cup \mathbf{C_B}$ to explain this t-separation. Moreover, this element cannot appear in both $\mathbf{C_A}$ and $\mathbf{C_B}$, or otherwise $|\mathbf{C_A}| + |\mathbf{C_B}| > |\mathbf{Z}| + 1$.

We call this element a *conditional choke point* for pairs $\{W_i, W_j\} \times \{X, Y\}$ given $\mathbf{Z}$. A conditional choke point can be a "left" or "right" choke point:

Definition (left conditional choke point) A vertex Z_0 is a left conditional choke point for pairs $\{W_i, W_j\} \times \{X, Y\}$ given $\mathbf{Z}$ in some causal graph $\mathcal{G}$ if

1. some $(\mathbf{C_A}; \mathbf{C_B})$ t-separates $\{W_i, W_j\}$ from $\{X, Y\}$ such that $\mathbf{C_A} \cup \mathbf{C_B} = \mathbf{Z} \cup \{Z_0\}$ and $\mathbf{C_A} \cap \mathbf{C_B} = \emptyset$;

2. for any trek $(P_1; P_2)$ from a vertex in $\{W_i, W_j\}$ to a vertex in $\{X, Y\}$ that contains no member of $\mathbf{Z}$, P_1 contains Z_0;

3. there is at least one trek $(P_1; P_2)$ from each vertex in $\{W_i, W_j\}$ to each vertex in $\{X, Y\}$ that contains no member of $\mathbf{Z}$;

A "right conditional choke point" follows an equivalent definition with respect to P_2. The literature has characterizations of *unconditional* choke points (Shafer et al., 1993; Sullivant et al., 2010), relating them to unconditional tetrad constraints. To the best of our knowledge, this is the first time that conditional choke points are explicitly defined and used.

Conditional choke points can be unintuitive. For instance, Z_0 is a (left and right) conditional choke point in Figure 3 for each $\{W_i, W_j\} \times \{X, Y\}$ given the empty set, since $(Z_0; Z_0)$ t-separates (say) W_i from Y. The conditional tetrad constraint holds, but once again the

Graphical Criteria does not as W_1 is not d-separated from Y given Z_0 in the graph that modifies Figure 3 by removing edge $X \to Y$.

This definition leads to the following characterization of Algorithm 2:

Theorem 2 *Let $\mathcal{C}$ be the outcome of Algorithm 2. Let $\beta_i \in \mathcal{C}$, and let $(W_i, W_j, \mathbf{Z})$ be the triplet that generated β_i according to passing Step 9. Then there exists a conditional choke point Z_0 corresponding to $\{W_i, W_j\} \times \{X, Y\}$ given $\mathbf{Z}$. Assuming rank faithfulness and that the edge $X \to Y$ exists in the causal graph, then*

1. *$\beta_i \neq \lambda_{yx}$ only if there is an active directed path from Z_0 to Y that does not include X, or there is at least one active path between W_i and Y that includes a collider;*

2. *if there is a directed path from Z_0 to Y that does not include X, then W_i does not satisfy the Graphical Criteria with respect to $X \to Y$ given $\mathbf{Z}$.*

Proof of Theorem 2. The existence of Z_0 follows from the discussion above. The other results are an almost immediate consequence of the existence of Z_0. Consider first the case where any directed path from Z_0 to Y includes X (including the case $Z_0 = X$). In this case, any treks between W_i and Y will either be blocked by $\mathbf{Z}$ or by X. If we remove edge $X \to Y$ from the graph, there will be no treks from W_i to Y that are unblocked given $\mathbf{Z}$, as by assumption all treks between W_i and X are into X and cannot be extended into a trek into Y without a directed path from X to Y. If there is also no active path between W_i and Y that includes a collider, then item 2 of the Graphical Criteria will be satisfied and $\beta_i = \lambda_{yx}$. In the case where there is a directed path from Z_0 to Y that does not include X, it can be combined with any trek between W_i and Z_0 that contains no member of $\mathbf{Z}$ (of which at least one exists) to form a trek from W_i to Y that contains no member of $\mathbf{Z} \cup \{X\}$. It follows that W_i cannot satisfy item 2 of the Graphical Criteria. □

Even if all elements of $\mathcal{C}$ are equal, it does not mean the value found is λ_{yx}. It is impossible to distinguish, based on covariance information, the graph in Figure 2(b) from a graph which includes an intermediate latent variable Z_0 that is a common parent of X and Y and blocks all directed paths from W_1 and W_2 to X.

This motivates the following definition:

Definition (locally covariance equivalent causal effect) We say that two ratios $\beta_{i(\mathbf{z})} \equiv \sigma_{w_i y.\mathbf{Z}}/\sigma_{w_i x.\mathbf{Z}}$ and $\beta_{j(\mathbf{z}')} \equiv \sigma_{w_j y.\mathbf{Z}'}/\sigma_{w_j x.\mathbf{Z}'}$ are *locally covariance equivalent causal effects* if there is some pair $\{W_k, W_l\}$, $W_i \neq W_k$, $W_j \neq W_l$, where $\beta_{i(\mathbf{z})} = \beta_{k(\mathbf{z})}$ and $\beta_{j(\mathbf{z}')} = \beta_{l(\mathbf{z}')}$.

IV-TETRAD$_\infty$ is sound and complete in the sense it returns all and only locally covariance equivalent causal effects, a class that is of interest if one is willing to adopt the following two criteria in any algorithm that learns causal effects λ_{yx} based on instrumental variables:

1. only substructures where two or more variables follow the Graphical Criteria conditioned on a common set $\mathbf{Z} \subset \mathbf{V}$ can be used to select conditional instrumental variables;

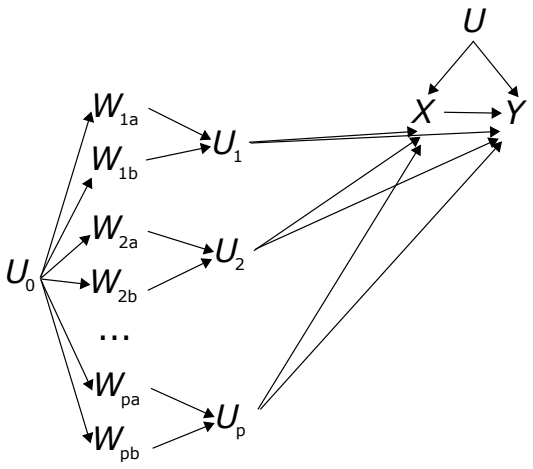


Figure 5: Vertices named $U_\star$ are latent variables, with X being the treatment and Y being the outcome. When given as input to algorithms for finding locally covariance equivalent solutions, the result will include a (potentially different) causal effect candidate for each of the pairwise groupings $\{W_{ia}, W_{ib}\}$, $i = 1, 2, \ldots, p$, where $p = |\mathbf{V}|/2$.

2. only covariance information is used.

The requirement for covariance information allows us to use only second moments, which is relatively statistically efficient. It will be relaxed in the next Section. The need for "two or more" variables in condition 1 is necessary, as the Graphical Criteria provides no testable implications for a single instrument, as discussed in the introduction. It is also desirable, as it consider variables that are jointly instrumental by conditioning on the same set $\mathbf{Z}$. To drop this criterion is to require a search space where different conditioning sets are required for each IV, resulting in an algorithm that effectively reconstructs a "non-local" graphical substructure that will require the identification of structural coefficients other than λ_{yx}. This is what is done, for instance, in one of the first published extensions of the Graphical Criteria (Brito and Pearl, 2002). However, identification is an easier problem than causal learning. Inferring structure from data is statistically challenging, and we claim that restricting a method to the desiderata above provides a baseline where stronger assumptions can be exchanged for greater specificity of the solutions returned.

The need to report an equivalence class is illustrated by the following result.

Proposition 3 *There exist problems where the output of* IV-TETRAD$_\infty$ *will contain* $\mathcal{O}(|\mathbf{V}|)$ *different elements.*

Proof of Proposition 3. This is illustrated by Figure 5: given $\mathbf{V} = \{W_{1a}, W_{1b}, W_{2a}, \ldots, W_{pb}\}$, $p = |\mathbf{V}|/2$, the structure shown in the Figure will generate solutions for all pairs $\{W_{ia}, W_{ib}\} \times \{X, Y\}$ with corresponding conditioning set $\mathbf{Z}_i \equiv \mathbf{V} \backslash \{W_{ia}, W_{ib}\}$. For parameters generated randomly and independently, it is clear each resulting λ_i will be unique. The corresponding

choke point in each case is U_i, and none of the returned causal effects will in general be correct. □

The output of IV-TETRAD$_\infty$ is an *equivalence class of causal effects*, differing from the usual output of causal discovery algorithms such as the PC algorithm (Spirtes et al., 2000): the PC algorithm returns a set of graphs of which some graphical features might be the same (for instance, they all share the same directed edge), and where some causal effects might be identified based on the common features. That is, the PC algorithm either gives (in the limit of infinite data) the correct answer, or it answers "I don't know." The same is true of the method introduced in Entner et al. (2012).

In contrast, IV-TETRAD$_\infty$ may return an empty set of solutions, but at the same time *all* solutions in a non-empty output set $\mathcal{C}$ may be wrong. To avoid this complication, one possibility is to assume *at least one* solution exists, that is, there exists at least one pair of IVs conditioned on the same set $\mathbf{Z}$. The resulting output $\mathcal{C}$ will automatically provide, among other summaries, an upper bound and a lower bound on the differential causal effect λ_{yx}. Assuming that one solution exists is a much weaker assumption than the one in IV-BY-MAJORITY/sisVIVE. However, it does require rank faithfulness and can only provide bounds and other summaries of an equivalence class. If in the application of the algorithm we have that all elements of $\mathcal{C}$ are the same, then from the assumption that there is at least one valid tuple it follows that we found the true causal effect. A discussion of what further assumptions are necessary in practical learning is postponed to Section 4.

3.3 Exploiting Non-Gaussianity Assumptions

This Section introduces a variant of IV-TETRAD$_\infty$ that can verify the validity of item 2a of the Graphical Criteria, as discussed in Section 2.3. We assume our causal model is a *LiNGAM model*, a linear structural equation model with independent, non-Gaussian error terms, which may include latent variables (Shimizu et al., 2006). We call this Algorithm IV-TETRAD$^+_\infty$, which is shown in Algorithm 3. We will also discuss the difficulties posed by 2b even under the assumption of non-Gaussianity. The motivation is again to reduce the size of the equivalence class by trading it off with the addition of more assumptions (non-Gaussianity) and a weakening of completeness (some solutions might be missed, as we will characterize).

