Joris M. Mooij,
Dominik Janzing,
Jonas Peters,
Tom Claassen,
Antti Hyttinen (Eds.)

Proceedings of the

UAI 2014 Workshop

Causal Inference: Learning and Prediction

Quebec City, Quebec, Canada

July 27, 2014

Preface

We are pleased to present the *Proceedings of the UAI 2014 Workshop on Causal Inference: Learning and Prediction*, held in Quebec City, Canada, on July 27, 2014, as a workshop of the 30th Conference on Uncertainty in Artificial Intelligence (UAI 2014). This workshop is the third in a series of UAI workshops on the topic of causality, following up on two successful predecessors, the *UAI Workshop on Causal Structure Learning 2012* and the *Approaches to Causal Structure Learning Workshop, UAI 2013*.

The aim of this workshop was to bring together researchers interested in the challenges of causal inference from observational and interventional data, especially when confounding variables, feedback loops or selection bias may be present. For this workshop, we decided to extend the scope from causal structure learning to include methods for making causal predictions, i.e., for predicting what happens under interventions. We especially encouraged contributions describing practical applications of causal methods.

There were 8 submissions, all full-length papers, each of which was peer-reviewed by two or three program committee members. We accepted five of these for oral presentation and for inclusion in these proceedings. The proceedings also include abstracts for three invited talks, including the two key-note talks by Robert Spekkens and Elias Bareinboim. Slides of most of the oral presentations are available on the workshop website:

`https://staff.fnwi.uva.nl/j.m.mooij/uai2014-causality-workshop/index.html`

We would like to thank the paper authors and presenters for their contributions and the program committee members for their reviewing service. We also appreciate the organizational support of the main UAI 2014 conference, in particular we would like to thank John Mark Agosta, Jin Tian and Ann Nicholson for their help. Further, we would like to thank Robin Evans, chair of the Approaches to Causal Structure Learning Workshop, UAI 2013, for his assistance. Finally, many thanks to the CEUR-WS team for hosting these proceedings.

October 2014

Joris M. Mooij (Chair)
Dominik Janzing
Jonas Peters
Tom Claassen
Antti Hyttinen

Organizing Committee

Joris M. Mooij	University of Amsterdam (Chair)
Dominik Janzing	Max Planck Institute for Intelligent Systems
Jonas Peters	ETH Zürich
Tom Claassen	Radboud University Nijmegen
Antti Hyttinen	California Institute of Technology

Program Committee

Thomas Richardson	University of Washington
Ricardo Silva	University College London
Markus Kalisch	ETH Zürich
Frederick Eberhardt	California Institute of Technology
Alain Hauser	ETH Zürich
Ilya Shpitser	University of Southampton
Robin Evans	University of Oxford
Kun Zhang	Max Planck Institute for Intelligent Systems
Eleni Sgouritsa	Max Planck Institute for Intelligent Systems
Aapo Hyvärinen	University of Helsinki
Jan Lemeire	Vrije Universiteit Brussel
James Robins	Harvard School of Public Health
Chris Meek	Microsoft Research
Preetam Nandy	ETH Zürich
Philipp Geiger	Max Planck Institute for Intelligent Systems
Nicholas Cornia	University of Amsterdam
Oliver Stegle	The European Bioinformatics Institute

Contents

How Occam's Razor Provides a Neat Definition of Direct Causation

Alexander Gebharter & Gerhard Schurz
Duesseldorf Center for Logic and Philosophy of Science
University of Duesseldorf
Universitaetsstrasse 1
40225 Duesseldorf, Germany

Abstract

In this paper we show that the application of Occam's razor to the theory of causal Bayes nets gives us a neat definition of direct causation. In particular we show that Occam's razor implies Woodward's (2003) definition of direct causation, provided suitable intervention variables exist and the causal Markov condition (CMC) is satisfied. We also show how Occam's razor can account for direct causal relationships Woodward style when only stochastic intervention variables are available.

1 INTRODUCTION

Occam's razor is typically seen as a methodological principle. There are many possible ways to apply the razor to the theory of causal Bayes nets. It could, for example, simply be interpreted to suggest preferring the simplest causal structure compatible with the given data among all compatible causal structures. The simplest causal structure could, for instance, be the one (or one of the ones) featuring the fewest causal arrows.