In Algorithm 3, function resProj$(V, \mathbf{S})$ is a function that returns the residual of the least-squares projection of V into the set of random variables $\mathbf{S}$ rearranged as a column vector,

$$\text{resProj}(V, \mathbf{S}) \equiv V - \mathbf{S} \times \mathbb{E}[\mathbf{S}^\intercal \mathbf{S}]^{-1} \mathbb{E}[\mathbf{S}^\intercal V].$$

3.3.1 Main Result

The validity of this variant holds "almost everywhere," in the sense it holds for all but a (Lebesgue) measure zero subset of the set of possible structural coefficients $\Lambda_\mathcal{G} = \{\lambda_{ij} \mid V_j \in par_\mathcal{G}(i)\}$. The motivation for this concept is analogous to the different variations of faithfulness, see the discussion on generic identifiability by Foygel et al. (2011) and Sullivant et al. (2010).

Algorithm 3 IV-TETRAD$^{+}_{\infty}$

1: **Input:** set of zero-mean random variables $\mathbf{V} \cup \{X, Y\}$
2: **Output:** $\mathcal{C}$, a set of candidate differential causal effects of X on Y
3: Initialize $\mathcal{C} \leftarrow \emptyset$
4: **for** each pair $\{W_i, W_j\} \subseteq \mathbf{V}$ **do**
5: **for** every set $\mathbf{Z} \subseteq \mathbf{V} \backslash \{W_i, W_j\}$ **do**
6: **if** $\sigma_{w_i x.\mathbf{z}} = 0$ **or** $\sigma_{w_j x.\mathbf{z}} = 0$ **then**
7: **next**
8: **end if**
9: **if** $\sigma_{w_i x.\mathbf{z}} \sigma_{w_j y.\mathbf{z}} \neq \sigma_{w_i y.\mathbf{z}} \sigma_{w_j x.\mathbf{z}}$ **then**
10: **next**
11: **end if**
12: $r_{W_i} \leftarrow \text{resProj}(W_i, \mathbf{Z} \cup \{W_j\})$
13: $r_{W_j} \leftarrow \text{resProj}(W_j, \mathbf{Z} \cup \{W_i\}\})$
14: $r_Y \leftarrow \text{resProj}(Y, \mathbf{Z} \cup \{W_i, W_j\})$
15: **if** $r_{W_i} \perp\!\!\!\perp r_Y$ **and** $r_{W_j} \perp\!\!\!\perp r_Y$ **then**
16: $\mathcal{C} \leftarrow \mathcal{C} \cup \{\sigma_{w_i y.\mathbf{z}} / \sigma_{w_i x.\mathbf{z}}\}$
17: **end if**
18: **end for**
19: **end for**
20: **return** $\mathcal{C}$

The main result of this section is the following:

Theorem 4 *Let $\mathbf{V} \cup \{Y\}$ be a subset of the variables in a zero-mean LiNGAM model, where Y has no descendants. Let $V_i \in \mathbf{V}$ and $\mathbf{Z} \subseteq \mathbf{V} \backslash \{V_i\}$, and let $r_i \equiv V_i - \mathbf{a}^{\intercal}\mathbf{Z}$ be the residual of the least-squares regression of V_i on $\mathbf{Z}$, with $\mathbf{a}$ being the corresponding least-squares coefficients. Analogously, let $r_y \equiv Y - b_i V_i - \mathbf{b}^{\intercal}\mathbf{Z}$ be the residual of the corresponding least-squares regression. Then, almost everywhere, $r_i \perp\!\!\!\perp r_y$ if and only if there are no active (with respect to $\mathbf{Z}$) non-directed paths between V_i and Y.*

The proof is given in the Appendix. This generalizes the main result of Entner et al. (2012), which considers the case where $\mathbf{Z}$ contains no descendant of V_i.

Even under non-Gaussianity, there are still choke points Z_0 that can induce error. However, the relationship between W_i and the choke point is constrained: no active non-directed paths can exist between W_i and Z_0. We call this type of choke point a "downstream conditional choke point," illustrated by vertex U in Figure 2(d). The name "downstream" denotes that this point is a descendant of $\{W_i, W_j\}$, and has no active back-door paths with them. By further restricting the set of possible choke points, we refine the explanation of possible disparities obtained by Algorithm 3. By knowing there is a *single* choke point per pair, which is *unconfounded* with the candidate IVs, *and* which lies on all unblocked directed paths from W_i to X, more sophisticated algorithms, combined with background knowledge, can be constructed that exploit this piece of information[1].

1. Notice that Steps 12-14 in IV-TETRAD$^{+}_{\infty}$ suggest that W_j should be included in the conditioning set for W_i and vice-versa. Including W_j in the conditioning set for W_i is acceptable, given that if W_i and

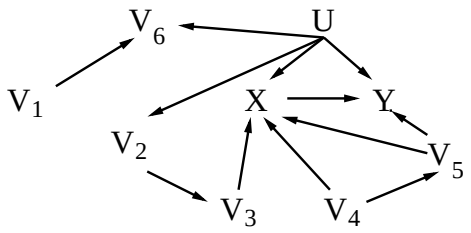


Figure 6: In this system, $\mathbf{W} \equiv \{V_1, V_2\}$ satisfies the conditional tetrad constraints $\{V_1, V_2\} \times \{X, Y\}$ given $\mathbf{Z} \equiv \{V_3, V_6\}$. The same is true for $\mathbf{W}' \equiv \{V_3, V_4\}$ for $\mathbf{Z}' \equiv \{V_2, V_5\}$. Only $\mathbf{W}'$ represents a set of (conditionally) valid IVs. The non-Gaussianity criteria of IV-TETRAD$^+_\infty$ helps to rule out the former.

3.3.2 Implications to Completeness

The following result is a direct consequence of the assumptions and the exhaustive search done by IV-TETRAD$^+_\infty$. The proof of this result follows immediately from Theorem 4:

Proposition 5 *If there is a pair of observable variables $\{W_i, W_j\}$ which are IVs conditioned on some $\mathbf{Z}$ according to the Graphical Criteria, and some $W \in \{W_i, W_j\}$ has an active non-directed path with X given $\mathbf{Z}$, then Algorithm 3 will not include $\{W_i, W_j\}$ in its output.*

The implication of this is as follows. A positive property of IV-TETRAD$^+_\infty$ is that it will exclude from the equivalence class some candidate IVs that do not obey the Graphical Criteria but which are returned by IV-TETRAD$_\infty$. See Figure 6 for an example. However, this algorithm is only complete, in the sense of finding the correct causal effect if the true model contains at least one pair of IVs conditioned on the same set, if we make further assumptions. In particular, suppose that for at least one pair $\{W_i, W_j\}$ the conditioning set $\mathbf{Z}$ also blocks active back-door/collider paths into X. This means, for example, that the algorithm will not find answers in models where W_i and X have common causes that cannot be blocked, even if W is a valid IV by not having common causes with Y. For example, W is a valid IV in the model with paths $W \to X \to Y$, $W \leftarrow U_1 \to X \leftarrow U_2 \to Y$, but W will be discarded due to the back-door path between W and Y that is unblocked by not conditioning on X. This trade-off can be implemented in different ways in practical algorithm, as we shall see next.

4. Practical Learning from Data

Testing tetrad constraints from data is difficult in practice, particularly without assuming Gaussianity and under conditioning. In order to search for candidate IVs, we will first adopt the stronger (but falsifiable) assumption that, if any set of valid conditional IVs exist, then

W_j each satisfy the Graphical Criteria with respect to $X \to Y$ given $\mathbf{Z}$, then W_i satisfies the Graphical Criteria with respect to $X \to Y$ given $\mathbf{Z} \cup \{W_j\}$. Moreover, if W_i and W_j share a hidden common cause that is independent of X and Y, not conditioning on W_j will induce an active back-door that will make the algorithm unnecessarily remove $\sigma_{w_i y.\mathbf{z}}/\sigma_{w_i x.\mathbf{z}}$ from the output.

Algorithm 4 IV-TETRAD$^+$

Input: $\mathcal{D}$, a sample from the joint distribution of random variables $\mathbf{V} \cup \{X, Y\}$; K, number of instruments to consider
Output: two sets of estimates of the differential causal effect of X on Y
Remove from $\mathbf{V}$ any element that is not in the Markov blanket of either X or Y
Let $\hat{\lambda}_i \leftarrow \hat{\sigma}_{v_i y.\mathbf{v}_{\setminus i}} / \hat{\sigma}_{v_i x.\mathbf{v}_{\setminus i}}$ for each $V_i \in \mathbf{V}$
Initialize $\mathcal{C}$ and $\mathcal{C}_R$ as empty sets, and $\mathbf{W}$ as $\mathbf{V}$
while $|\mathbf{W}| \geq K$ **do**
  Sort $\mathbf{W}$ as $W^{(1)}, \ldots, W^{(|\mathbf{W}|)}$ according to the corresponding $\{\hat{\lambda}\}$
  Find $j \in \{1, 2, \ldots, |\mathbf{W}|\}$ that minimizes ScoreTetrads$(W^{(j:j+K-1)}, \mathbf{V}, X, Y, \mathcal{D})$
  $\mathbf{W}_j \leftarrow$ ExpandTetrads$(j, K, \mathbf{W}, \mathbf{V}, X, Y, \mathcal{D})$
  IF TestTetrads$(\mathbf{W}_j, \mathbf{V}, X, Y, \mathcal{D})$ **THEN**
    $\hat{\lambda}_{TSLS} \leftarrow$ TSLS$(\mathbf{W}_j, \mathbf{V} \setminus \mathbf{W}_j, X, Y, \mathcal{D})$
    IF TestResiduals$(\mathbf{W}_j, \mathbf{V}, X, Y, \mathcal{D})$ **THEN**
      $\mathcal{C} \leftarrow C \cup \hat{\lambda}_{TSLS}$
    ELSE
      $\mathcal{C}_R \leftarrow C_R \cup \hat{\lambda}_{TSLS}$
    END IF
  END IF
  $\mathbf{W} \leftarrow \mathbf{W} \setminus \mathbf{W}_j$
end while
return $(\mathcal{C}, \mathcal{C}_R)$

there is at least one such set $\mathbf{W}$ of size K, which remains valid by conditioning on $\mathbf{V} \setminus \mathbf{W}$. One motivation is to avoid a combinatorial search for conditioning sets, while still having the option of rejecting a solution if confounding or collider bias remain by doing a test at the end. The other motivation is the statistical unreliability of candidate sets of small size: in a large system where the treatment may have many observed causes, instruments will in general be weakly associated with the treatment, leading to high variance estimates. This issue of "weak instruments" is pervasive in real problems and one mitigation is to consider instrument sets of a minimum size.

The algorithm remains conservative in the sense of missing possible valid IV sets of smaller size, or which require a different conditioning set, while testing whether the proposed solutions satisfy the relevant tetrad and residual independence constraints. The algorithm, IV-TETRAD$^+$ is shown in Algorithm 4.

The algorithm makes use of TSLS$(\mathbf{W}, \mathbf{Z}, X, Y, \mathcal{D})$, the two-stage least squares (TSLS) estimator for some set of conditional instruments $\mathbf{W}$ and conditioning set $\mathbf{Z}$, given a dataset $\mathcal{D}$. Assuming $\mathcal{D}$ is centered to have zero empirical mean, this is defined as follows: first, as an abuse of notation, let $\mathbf{W}$, $\mathbf{X}$, $\mathbf{Y}$ denote the corresponding empirical residuals of the least-squares regression of $\mathbf{W}$, X and Y on $\mathbf{Z}$. The projection matrix $P_{\mathbf{W}}$ is defined as $\mathbf{W}(\mathbf{W}^\top \mathbf{W})^{-1} \mathbf{W}^\top$. Let $\hat{\mathbf{X}}$ denote the predicted value $\hat{\mathbf{X}} \equiv P_{\mathbf{W}} \mathbf{X}$. The two-stage least squares coefficient is then given by the least-squares regression of $\mathbf{Y}$ on $\hat{\mathbf{X}}$, resulting in $\hat{\lambda} \equiv (\mathbf{X}^\top P_{\mathbf{W}} \mathbf{X})^{-1} \mathbf{X}^\top P_{\mathbf{W}} \mathbf{Y}$.

The details of the algorithm steps are as follows:

- In Line 2, we remove variables from $\mathbf{V}$ that are considered to be conditionally independent of X or Y. In our experiments in the next Section, this is implemented by running LARS (Efron et al., 2004), regressing separately X on $\mathbf{V}$ and Y on $\mathbf{V}$. We then remove from $\mathbf{V}$ any element that is given a coefficient of value zero in either regression. Moreover, we also remove from $\mathbf{V}$ any element marginally uncorrelated with X or Y by a test of vanishing Spearman correlations at a 0.05 level;

- In Line 4, $\hat{\sigma}_{ab.\mathbf{z}}$ corresponds to the conditional covariance as given by the empirical covariance matrix derived from $\mathcal{D}$, and $\mathbf{V}_{\setminus i} \equiv \mathbf{V} \setminus \{V_i\}$;

- In Line 8, function ScoreTetrads$(W^{(j:j+K-1)}, \mathbf{V}, X, Y, \mathcal{D})$ is defined as by first calculating $\hat{\lambda}_{TSLS} \equiv$TSLS $(\{W^{(j)}, \ldots, W^{(j+K-1)}\}, \mathbf{V} \setminus \{W^{(j)}, \ldots, W^{(j+K-1)}\}, X, Y, \mathcal{D})$. Function ScoreTetrads returns the median of the absolute differences between $\hat{\lambda}_{TSLS}$ and each $\lambda^{(j)}, \ldots \lambda^{(j+K-1)}$.

- In Line 9, we perform a greedy search on whether we should expand the current set $\{W^{(j)}, \ldots, W^{(j+K-1)}\}$ as either $\{W^{(j-1)}, \ldots, W^{(j+K-1)}\}$ or $\{W^{(j)}, \ldots, W^{(j+K)}\}$ until a local maximum for ScoreTetrads is found;

- In Line 10, we perform Wishart tests of conditional tetrad constraints (Wishart, 1928; Spirtes et al., 2000) for every possible pair in $\mathbf{W}_j$ against pair $\{X, Y\}$. We ignore that the test distribution in Wishart's test assumes Gaussianity and marginal covariances. As discussed by Spirtes (2013), alternative tests that do not assume Gaussianity (Bollen, 1990) seen to have no clear advantage compared to Wishart's. Function TestTetrads returns true if and only if more than half of the tested pairs return a test p-value greater than a particular pre-defined level. In our experiments in the next Section, this level was set at 0.05;

- Finally, in Line 12, for each $W \in \mathbf{W}_j$, we obtain the residual r_W of the least-squares regression of W on $\mathbf{V} \setminus W$, as well as the residual r_X of the regression of X on $\mathbf{V}$ (alternatively we could calculate r_Y, but we expect X to be more strongly associated with W as in theory there is no edge between W and Y). We perform a marginal independence test of these two residuals using HSIC (Gretton et al., 2007). As with the tetrad tests, TestResiduals will return true if and only if more than half of the p-values are above a particular threshold (again 0.05 in our experiments).

The rationale for Algorithm 4 is as follows. In the limit of infinite data, any set of K variables which satisfies the corresponding tetrad constraints will return a value of zero for ScoreTetrads, which will then be expanded by ExpandTetrads to include the remaining candidate IVs that imply the same ratio λ. Sets of size K that satisfy the Graphical Criteria will contribute with the true causal effect to $\mathcal{C}$, as will also sets of size K or larger corresponding to locally covariance equivalent causal effects. The algorithm can fail to produce information that a more exhaustive algorithm would produce if subsets of conditioning sets $\mathbf{V} \setminus W_j$ were searched (for instance, in the limit of infinite data, IV-TETRAD$^+$ will return an empty $\mathcal{C}_R$ for data generated by the model in Figure 3, which is in theory avoidable).

The separation between $\mathcal{C}$ and $\mathcal{C}_R$ is due to the fact that full automation of effect discovery is not recommended, as there are unidentifiability trade-offs. Recall that in the graph where

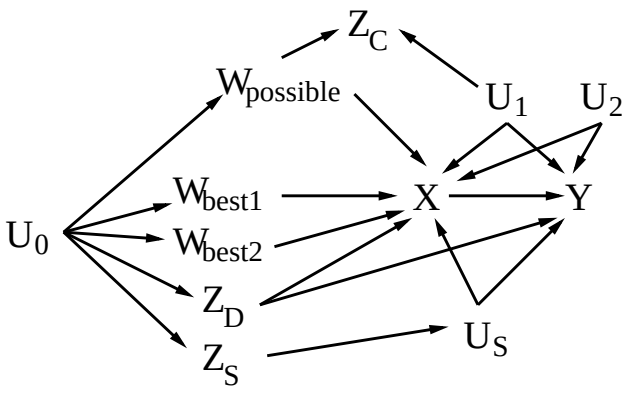


Figure 7: Example of a synthetic graph generated by the template used in Section 5. W_{best_1}, W_{best_2} and $W_{possible}$ are valid IVs conditioned on Z_D and Z_S only, as conditioning on Z_C activates the path $W_{possible} \rightarrow Z_C \leftarrow U_1 \rightarrow Y$, which invalidates that instrument. However, W_{best_1} and W_{best_2} remain valid even after conditioning on all other observed variables.

W_1 and W_2 are valid IVs, but where W_1 and X share an unmeasured confounder, that the residual test will asymptotically reject this candidate pair. It is to remind the user of this possibility that the algorithm separates in a second set the outcomes filtered by a more stringent criteria.

5. Experiments

We assess how IV-TETRAD$^+$ compares to other methods in a series of simulations. We then provide an example on how the method can be used in an empirical study and assess its robustness.

5.1 Synthetic Studies for Finding a Single Causal Effect assuming IVs Exist

We generate synthetic models of particular structures, intended to capture features such as confounding and collider bias that should be avoided by proper conditioning. In order to allow for a direct a comparison of IV-TETRAD$^+$ against alternative methods that return a single causal effect estimate, we pragmatically force our algorithm to choose a single element in each output $\mathcal{C} \cup \mathcal{C}_R$[2] for the sake of simplifying comparisons. The choice is the effect estimate corresponding to the set $\mathbf{W}_j$ that minimizes ScoreTetrads at the end of Line 9, which is then contrasted against the true effect[3].

2. Also, in this setup, the non-Gaussianity test of Line 12 does not play a role in the final output.
3. It is theoretically possible that this criterion will make IV-TETRAD$^+$ return the wrong causal effect in the limit of infinite data, as by design we introduce locally covariance equivalent causal effects. However, with finite data this criterion typically chooses solutions corresponding to correct IVs: in our setup, the non-IV vertices that imply locally covariance equivalent effects tend to be more weakly associated with outcome Y than the true IVs, as paths from such vertices to outcome Y pass through latent choke

Confounding	$\lambda_{xu} = \lambda_{yu} = 0.125$			$\lambda_{xu} = \lambda_{yu} = 0.25$		
N	1000	5000	10000	1000	5000	10000
NAIVE1	0.25, 0.79	0.25, 0.81	0.25, 0.81	0.50, 0.58	0.50, 0.58	0.50, 0.59
NAIVE2	0.16, 0.84	0.16, 0.84	0.15, 0.85	0.37, 0.68	0.36, 0.66	0.35, 0.69
NAIVE3	0.54, 0.56	0.54, 0.56	0.54, 0.56	0.78, 0.43	0.78, 0.40	0.78, 0.42
ORACLE	0.04, 0.94	0.02, 0.95	0.01, 0.98	0.13, 0.88	0.04, 0.95	0.02, 0.98
S-ORACLE	0.06, 0.92	0.03, 0.95	0.02, 0.97	0.17, 0.78	0.06, 0.93	0.03, 0.93
W-ORACLE	0.25, 0.80	0.23, 0.79	0.21, 0.77	0.52, 0.58	0.47, 0.60	0.44, 0.63
SISVIVE	0.34, 0.73	0.35, 0.72	0.31, 0.75	0.64, 0.56	0.67, 0.56	0.68, 0.57
IV-TETRAD$^+$	0.23, 0.78	0.12, 0.86	**0.07, 0.86**	**0.56, 0.61**	0.26, 0.77	**0.10, 0.79**

Table 1: Experimental results for 8 methods as described in the text, including three oracle competitors that illustrate idealized scenarios where parts of the structure are known. Two experimental parameters are varied: sample size N and the amount $\lambda_{xu} = \lambda_{yu}$ of unmeasured confounding between X and Y. For each pair of method and experimental condition, we report the summarized performance over 100 trials in two ways: the median absolute difference between the estimated causal effect and true λ_{yx} (the smaller, the better); and the proportion in which the sign of the estimated effect and the sign of the true effect agree (the higher, the better). IV-TETRAD$^+$ beats all non-oracle competitors at sample size 10000, with a p-value $p < 0.001$ according to a binomial test (in bold). Notice that NAIVE2 is actually the closer competitor to IV-TETRAD$^+$, while in this particular setup sisVIVE does not improve even with large sample sizes.