In this paper, however, we are interested in a slightly different application of Occam's razor: Our interpretation of Occam's razor asserts that given a causal structure is compatible with the data, it should only be chosen if it satisfies the causal minimality condition (Min) in the sense of Spirtes et al. (2000, p. 31), which requires that no causal arrow in the structure can be omitted in such a way that the resulting substructure would still be compatible with the data. When speaking of a causal structure being compatible with the data, we have a causal structure and a probability distribution satisfying the causal Markov condition (CMC) in mind. (For details, see sec. 5.) In the following, applying Occam's razor always means to assume that the causal minimality condition is satisfied.

In this paper we give a motivation for Occam's razor that goes beyond its merits as a methodological principle dictating that one should always decide in favor of minimal causal models. In particular, we show that Occam's razor provides a neat definition of direct causal relatedness in the sense of Woodward (2003), provided suitable intervention variables exist and CMC is satisfied. Note the connection of this enterprise to Zhang and Spirtes' (2011) project. Zhang and Spirtes prove that CMC and an interventionist definition of direct causation a la Woodward (2003) together imply minimality. So Occam's razor is well-motivated within a manipulationist framework such as Woodward's. We show, vice versa, that CMC and minimality together imply Woodward's definition of direct causation. So if one wants a neat definition of direct causation, it is reasonable to apply Occam's razor in the sense of assuming minimality.

The paper is structured as follows: In sec. 2 we introduce the notation we use in subsequent sections. In sec. 3 we present Woodward's (2003) definition of direct causation and his definition of an intervention variable. In sec. 4 we give precise reconstructions of both definitions in terms of causal Bayes nets. We also provide a definition of the notion of an intervention expansion, which is needed to account for direct causal relations in terms of the existence of certain intervention variables. In sec. 5 we show that Occam's razor gives us Woodward's definition of direct causation if CMC is assumed and the existence of suitable intervention variables is granted (theorem 2). In sec. 6 we go a step further and show how Occam's razor allows us to account for direct causation Woodward style when only stochastic intervention variables (cf. Korb et al., 2004, sec. 5) are available (theorem 3). We conclude in sec. 7.

Note that though the main results of the present paper (i.e., theorems 2 and 3) can be used for causal discovery, the goal of this paper is not to provide a method for uncovering direct causal connections among variables in a set of variables $\mathbf{V}$ of interest. The goal of this paper is to establish a connection between Woodward's (2003) intervention-based notion of direct causation and the presence of a causal arrow in a minimal causal Bayes net, which

can be interpreted as support for Occam's razor. Because of this, the present paper does not discuss the relation of theorems 2 and 3 to results about causal discovery by means of interventions such as, e.g., (Eberhardt and Scheines, 2007) or (Nyberg and Korb, 2007).

2 NOTATION

We represent causal structures by graphs, i.e., by ordered pairs $\langle \mathbf{V}, E\rangle$, where $\mathbf{V}$ is a set of variables and E is a binary relation on $\mathbf{V}$ ($E \subseteq \mathbf{V} \times \mathbf{V}$). $\mathbf{V}$'s elements are called the graph's "vertices" and E's elements are called its "edges". "$X \rightarrow Y$" stands short for "$\langle X, Y\rangle \in E$" and is interpreted as "X is a direct cause of Y in $\langle \mathbf{V}, E\rangle$" or as "$Y$ is a direct effect of X in $\langle \mathbf{V}, E\rangle$". $Par(Y)$ is the set of all $X \in \mathbf{V}$ with $X \rightarrow Y$ in $\langle \mathbf{V}, E\rangle$. The elements of $Par(Y)$ are called Y's parents. We write "$X - Y$" for "$X \rightarrow Y$ or $X \leftarrow Y$". A path $\pi : X - ... - Y$ is called a (causal) path connecting X and Y in $\langle \mathbf{V}, E\rangle$. A causal path π is called a directed causal path from X to Y if and only if ("iff" for short) it has the form $X \rightarrow ... \rightarrow Y$. X is called a cause of Y and Y an effect of X in that case. A causal path π is called a common cause path iff it has the form $X \leftarrow ... \leftarrow Z \rightarrow ... \rightarrow Y$ and no variable appears more often than once on π. Z is called a common cause of X and Y lying on path π in that case. A variable Z lying on a path $\pi : X - ... \rightarrow Z \leftarrow ... - Y$ is called a collider lying on this path. A variable X is called exogenous iff no arrow is pointing at X; it is called endogenous otherwise.