Simulations are performed as follows. Treatment X and outcome Y have two latent parents, U_1 and U_2. U_1 is connected to observed covariates as described below, so its variance decreases by conditioning. U_2 is not directly connected to the observed covariates. We then generate synthetic graphs by splitting the observed covariates in different groups.

Group $\mathbf{W}$ are variables which can be used as conditional IVs. This group has two subgroups, $\mathbf{W}_{best}$ and $\mathbf{W}_{possible}$. The former remains a set of valid IVs conditioned on all other observed covariates. The latter will have active collider paths with treatment X that can be in principle deactivated by a combinatorial search, which is not performed by any algorithm in our benchmark. Group $\mathbf{Z}_S$ is a set of covariates with a common latent child U_S which is also a parent of X and Y, making U_S a spurious choke point between $\mathbf{Z}_S$ and $\{X, Y\}$. Group $\mathbf{Z}_D$ is a set of covariates which are parents of both X and Y and as such are not blocked by choke points, latent or not. Group $\mathbf{Z}_C$ are variables which are children of $\mathbf{W}_{possible}$ and U_1. Finally, all variables in $\mathbf{WZ} \equiv \mathbf{W} \cup \mathbf{Z}_S \cup \mathbf{Z}_D$ have a common latent parent U_0, making conditioning on $\mathbf{WZ} \backslash W$ required for $W \in \mathbf{W}_{best}$ to be a valid IV. Figure 7 shows an example where $|\mathbf{W}_{best}| = 2, |\mathbf{W}_{possible}| = 1, |\mathbf{Z}_S| = 2, |\mathbf{Z}_D| = 1$ and $|\mathbf{Z}_C| = 1$.

The methods we compare against are:

points. This inflates the variance of the corresponding $\hat{\lambda}_i$, leading to higher disparity compared to the effect estimated by two-stage least squares.

1. NAIVE1, the least-squares regression coefficient of Y on X;

2. NAIVE2, two-stage least squares (TSLS) of Y on X using all variables $\mathbf{V} = \mathbf{W} \cup \mathbf{Z}_S \cup \mathbf{Z}_C \cup \mathbf{Z}_D$ as instruments;

3. NAIVE3, least-squares regression of Y on X and $\mathbf{V}$;

4. ORACLE, TSLS estimation using the correct set of IVs $\mathbf{W}$ and correct adjustment set $\mathbf{Z}_S \cup \mathbf{Z}_D$;

5. W-ORACLE, TSLS using $\mathbf{W}$ as IVs, but conditioning on all of the other variables $\mathbf{Z}_S \cup \mathbf{Z}_D \cup \mathbf{Z}_C$;

6. S-ORACLE, the Kang et al. (2015) algorithm performed by first correctly removing the set $\mathbf{Z}_C$ from the input;

7. SISVIVE, the Kang et al. (2015) algorithm taking all variables $\mathbf{V}$ as input;

8. IV-TETRAD$^+$, our method, using all variables $\mathbf{V}$ as input and $K = 10$.

All error variables and latent variables are zero-mean Laplacian distributed, and coefficients are sampled from standard Gaussians and re-scaled such that the observed variables have a variance of 1. A simulation is rejected until $|\lambda_{yx}| > 0.05$. Coefficients between $\{X, Y\}$ and latent variables $\{U_1, U_2\}$ are set such that $\lambda_{xu_1} = \lambda_{yu_1} = \lambda_{yu_2} = \lambda_{yu_2}$, at two levels, $(0.125, 0.25)$. The difficulty of the problem increases with λ_{xu_1}, as this makes unmeasured confounding stronger. Comparisons are shown in Table 1, with the setup $|\mathbf{W}_{best}| = 15$, $|\mathbf{W}_{possible}| = 10$, $|\mathbf{Z}_S| = 10$, $|\mathbf{Z}_D| = 5$, $|\mathbf{Z}_C| = 10$. This satisfies the criterion of $|\mathbf{W}|$ being half the number of remaining variables, although only the 15 variables $\mathbf{Z}_S \cup \mathbf{Z}_D$ should be used as a conditioning set.

We generate 100 synthetic problems with a dataset of 10,000 points each and assess methods using also subsets of size 1000 and 5000. For $\lambda_{xu} = 0.125$, the empirical distribution of effects λ_{yx} had a median of 0.34, an upper quartile of 0.55 and a maximum of 1.05. For $\lambda_{yu} = 0.25$, a median of 0.35, an upper quartile of 0.75 and a maximum of 1.28. For assessment, we use two measures that summarize absolute deviance from λ_{yx} and agreement with its sign.

The message in Table 1 is as follows. The parameter estimation problem is easy if one knows the right IVs and background variables (ORACLE), but it shows already its difficulties if one has potentially right IVs but conditions on the wrong set (W-ORACLE). sisVIVE can work very well if one knows in advance which conditioning variables to discard (S-ORACLE). In general we will not have this knowledge, and sisVIVE can potentially behave badly, even worse than a naïve method (compare NAIVE3 to SISVIVE). The large contrast between S-ORACLE and SISVIVE is a warning against ignoring the effects of incorrect conditioning. Of all non-oracle methods, IV-TETRAD$^+$ is the clear winner, although as expected it is a high variance estimator and should be advantageous only for relatively large sample sizes. In Section 5.3 we illustrate ways in which the uncertainty of this estimator can be assessed.

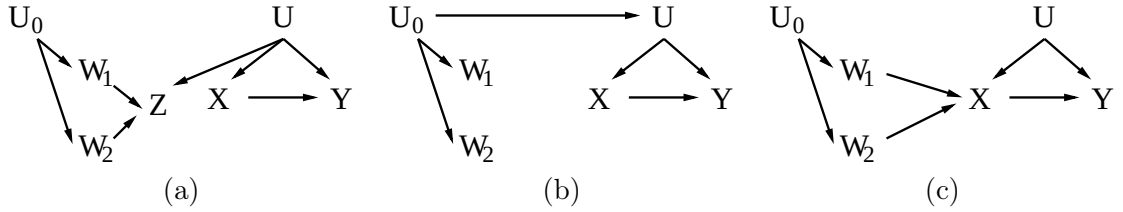


Figure 8: The three different configurations used in the assessment performed in Section 5.2.

5.2 Residual Test Assessment

We assess the robustness of the residual test validations of IV-TETRAD$^+$ in a series of simulations. We designed three scenarios as illustrated in Figure 8, where $\{U_0, U\}$ are latent variables:

S_1: The graph is given by Figure 8(a). The ideal output is to falsify the null hypothesis of residual independence between W_1 and X given $\{W_2, Z\}$, and W_2 and X given $\{W_1, Z\}$.

S_2: The graph is given by Figure 8(b). The ideal output is to falsify the null hypothesis of residual independence between W_1 and X given W_2, and W_1 and X given W_1.

S_3: The graph is given by Figure 8(c). The ideal output is *not* to falsify the null hypothesis of residual independence between W_1 and X given W_2, and W_2 and X given W_1.

In each scenario, we simulate 100 synthetic problems and generate a dataset of size 10,000 from each. The simulation of parameters is similar to Section 5.1, also including two levels of confounding by setting $\lambda_{xu} = \lambda_{yu}$ to either 0.125 or 0.25. We also increase the non-Gaussianity of the error terms: each error term is generated by sampling a Laplace distributed random variable as before, and then raising it to the power of 1.5, preserving the sign, and rescaling it back to have the same variance required so that each latent and observed variable has a marginal variance of 1. We do not claim that the tests proposed are particularly strong when the sample size is not large or the random variables are not clearly non-Gaussian. We do illustrate that, according to the theory, we should get better decisions with increasing sample sizes.

A joint decision is given for the pair of variables $\{W_1, W_2\}$, as in the definition of TestResiduals used in Algorithm 4. This is done by voting: for each candidate IV $W \in \{W_1, W_2\}$, we test the independence of r_W against r_X using HSIC (Gretton et al., 2007) conditioning on the remaining observed variables. If more than half of the candidate IVs result in a corresponding p-value greater than the chosen level of 0.05 (in our setup, this means that both W_1 and W_2 pass the test), then the model passes the test and the IVs are not falsified. Otherwise, the model is rejected. In our simulation, we are correct in each simulated instance if we reject the candidate instrument pair in scenarios S_1 and S_2, or if we do not reject it in scenario S_3.

Confounding	$N = 1000$			$N = 5000$			$N = 10000$		
	S_1	S_2	S_3	S_1	S_2	S_3	S_1	S_2	S_3
$\lambda_{yu} = 0.125$	0.54	0.28	0.87	0.77	0.85	0.89	0.82	0.93	0.83
$\lambda_{yu} = 0.25$	0.63	0.59	0.79	0.85	0.98	0.83	0.86	1.00	0.83

Table 2: Experimental results for assessing the detection of invalid instrument candidates based on the dependencies between least-squares residuals of instruments and treatments on a common pool of covariates. Within each sample size and amount of confounding, we show the proportion of times we correctly decided whether the corresponding pair of candidate IVs are indeed valid or not (out of 100 simulations). The three problems S_1, S_2 and S_3 are discussed in the main text. The correct decisions are to reject the candidates in problems of type S_1 and S_2, and not reject them in S_3.

We use a Shapiro-Wilk test of Gaussianity followed by the HSIC test as implemented in the R package dHSIC[4], both at a level 0.05. That is, if for candidate instrument W the Gaussianity assumption cannot be rejected for the residuals r_W or r_X, then W by default is not rejected as a plausible instrument[5]. Only default dHSIC hyperparameters were used to make the results more conservative: we believe better results can be obtained by more sophisticated approaches for hyperparameter selection. Similarly, the p-value threshold of 0.05 should not be seen the level of the test but as a regularization parameter, and should be adapted as the sample size increases (Kalisch and Bühlmann, 2007). Results are summarized in Table 2.

As expected, correctly rejecting the model in scenarios S_1 and S_2 is easier when confounding is stronger, which is precisely when we would like to be more conservative regarding our choice of instruments. Scenario S_2 is particularly hard under smaller sample sizes, as correlations quickly go to zero given a single path of length 3 between candidate instruments and treatment/outcome. The performance in Scenario S_3 plateaus, meaning that in this particular simulated distribution of parameters we still erroneously reject the valid IVs approximately 17% of the time. The Type I error probability is not that straightforward to calculate due to the voting mechanism and the dependency of the least-squares rejection estimates and the residual estimates. However, as it is known, in any constraint-based search algorithm we should decrease the test level hyperparameter from 0.05 towards zero as sample size increases if we want consistency. This is analogous to the behavior of other algorithms such as the PC algorithm (Spirtes et al., 2000).

To conclude, despite the noted shortcomings of requiring detectable non-Gaussianity and large sample sizes, the tests do provide a validation of proposed instruments as advertised. Moreover, they are expected to be more effective in discarding possibly invalid IVs as unmeasured confounding gets stronger, which is a desired behavior. However, we still recommend

4. Available at the Comprehensive R Archive Network, CRAN, `https://cran.r-project.org/`.
5. In theory the goal is to save computational time. In this cases the independence assumption is likely not to be rejected anyway, but the HSIC test can be costly if done many times so we want to avoid performing it whenever possible.

that we output candidate effects both with and without the residual independence tests as done by IV-TETRAD$^+$, for the reasons explained.

5.3 Empirical Illustration

We consider the application of our method to the study by Sachs et al. (2005). That study collected cell activity measurements (concentration of proteins and lipids) for single cell data under a variety of conditions. While searching for instrumental variables under a combination of experimental conditions is an interesting topic, we will focus on a single condition (described in the paper as stimulation with anti-CD3 and anti-CD28). There are 11 variables and 853 data points, of which we selected the manipulation of concentration levels of molecule *Erk* as the treatment, and concentration of *Akt* as the outcome. We use as background knowledge the model inference result shown as Figure 3A of (Sachs et al., 2005), encoding that the other 9 cell products are not causally affected by either *Erk* or *Akt*. Although previously unknown, evidence for this causal link was also given support by an experiment performed by Sachs et al. (2005).

The data shows some weak correlations, but we will assume for simplicity that no conditional independencies will hold between the treatment variable and the remaining 10 variables. The motivation for Step 3 in Algorithm 4 was primary as a variance reduction technique. Instead, here we will be primarily concerned about illustrating how we could assert uncertainty in our estimates using the modified Bayesian approach introduced by Silva and Kalaitzis (2015).

Running the standard regression adjustment NAIVE3 on this data, we get a differential causal effect of 1.36. sisVIVE reports an effect of 1.58, where all variables were chosen as instruments and as such this estimate here is the same as NAIVE2. This suggests that unmeasured confounding may be weakening the association between treatment and outcome, but sampling variability may be high as discussed in the next section. We run IV-TETRAD$^+$ as in Section 5.1 (and skipping Step 3), where a single effect of 1.43 is returned by minimizing ScoreTetrads. For this run we chose a window size $K = 3$, as $K = 2$ is in general too noisy of a choice since it cannot exploit any redundancies of the tetrad constraints implied by the candidate instrument sets. TestResiduals does not reject any candidate set proposed by the tetrad search. For $K = 4$ and $K > 4$ the results were 1.38 and 1.58 respectively, illustrating some of the difficulties of working with a relatively small sample size (853 points) using an algorithm that has been shown to require large samples. To illustrate what a practitioner should do in this case, we describe ways of assessing the uncertainty of the output of IV-TETRAD$^+$.

5.3.1 Exploration with the Bootstrap

First, we show a simple comparison based on the bootstrap: we sample the data with replacement and show the corresponding (bootstrapped) sampling distribution of the estimates obtained by NAIVE3, sisVIVE and the single-output IV-TETRAD$^+$ with $K = 3$. A smoothed estimate of the respective bootstrapped outcomes is shown in Figure 9. It is evident that as the method gets more flexible, the entropy of the respective sampling distribution of the estimates also increases. There is some evidence of bimodality of the sampling distribution, with IV-TETRAD$^+$ being particularly more extreme.

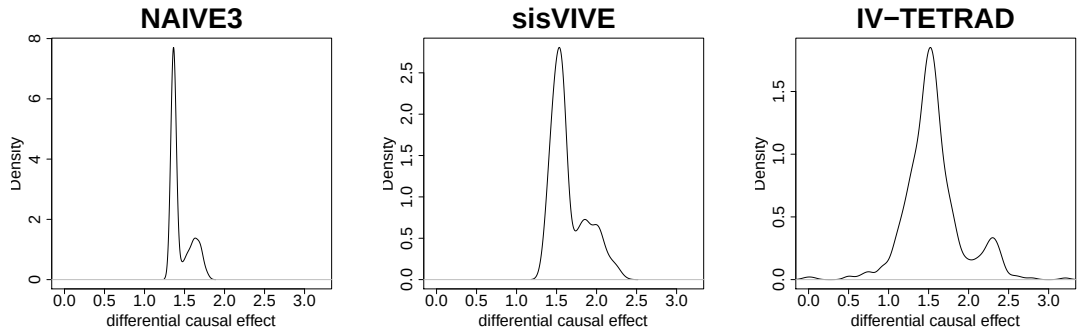


Figure 9: Smoothed sampling distribution of the estimated causal effects for the three discussed methods using 500 bootstrapped samples and kernel density estimation.

Bootstrap can in principle be used to find confidence intervals for the corresponding causal effects. However, it is relevant also to discuss Bayesian approaches for quantifying uncertainty. This is particularly interesting for a method like IV-TETRAD$^+$, which potentially returns sets of causal effects of varying sizes. It is not clear what a confidence interval (or confidence set) would be here. Returning a summary such as the minimum and maximum elements of the set, and building confidence intervals on bounds of the causal effect, may be too conservative and not very robust to outliers. In the next section, we will show a Bayesian solution to that problem. First, let us start by still considering the case where IV-TETRAD$^+$ returns a single element as chosen by ScoreTetrads.

Obtaining a posterior distribution over the causal effect of interest requires a likelihood function that includes all observed variables. The usual Bayesian approach for graphical modeling would require exploring a space of conditional tetrad and residual independence constraints, where any chosen set of constraints defines a model. We would then evaluate the marginal likelihood of each model and search for the maximum a posteriori or list several models of high probability. Alternatively, we in principle could perform Markov chain Monte Carlo (MCMC) sampling in the space of parameters and model constraints. A very different algorithm from IV-TETRAD$^+$ would be necessary, as it does not search among complete models. Such an algorithm would be computationally very intense, and it is not clear how to parameterize a joint likelihood for any given set of constraints, whether latent variables are explicitly introduced or not.

5.3.2 Learning by Bayesian Projections

Instead, we adopt the much simpler "Bayesian projections" approach suggested by Silva and Kalaitzis (2015), which has a relation to several other approaches for model selection that avoid introducing constraints directly in the likelihood function, such as the one by Goutis and Robert (1998). We define a black-box non-Gaussian likelihood function for the observed variables, generate posterior samples from it by MCMC, and perform a causal effect search algorithm similar to IV-TETRAD$^+$. The main difference here is that instead of performing statistical tests, we instead reject constraints if they violate a threshold of misfit.

To make it more explicit, we define an algorithm which we will call IV-TETRAD^{++}. In this algorithm, we are given M posterior samples of the model parameters from a black-box mixture of Gaussians likelihood. Within each sample $1 \leq m \leq M$, we run IV-TETRAD^{+} with the following modifications:

1. Empirical covariances $\hat{\sigma}_{ab.\mathbf{s}}$ are replaced with the corresponding model covariances $\sigma^{(m)}_{ab.\mathbf{s}}$, which can be computed analytically for a mixture of Gaussians model;

2. We ignore Step 3 of IV-TETRAD^{+}, the one where we originally estimated Markov blankets. Instead, we discard every λ_i such that $|\lambda_i| > \alpha_{cut}$. Hyperparameter α_{cut} is given as an input, implying the removal from the output of unusually high causal effects, including some of those resulting from variables V_i such that $\sigma_{v_i x.\mathbf{v}_{\setminus i}} \approx 0$.

3. TestTetrads is modified to reject a tetrad constraint if $|\rho^{(m)}_{ix.\mathbf{v}_{\setminus i}}\rho^{(m)}_{jy.\mathbf{v}_{\setminus j}} - \rho^{(m)}_{iy.\mathbf{v}_{\setminus i}}\rho^{(m)}_{jx.\mathbf{v}_{\setminus j}}| > \tau$. Hyperparameter τ is given as an input, and provides the rule that decides on whether a tetrad constraint fits the model given by the m-th MCMC sample with model partial correlations $\rho^{(m)}_{ab.\mathbf{s}}$.

4. For simplicity, we do not perform TestResiduals. It could in principle be implemented by sampling a large synthetic dataset from model m, estimating a measure of dependence between the residuals, and rejecting independence with this measure is larger than a given hyperparameter. Due to its computational cost and the fact it does not change the message of this section, we will ignore this step.

Silva and Kalaitzis (2015) discuss the shortcomings of an approach based on Bayesian projections compared to a traditional Bayesian approach where the likelihood enforces constraints explicitly. However, the computational advantages are major, as sampling from the posterior of a more standard black-box model is much simpler than searching within a complex likelihood space. As a matter of fact, at the present time it is not known how to define such a likelihood function. A Bayesian projection approach has some similarities to frequentist bootstrap, where a black-box object (the empirical distribution, in the nonparametric bootstrap case) provides the source of variability (the sampling distribution, in the bootstrap case). This allows us to leverage constraint-based algorithms with few modifications.

The most delicate design choices are the choice of α_{cut} and τ. One possibility to assess its impact is by sampling from the prior and plotting the implied distribution on differential causal effects. This is done in Figure 10, where the black-box distribution is a mixture of 5 Gaussians with the following priors: the mixing distribution is given a Dirichlet prior with hyperparameter $(2, 2, 2, 2, 2)$; the covariance matrix Σ_c of each mixture component c is given an inverse Wishart prior with $\nu = 11 + 3$ degrees of freedom and a $\nu\mathbf{I}$ scale matrix, with $\mathbf{I}$ being the 11×11 identity matrix; the mean vector μ_c of each component is given a multivariate Gaussian prior with zero mean and covariance matrix Σ_c.