A graph $\langle \mathbf{V}, E\rangle$ is called a directed graph in case all edges in E are one-headed arrows "$\rightarrow$". It is called cyclic iff it features a causal path of the form $X \rightarrow ... \rightarrow X$ and acyclic otherwise. A causal structure $\langle \mathbf{V}, E\rangle$ together with a probability distribution P over $\mathbf{V}$ is called a causal model $\langle \mathbf{V}, E, P\rangle$. P is intended to provide information about the strengths of causal influences represented by the arrows in $\langle \mathbf{V}, E\rangle$. A causal model $\langle \mathbf{V}, E, P\rangle$ is called cyclic iff its graph $\langle \mathbf{V}, E\rangle$ is cyclic; it is called acyclic otherwise. In the following, we will only be interested in acyclic causal models.

We use the standard notions of (conditional) probabilistic dependence and independence:

Definition 1 (conditional probabilistic (in)dependence) *X and Y are probabilistically dependent conditional on Z iff there are X-, Y-, and Z-values x, y, and z, respectively, such that $P(x|y,z) \neq P(x|z) \wedge P(y,z) > 0$.*

X and Y are probabilistically independent conditional on Z iff X and Y are not probabilistically dependent conditional on Z.

Probabilistic independence between X and Y conditional on Z is abbreviated as "$Indep(X, Y|Z)$", probabilistic dependence is abbreviated as "$Dep(X, Y|Z)$". Unconditional probabilistic (in)dependence between X and Y $(In)Dep(X, Y)$ is defined as $(In)Dep(X, Y|\emptyset)$. X, Y, and Z in definition 1 can be variables or sequences of variables. When $X, Y, Z, ...$ are sequences of variables, we write them in bold letters. We write also the values $\mathbf{x}, \mathbf{y}, \mathbf{z}, ...$ of sequences $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, ...$ in bold letters. The set of values $\mathbf{x}$ of a sequence $\mathbf{X}$ of variables $X_1, ..., X_n$ is $val(X_1) \times ... \times val(X_n)$, where $val(X_i)$ is the set of X_i's possible values.

3 WOODWARD'S DEFINITION OF DIRECT CAUSATION

Woodward's (2003) interventionist theory of causation aims to explicate direct causation w.r.t. a set of variables $\mathbf{V}$ in terms of possible interventions. Woodward (2003, p. 98) provides the following definition of an intervention variable:

Definition 2 ($\mathbf{IV}_W$) *I is an intervention variable for X with respect to Y if and only if I meets the following conditions:*
I1. I causes X.
I2. I acts as a switch for all the other variables that cause X. That is, certain values of I are such that when I attains those values, X ceases to depend on the values of other variables that cause X and instead depends only on the value taken by I.
I3. Any directed path from I to Y [if there exists one] goes through X [...].
I4. I is (statistically) independent of any variable Z that causes Y and that is on a directed path that does not go through X.

($\mathbf{IV}_W$) is intended to single out those variables as intervention variables for X w.r.t. Y that allow for correct causal inference according to Woodward's (2003) definition of direct causation. For I to be an intervention variable for X w.r.t. Y it is required that I is causally relevant to X (condition I1), that X is only under I's influence when $I = on$ (condition I2), and that a correlation between I and Y can only be due to a directed causal path from I to Y going through X (conditions I3 and I4). For a detailed motivation of I1-I4, see (Woodward, 2003, sec. 3.1.4). For problems with Woodward's definitions, see (Gebharter and Schurz, ms).

An intervention on X w.r.t. Y (from now on we refer to X as the intervention's "target variable" and to Y as the "test variable") is then straightforwardly defined as an intervention variable I for X w.r.t. Y taking one of its on-values, which forces X to take a certain value x. We will call interventions whose on-values force X to take certain values x "deterministic interventions" (cf. Korb et al., 2004, sec. 5).

Note that Woodward's (2003) notion of an intervention is, on the one hand, strong because it requires interventions to be deterministic interventions. It is, on the other hand, weak in another respect: In contrast to structural or surgical interventions (cf. Eberhardt and Scheines, 2007, p. 984; Pearl, 2009) Woodward's interventions are allowed to be direct causes of more than one variable as long as the intervention's direct effects which are non-target variables do not cause the test variable over a path not going through the intervention's target variable (intervention condition I3).