Given the prior for the base distribution, Figure 10 shows a smoothed depiction of 1000 samples from the priors defined by sampling the mixture model parameters from the base distribution and passing them through IV-TETRAD^{++}, where we choose the single element that minimizes ScoreTetrads. To generate the figure, we used the the following configurations: i. $(\alpha_{cut} = 10, \tau = 0.01)$; ii. $(\alpha_{cut} = 3, \tau = 0.01)$ and iii. $(\alpha_{cut} = 10, \tau =$

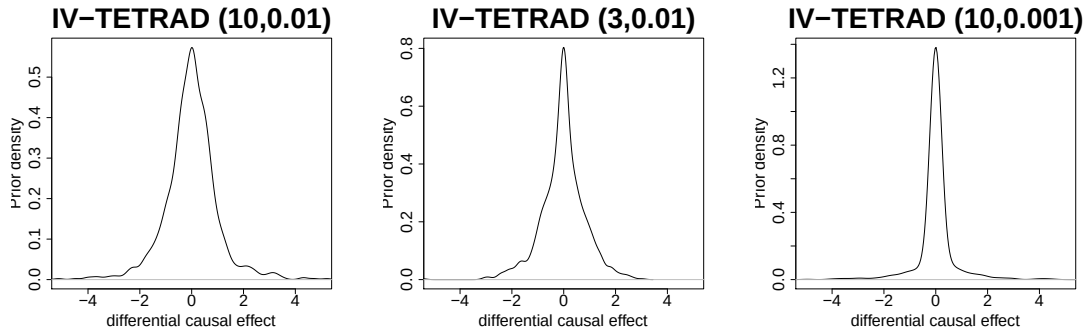
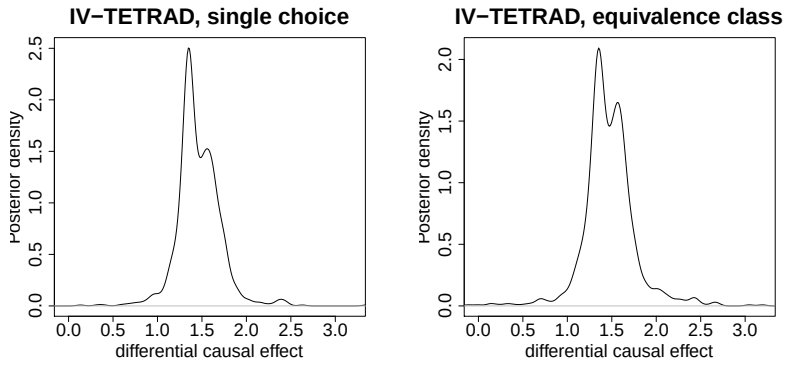


Figure 10: Examples of the implied prior on the differential causal effect using the Bayesian projections algorithm IV-TETRAD^{++} with three choices of hyperparameters ($\alpha_{cut} = 10, \tau = 0.01$), ($\alpha_{cut} = 3, \tau = 0.01$) and ($\alpha_{cut} = 10, \tau = 0.001$).



(a) (b)

Figure 11: (a) The posterior distribution when a single value is chosen as the output of IV-TETRAD^{++}. (b) The resulting posterior where we uniformly sample one of the elements of the equivalence class.

0.001). As expected, decreasing these two hyperparameters makes the prior narrower. At the same time, since the effect is given by the ratio of covariance entries, we expect the implied prior to be heavy-tailed. For illustration purposes, we proceed to analyze the Sachs et al. data by choosing ($\alpha_{cut} = 10, \tau = 0.01$) as the prior configuration.

5.3.3 Results

Figure 11(a) shows the resulting posterior after running MCMC for the mixture of Gaussians for 10,000 iterations and thinning it down to 1000 samples. The evidence that 5 mixture components is enough is given by the fact that the component of smallest posterior probability was assigned almost zero mass. The figure illustrates the difficulty of the problem and

why a single point estimate might not be a good summary of the outcome even in the case where a single causal effect is returned.

More interestingly, the Bayesian approach allows us to generate posteriors over equivalence classes in a relatively simple way, without resorting to potentially uninformative summaries based on bounds. This however requires subject matter information, as the data cannot distinguish the elements in an equivalence class. Following the philosophy of Richardson et al. (2011), we separate priors for elements which are informed by the data (equivalence class) from priors for elements which cannot be distinguished (the choice *within* the equivalence class, given the equivalence class). Figure 11(b) shows the posterior distribution that, for each sample m of the Markov chain, generates a sample causal effect by sampling uniformly at random from the elements in the corresponding equivalence class of sample m. That is, the extra information here is the uniform prior, which may be taken as a default choice. In principle, another prior can be used. For instance, if it is agreed that the choice should be made based, for instance, on a distribution that assigns mass in a way that is inversely proportional to the magnitude of each element in the class (this assumes that smaller effects are more likely than large effects). In any case, the important lesson is the understanding that this final component of the posterior is independent of the data given the equivalence class[6]. The conclusion from Figure 11(b) is that the bimodality is explained not only by the data, but also by the indeterminacy of the solution, in contrast to the bimodality found in Figure 9, which is an artifact of the sampling distribution of the corresponding estimators.

6. Related Work

The notion of "equivalence class of causal effects" is not new, even if not previously named as such. Hoyer et al. (2008) discuss how causal effects in LiNGAM models can be partially identified if the number of latent confounders is known. Their method also returns sets of causal effects, but without any need for further covariates and instruments. However, the size of the equivalence class of causal effects grows rapidly with the number of assumed latent variables and it is unclear how to determine the number of unmeasured confounders. Maathuis et al. (2009) present extensions of the PC algorithm which return sets of causal effects that follow from different members of an equivalence class of DAGs compatible with the observable independence constraints.

Didelez and Sheehan (2007) present a survey on instrumental variables from the point of view of applications in Mendelian randomization. A complementary method to sisVIVE based on similar assumptions, but performing estimation by the reweighted median of covariance ratios, is discussed by Bowden et al. (2016). The difficulty of estimating effects with weak instruments is well-known in the statistics literature. A review that covers some background on the estimation with weak instruments is provided by Burgess et al. (2017).

A different way of inducing "approximate" instrumental variables was introduced by Silva and Evans (2016). In that case, only (approximate) independence constraints are exploited. The same machinery in principle could include approximate tetrad constraints, although

6. This idea could also be adapted to the case where the empty set is returned in some of the MCMC samples: the posterior would be a mixture model with positive mass on the empty set and the remaining mass spread over the values found.

further research on that would be necessary. The problem of equivalence class of causal effects is also present in that approach. In such a scenario, however, there are infinitely many alternatives, which can be represented as an interval. The width of the interval is typically large and uninformative in classical bounding methods for instrumental variables (Didelez and Sheehan, 2007), but the approach introduced by Silva and Evans (2016) allows for the introduction of stronger assumptions as a way of reducing the size of the equivalence class. This borrows some concepts from earlier work such as (Ramsahai, 2012).

The use of tetrad constraints for testing the validity of particular edge exclusions in linear causal models has a long history, dating back at least to Spearman (1904). Although the generalizations discussed by Sullivant et al. (2010) were presented in the context of Gaussian models, this distributional assumption is not necessary, with their rank constraint results exploited even in the context of partially non-linear models by Spirtes (2013). More recently, tetrad constraints have been used in the discovery of latent variable model structure (Silva et al., 2006; Spirtes, 2013), where structures such as Figure 2(c) emerge but no direct relationships among observables (such as $X \rightarrow Y$) are discoverable. The combination of tetrad constraints and non-Gaussianity assumptions has been exploited by Shimizu et al. (2009), again with the target being relationships among latent variables. Tetrad tests for the validity of postulated IVs were discussed by Kuroki and Cai (2005).

The literature on learning algorithms allowing for latent variables has been growing steadily, including the Fast Causal Inference algorithm of Spirtes et al. (2000) and more recent methods that exploit constraints other than independence constraints (Tashiro et al., 2014; Nowzohour et al., 2015), but none of these methods allow for the estimation of the causal effect of X and Y when there is an unblocked unmeasured confounder between them. Phiromswad and Hoover (2013) introduced an algorithm for IV discovery, but it does not take into account unidentifiability issues that can be solved by exploring constraints other than covariance matrix constraints. It also returns equivalence classes of graphs and requires searching for multiple causal effects at the same time, contrary to our goal of focusing on a given causal effect λ_{yx}.

7. Conclusion

Finding instrumental variables is one of the most fundamental problems in causal inference. To the best of our knowledge, this paper provides the first treatment on how this can be systematically achieved by exploiting rank constraints and clarifying to which extent an equivalence class of solutions remains. We then proceeded to show how non-Gaussianity can be exploited in a pragmatic way, by adapting a state-of-the-art algorithm. Finally, we illustrated how empirical improvement can be obtained.

We expect that theoretical challenges in instrumental variable discovery can be further tackled by building on the findings shown here. In particular, as also hinted by Kang et al. (2015), some of the ideas here raised extend to non-linear (additive), heterogeneous effects and binary models. Methods developed by Peters et al. (2014) can potentially provide a starting point on how to allow for non-linearities in the context of instrumental variables. As discussed by Spirtes (2013), linearity is only really needed "downstream" of the choke point: that is, it would be enough that only the structural equation for outcome Y is

linear. Theoretical details need to be sorted out, but IV-TETRAD$^+$ should in principle be practically applicable to non-linear models with linearity for Y only.

More sophisticated graphical criteria for the identification of causal effects in linear systems were introduced by Brito and Pearl (2002). Further work has led to rich graphical criteria to identify causal effects in confounded pairs (Foygel et al., 2011), going beyond the IV criteria discussed here. This opens up the possibility elaborated discovery algorithms where back-door blocking (Entner et al., 2012) and the methods in this paper cannot provide a solution. How to perform this task in a computationally and statistically tractable way remains an open question.

Code for all experiments is available at `http://www.homepages.ucl.ac.uk/~ucgtrbd/code/iv_discovery`.

Acknowledgments

We thank Robin Evans, Peter Spirtes and Kun Zhang for several useful discussions. Part of the discussions were possible by travel funding supporting the attendance of RS and SS of the Short Thematic Program "Statistical Causal Inference and Applications to Genetics" organized and sponsored by the Centre the Recherches Mathématiques, University of Montreal, Canada. RS's visit was also partially supported by a grant from Adobe Research. RS also acknowledges travel support from the Center for Causal Discovery in Pittsburgh, PA, where part of the discussions that originated this paper took place.

References

J. Angrist and J-S Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton University Press, 2009.

K. Bollen. *Structural Equations with Latent Variables.* John Wiley & Sons, 1989.

K. Bollen. Outlier screening and a distribution-free test for vanishing tetrads. *Sociological Methods and Research*, 19:80–92, 1990.

J. Bowden, G. Smith, P. Haycock, and S. Burgess. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic Epidemiology*, 40:304–314, 2016.

C. Brito and J. Pearl. Generalized instrumental variables. *Proceedings of 18th Conference on Uncertainty in Artificial Intelligence*, 2002.

S. Burgess, D. Small, and S. Thompson. A review of instrumental variable estimators for mendelian randomization. *Statistical Methods in Medical Research*, To appear, 2017.

B. Chen, J. Tian, and J. Pearl. Testable implications of linear structural equations models. *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI)*, pages 2424–2430, 2014.

T. Chu, R. Scheines, and P. Spirtes. Semi-instrumental variables: a test for instrument admissibility. *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence (UAI 2001)*, pages 83–90, 2001.

G. Cooper. A simple constraint-based algorithm for efficiently mining observational databases for causal relationships. *Data Mining and Knowledge Discovery*, 2, 1997.

G. Darmois. Analyse générale des liaisons stochastiques. *Review of the International Statistical Institute*, 21:2–8, 1953.

V. Didelez and N. Sheehan. Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research*, 16:309–330, 2007.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32:407–499, 2004.

D. Entner, P.O. Hoyer, and P. Spirtes. Statistical test for consistent estimation of causal effects in linear non-Gaussian models. *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS 2012)*, pages 364–372, 2012.

R. Foygel, J. Draisma, and M. Drton. Half-trek criterion for generic identifiability of linear structural equation models. *Annals of Statistics*, 40:1682–1713, 2011.

C. Goutis and C. Robert. Model choice in generalised linear models: a Bayesian approach via kullback-leibler projections. *Biometrika*, 85:28–37, 1998.

A. Gretton, K. Fukumizu, C. Teo, L. Song, B. Schölkopf, and A. Smola. A kernel statistical test of independence. *Advances in Neural Information Processing Systems*, 20:585–592, 2007.

P. Hoyer, S. Shimizu, A. Kerminen, and M. Palviainen. Estimation of causal effects using linear non-Gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49:362–378, 2008.

M. Kalisch and P. Bühlmann. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*, 8:613–636, 2007.

H. Kang, A. Zhang, T. Cai, and D. Small. Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. *Journal of the American Statistical Association*, page To appear, 2015.

M. Kuroki and Z. Cai. Instrumental variable tests for directed acyclic graph models. *Tenth workshop on Artificial Intelligence and Statistics (AISTATS 2005)*, 2005.

M. H. Maathuis, M. Kalisch, and P. Bühlmann. Estimating high-dimensional intervention effects from observational data. *Annals of Statistics*, 37:3133–3164, 2009.

S. Mani, G. Cooper, and P. Spirtes. A theoretical study of Y structures for causal discovery. *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence (UAI2006)*, pages 314–323, 2006.

S. Morgan and C. Winship. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press, 2015.

C. Nowzohour, M. Maathuis, and P. Bühlmann. Structure learning with bow-free acyclic path diagrams. *arXiv:1508.01717*, 2015.

J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2000.

J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15:2009–2053, 2014.

P. Phiromswad and K. Hoover. Selecting instrumental variables: A graph-theoretic approach. *Working paper. Available at SSRN: http://ssrn.com/abstract=2318552 or http://dx.doi.org/10.2139/ssrn.2318552*, 2013.

R. Ramsahai. Causal bounds and observable constraints for non-deterministic models. *Journal of Machine Learning Research*, pages 829–848, 2012.

T. Richardson and P. Spirtes. Ancestral graph Markov models. *Annals of Statistics*, 30: 962–1030, 2002.

T. Richardson, R. Evans, and J. Robins. Transparent parameterizatios of models for potential outcomes. In J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West, editors, *Bayesian Statistics 9*, pages 569–610. Oxford University Press, 2011.

K. Sachs, O. Perez, D. Pe'er, D. Lauffenburger, and G. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308, 2005.

G. Shafer, A. Kogan, and P.Spirtes. Generalization of the tetrad representation theorem. *DIMACS Technical Report*, 1993.

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

S. Shimizu, P. Hoyer, and A. Hyvärinen. Estimation of linear non-Gaussian acyclic models for latent factors. *Neurocomputing*, 72:2024–2027, 2009.

R. Silva and R. Evans. Causal inference through a witness protection program. *Journal of Machine Learning Research*, 17(56):1–53, 2016.

R. Silva and A. Kalaitzis. Bayesian inference via projections. *Statistics and Computing*, 25: 739–753, 2015.

R. Silva, R. Scheines, C. Glymour, and P. Spirtes. Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7:191–246, 2006.

W. Skitovitch. On a property of the normal distribution. *Doklady Akademii Nauk SSSR*, 89:217—-219, 1953.

C. Spearman. "General intelligence," objectively determined and measured. *American Journal of Psychology*, 15:210–293, 1904.

P. Spirtes. Calculation of entailed rank constraints in partially non-linear and cyclic models. *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence (UAI 2013)*, pages 606–615, 2013.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction and Search*. Cambridge University Press, 2000.

S. Sullivant, K. Talaska, and J. Draisma. Trek separation for Gaussian graphical models. *Annals of Statistics*, 38:1665–1685, 2010.

T. Tashiro, S. Shimizu, A. Hyvärinen, and T. Washio. ParceLiNGAM: A causal ordering method robust against latent confounders. *Neural Computation*, 26:57–83, 2014.

J. Wishart. Sampling errors in the theory of two factors. *British Journal of Psychology*, 19: 180–187, 1928.

APPENDIX: Proof of Theorem 4

The result for Theorem 4 depends on this standard theorem (Darmois, 1953; Skitovitch, 1953):

Theorem 9. (Darmois-Skitovitch Theorem) *Let $e_1, \ldots, e_n$ be independent random variables, $n \geq 2$. Let $v_1 = \sum_i \alpha_i e_i$, $v_2 = \sum_i \beta_i e_i$ for some coefficients $\{\alpha_i\}$, $\{\beta_i\}$. If v_1 and v_2 are independent, then those e_j for which $\alpha_j \neq 0$, $\beta_j \neq 0$ are Gaussian.*

The idea is that if we assume $\{e_i\}$ are not Gaussian, $\{V_1, V_2\}$ share a common source if and only if they are dependent. See (Shimizu et al., 2006; Entner et al., 2012) for a deeper discussion on how this theorem is used in causal discovery.

For the main results, we will assume that particular algebraic (polynomial) identities implied by the model graph do not vanish at the particular parameter values of the given model (which we called "almost everywhere" results in the theorem). We will in particular consider ways of "expanding" the structural equations of each vertex according to *exogenous* variables, that is, any variable which is either an error term or latent variable (assuming without loss of generality that latent variables have no parents).

For each vertex V_k in the model, and each exogenous ancestor E_m of V_k, let $\mathcal{P}_{km}$ be the set of all directed paths from E_m to V_k. For each path $p \in \mathcal{P}_{km}$, define $\phi_{kmp} = \prod_j \lambda_{jj'}$, the product of all coefficients along this path for $V_j \in \mathbf{V} \cap p$ where $V_{j'} = p \cap par_{\mathcal{G}}(j)$ (that is, $V_{j'}$ is the parent of V_j in this path. We multiply coefficients following a sequence $E_m \to \cdots \to V_{j'} \to V_j \to \cdots \to V_k$). From this,

$$V_k = \sum_{E_m \in \mathcal{A}_{\mathcal{G}}(k)} \sum_{p \in \mathcal{P}_{km}} \phi_{kmp} E_m, \tag{4}$$

where $A_{\mathcal{G}}(k)$ is the set of exogenous ancestors of V_k, where for $E_m = e_k$ we have $\phi_{kmp} \equiv 1$ and path p is given by the single edge $e_k \to V_k$. We refer to the idea of expansion a few times in the proofs as a way of describing how the models can be written as polynomial functions of the coefficients $\Lambda_{\mathcal{G}} = \{\lambda_{ij} \mid V_j \in par_{\mathcal{G}}(i)\}$.

Overall, for a LiNGAM model $\mathcal{M}$ with DAG $\mathcal{G}$, we denote by $\mathcal{X}_{\mathcal{G}}$ the set of exogenous variables of $\mathcal{M}$, and by the *expanded graph* of $\mathcal{M}$ the graph $\mathcal{G}$ augmented with the error terms and the corresponding edges $e_i \to V_i$ for all observable vertices V_i in $\mathcal{G}$.

The main result used in the proof of Theorem 4 comes from the following Lemma. Notice that the non-Gaussianity assumption and the Darmois-Skitovitch Theorem are not necessary for its proof.

Lemma 10. *Let $\mathbf{V} \cup \mathbf{U}$ be the set of variables in a zero-mean LiNGAM model $\mathcal{M}$, where $\mathbf{U}$ are the latent variables of the model. Let $V_i \in \mathbf{V}$, and define $\mathbf{V}_{\setminus i} \equiv \mathbf{V} \setminus \{V_i\}$. Let $r_i \equiv V_i - \mathbf{a}^T \mathbf{V}_{\setminus i}$ be the residual of the least-squares regression of V_i on $\mathbf{V}_{\setminus i}$, with $\mathbf{a}$ being the corresponding least-squares coefficients. Then, almost everywhere, r_i can be written as a linear function of the exogenous variables of $\mathcal{M}$, $r_i = \sum_{E_m \in \mathcal{X}_\mathcal{G}} c_m E_m$, where $c_m \neq 0$ if and only if V_i is d-connected to E_m given $\mathbf{V}_{\setminus i}$ in the expanded graph of $\mathcal{M}$.*

Proof of Lemma 10. Without loss of generality, assume that each latent variable in $\mathbf{U}$ has no parents. We will sometimes use X_k as another representation of any particular model variable (observable, latent or error term), with the index k indicating particular variables in $\mathbf{V} \cup \mathbf{U}$ and error terms, depending on the context.

One way of obtaining r_i is by first performing least-squares regression of each model variable X_k on V_j, for some $V_j \neq V_i$ in $\mathbf{V}$, and calculating residuals $X_k^{(1)}$. Define $\mathbf{V}^{(1)}$ as the set of all residuals $\{V_k^{(1)}\}$, $k \neq j$. We then repeat the process by regressing on some element of $\mathbf{V}^{(1)} \setminus \{V_i^{(1)}\}$, iterating until we are left with $\mathbf{V}^{(n-1)}$ containing the single element $V_i^{(n-1)}$, where n is the size of $\mathbf{V}$ and $V_i^{(n-1)} = r_i$. The elimination sequence can be arbitrary.

Let V_j be a vertex in $\mathbf{V}_{\setminus i}$. Let λ_{km} be the structural coefficient between V_k and any $X_m \in \mathbf{V} \cup \mathbf{U}$. We define $\lambda_{kj} \equiv 0$ if X_j is not a parent of V_k. Since

$$V_k = \lambda_{kj} V_j + \sum_{X_m \in par_\mathcal{G}(k) \setminus V_j} \lambda_{km} X_m + e_k,$$

we have

$$\sigma_{kj} = \lambda_{kj} \sigma_{jj} + \sum_{X_m \in par_\mathcal{G}(k) \setminus V_j} \lambda_{km} \sigma_{mj} + \sigma_{e_k j},$$

where $\sigma_{e_k j}$ is the covariance of e_k and V_j and σ_{mj} here represents the covariance of X_m and V_j. This implies,

$$a_{kj}^{(1)} = \lambda_{kj} + \sum_{X_m \in par_\mathcal{G}(k) \setminus V_j} \lambda_{km} a_j^{(1)} + a_{e_k j}^{(1)}. \tag{5}$$

where $a_{e_k j}^{(1)}$ is the least-squares regression coefficient of e_k on V_j. This means $V_k^{(1)} = V_k - a_{kj}^{(1)} V_j$ can be written as

$$V_k^{(1)} = \sum_{X_m \in par_\mathcal{G}(k) \setminus V_j} \lambda_{km} X_m^{(1)} + e_k^{(1)} \tag{6}$$

with $X_m^{(1)}$ and $e_k^{(1)}$ defined analogously.

We can iterate this process until we are left with r_i:

$$r_i = \sum_{U_k \in par_\mathcal{G}(i) \cap \mathbf{U}} \lambda_{ik} U_k^{(n-1)} + e_i^{(n-1)}, \tag{7}$$

where $|\mathbf{V}| = n$. Variable $U_k^{(n-1)}$ is the residual of the regression of U_k on $\mathbf{V}_{\setminus i}$, similarly for $e_i^{(n-1)}$.

What we will show next is that within (7) each $U_k^{(n-1)}$ and $e_i^{(n-1)}$ can be expanded as polynomial functions of $\Lambda_{\mathcal{G}}$ and $\mathcal{X}_{\mathcal{G}}$, and the end result will contain non-vanishing monomials that are a (linear) function of only the exogenous variables E_m which are d-connected to V_i given $\mathbf{V}\setminus V_i$ in the expanded graph of $\mathcal{M}$. Since the monomials cannot vanish except for a strict subset of lower dimensionality than that of the set of possible $\Lambda_{\mathcal{G}}$, the result will hold almost everywhere.

Since we are free to choose the elimination ordering leading to r_i, as they all lead to the same equivalent relation (7), let us define it in a way that a vertex can be eliminated at stage t only when its has no ancestors in $\mathbf{V}^{(t-1)}$ (where $\mathbf{V}^{(0)} \equiv \mathbf{V}_{\setminus i}$).

For $t = 1$, the only exogenous variables which will have a non-zero coefficient multiplying V_j in the least-squares regression are the parents of V_j in the expanded graph, since V_j has no other ancestors[7]. Let $U_k^{(1)}$ be the residual of some latent parent of V_j,

$$U_k^{(1)} = U_k - a_{kj}\left(\sum_{E_m \in \mathcal{A}_{\mathcal{G}}(j)} \sum_{p \in \mathcal{P}_{jm}} \phi_{jmp} E_m\right), \tag{8}$$

where $\phi_{jmp} \equiv \lambda_{jm}$ if E_m is a latent variable, or 1 if $E_m = e_j$. Moreover, $a_{kj} = \lambda_{jk} v_{kk}/\sigma_{jj}$, where v_{kk} is the variance parameter of U_k and σ_{jj} is a polynomial function of $\Lambda_{\mathcal{G}}$. We can multiply both sides of the equation above by σ_{jj} (as well all equations referring to any $V_k^{(1)}$ or $X_m^{(1)}$ such as (6)) to get a new system of variables that is polynomial in $\Lambda_{\mathcal{G}}$. We will adopt this step implicitly and claim that from (8) we have that $U_k^{(1)}$ can be expanded as parameters that are polynomial functions of $\Lambda_{\mathcal{G}}$. Moreover, it is clear from (8) that there will be at least one non-vanishing monomial containing each E_m. In what follows, we refer to any expression analogous to (8) as the *expansion* of $U_k^{(t)}$ for $t = 1, 2, \ldots, n-1$.

We define a DAG $\mathcal{G}^{(1)}$ with vertices $X_m^{(1)}$, where $U_k^{(1)}$ will assume as children all and only the $V_k^{(1)}$ such that U_k is a parent of V_k in the original extended graph of the model. That is, $\mathcal{G}^{(1)}$ is the extended graph over residuals after the first regression. The respective model $\mathcal{M}^{(1)}$ is given by equations of type (6) with parameters coming from $\Lambda_{\mathcal{G}}$[8].

For any $t > 1$, let $V_j^{(t)}$ be the vertex being eliminated. Each $U_k^{(t)}$ in which $U_k^{(t-1)}$ is a parent of $V_j^{(t-1)}$ in $\mathcal{G}^{(t-1)}$ will be a polynomial function of $\Lambda_{\mathcal{G}}$ and a linear function of the union of the exogenous variables present in the expansion of each parent of $V_j^{(t-1)}$: the expansion analogous to (8) in the new model will always introduce new symbols $\lambda_{j\star}$ into existing monomials, or create new monomials with e_j, as vertex $V_j^{(t-1)}$ had no eliminated descendants up to iteration t. As such, no exogenous variable will be eliminated from the algebraic expansion of the respective $U_k^{(t)}$.

7. Assuming V_j is not a child of V_i. In this case, without loss of generality we assume that the parents of V_i are added to the parents of V_j, and remove V_i from the model at any iteration t.
8. To be more precise, polynomial functions of such parameters, as we are implicitly multiplying each equation by σ_{jj}.

Finally, the expansion of $\lambda_{ik}U_k^{(n-1)}$ in (7) will not cancel any monomial in the expansion of some other $\lambda_{ik'}U_{k'}^{(n-1)}$: since U_k and $U_{k'}$ are both parents of V_i, no monomial in the expansion of U_k can differ from a monomial in the expansion of $U_{k'}$ by a factor of $\lambda_{ik}\lambda_{ik'}$. So (7) will depend algebraically on the union of the exogenous terms leading to each $U_k^{(n-1)}$.

To prove the Lemma, we start by pointing out that $U_k^{(n-1)}$ will have a latent/error parent of some V_j in its expansion if and only if there is at least one sequence of vertices $(V_c, \ldots, V_j)$ where V_c is an observable child of U_k and any two consecutive elements in this sequence have at least one common latent parent in $\mathcal{G}$ (the sequence can be a singleton, $V_c = V_j$). To see this, notice that the different $U_k^{(t)}$ form an equivalence relation: each $U_k^{(t)}$ with a $V_j^{(t-1)}$ child which is being eliminated at iteration t will include into its expansion the exogenous variables found in the expansion of the other parents of $V_j^{(t-1)}$. This partitions $\mathbf{V}_{\setminus i}$ into sets in which each vertex V_j can "reach" some other vertex V_k by first moving to some $V_{j'}$ which shares a latent parent with V_j and which can "reach" V_k. The latent parents of $\mathbf{V}$ are then partitioned according to their observed children.

To finalize the proof, suppose V_i is d-separated from a latent/error parent E_m of V_j given $\mathbf{V}\backslash V_i$. This happens if and only if all latent parents of V_i (and e_i) are d-separated from E_m given $\mathbf{V}\backslash V_i$. Let U_k be a latent parent of V_i (or its error term). Then $U_k^{(n-1)}$ cannot have E_m in its expansion. If this was the case, by the previous paragraph U_k would be d-connected to all latent parents of V_j, meaning V_i would be d-connected to them. This implies $c_m = 0$. Conversely, suppose V_i is d-connected to the error term or a latent parents of V_j given $\mathbf{V}\backslash V_i$. Then again by the previous paragraph, for any latent parent U_k of V_i, $U_k^{(n-1)}$ will have the latent parents of V_j as terms in its expansion, implying $c_m \neq 0$ almost everywhere. □

We can now prove Theorem 4.

Proof of Theorem 4. Considering the system for $\{V_i, Y\} \cup \mathbf{Z}$, we can represent the model in an equivalent way where all latent variables are exogenous. Applying Lemma 10 to both r_i and r_y, and by Theorem 9, these variables will be dependent if and only if they are a non-trivial linear function of at least one common exogenous variable E_m in the model. By Lemma 6, this happens if and only if V_i is d-connected to E_m given $\mathbf{Z}$ and Y is d-connected to E_m given V_i and $\mathbf{Z}$. If Y is d-connected to E_m given $\mathbf{Z}$ only, and since the concatenation of the (V_i, E_m) path with (Y, E_m) path must be by either colliding at the same child of E_m, or connected through some $V_x \leftarrow E_m \rightarrow V_y$, where V_x is in the path connected to V_i (which needs to be into V_x) and V_y is in the path connect to Y (which is into V_y), the theorem holds. If Y is not d-connected to E_m given $\mathbf{Z}$ only, then Y must be d-connected to V_i given $\mathbf{Z}$ by a path that is into V_i, and the claim again follows. □

Remarks: The assumptions are stronger than, for instance, the ones used in the proofs of Tashiro et al. (2014). A closely related result in that paper is its Lemma 2, a result identifying the dependence between the residual of the regression of a variable on its children. It does not use any variation of the faithfulness assumption. This is because, in their context, it is enough to detect the dependence between the residual and *some* children. So if some path cancellations take place, some other path cancellations cannot occur. But we need the dependence of our r_i and every relevant error term, because we cannot claim that r_i depends

on *some* error terms or latent variables, while r_y depends on *some* error terms or latent variables, if these two sets do not overlap. Although some of the ideas by Tashiro et al. (2014) could be used in our context to build partial models and from them deduce instrumental variables, it goes against our framework of solving a particular prediction problem (causal effect of a target treatment-outcome) directly, instead of doing it by recovering parts of a broader causal graph.

Finally, we have not provided an explicit discussion on how to validate the non-Gaussianity assumption by testing the non-Gaussianity of the residuals, as done by Entner et al. (2012). Or, more precisely, showing which assumptions are necessary so that testing non-Gaussianity of the residuals is equivalent to testing non-Gaussianity of the error terms. This is left as future work.