Based on his notion of an intervention, Woodward (2003, p. 59) gives the following definition of direct causation w.r.t. a variable set $\mathbf{V}$:

Definition 3 ($\mathbf{DC}_W$) *A necessary and sufficient condition for X to be a (type-level) direct cause of Y with respect to a variable set $\mathbf{V}$ is that there be a possible intervention on X that will change Y or the probability distribution of Y when one holds fixed at some value all other variables Z_i in $\mathbf{V}$.*

($\mathbf{DC}_W$) neatly explicates direct causation w.r.t. a variable set $\mathbf{V}$ in terms of possible interventions: X is a direct cause of Y w.r.t. $\mathbf{V}$ if Y can be wiggled by wiggling X; and if X is a direct cause of Y w.r.t. $\mathbf{V}$, then there are possible interventions by whose means one can influence Y by manipulating X.[1]

Note that **($\mathbf{DC}_W$)** may be too strong because many domains involve variables one cannot control by deterministic interventions. Scenarios of this kind include, for example, the decay of uranium or states of entangled systems in quantum mechanics. The decay of uranium can only be probabilistically influenced, and any attempt to manipulate the state of one of two entangled photons, for example, would destroy the entangled system. Glymour (2004) also considers variables for sex and race as not manipulable by means of intervention variables in the sense of **($\mathbf{IV}_W$)**.

To avoid all problems that might arise for Woodward's (2003) account due to variables that are not manipulable by deterministic interventions, we will reconstruct Woodward's **($\mathbf{DC}_W$)** as a partial definition in sec. 4. In particular, we will define direct causation only for sets of variables $\mathbf{V}$ for which suitable intervention variables exist.

4 RECONSTRUCTING WOODWARD'S DEFINITION

In this section we reconstruct Woodward's (2003) definition of direct causation in terms of causal Bayes nets. The reconstruction of **($\mathbf{IV}_W$)** is straightforward:

[1]Note that Woodward (2003) does not require the intervention variables I to be elements of the set of variables $\mathbf{V}$ containing the target variable X and the test variable Y.

Definition 4 (IV) *$I_X \in \mathbf{V}$ is an intervention variable for $X \in \mathbf{V}$ w.r.t. $Y \in \mathbf{V}$ in a causal model $\langle \mathbf{V}, E, P\rangle$ iff*
(a) I_X is exogenous and there is a path $\pi : I_X \rightarrow X$ in $\langle \mathbf{V}, E\rangle$,
(b) for every on-value of I_X there is an X-value x such that $P(x|I_X = on) = 1$ and $Dep(x, I_X = on|\mathbf{z})$ holds for every instantiation $\mathbf{z}$ of every $\mathbf{Z} \subseteq \mathbf{V}\backslash\{I_X, X\}$,
(c) all paths $I_X \rightarrow ... \rightarrow Y$ in $\langle \mathbf{V}, E\rangle$ have the form $I_X \rightarrow ... \rightarrow X \rightarrow ... \rightarrow Y$,
(d) I_X is independent from every variable C (in $\mathbf{V}$ or not in $\mathbf{V}$) which causes Y over a path not going through X.

Note that **(IV)** still allows for intervention variables I_X that are common causes of their target variable X and other variables in $\mathbf{V}$. Condition (a) requires I_X to be exogenous. This is, though it is a typical assumption made for intervention variables, not explicit in Woodward's (2003) original definition **($\mathbf{IV}_W$)**. One problem that might arise for Woodward's account when not making this assumption is that I_X in a causal structure $Y \rightarrow I_X \rightarrow X$ may turn out to be an intervention variable for X w.r.t. Y. If Y then depends on $I_X = on$, **($\mathbf{DC}_W$)** would falsely determine X to be a cause of Y (cf. Gebharter and Schurz, ms). $I_X \rightarrow X$ in condition (a) is a harmless simplification of I1. Condition (b) captures Woodward's requirement that interventions have to be deterministic, from which I2 follows. X is assumed to be under full control of I_X when I_X is on. This does not only require that for every on-value of I_X there is an X-value x such that $P(x|I_X = on) = 1$, but also that $I_X = on$ actually has an influence on x in every possible context, i.e., under conditionalization on arbitrary instantiations $\mathbf{z}$ of all kinds of subsets $\mathbf{Z}$ of $\mathbf{V}\backslash\{I_X, X\}$. Condition (c) directly mirrors I3. Condition (d) mirrors Woodward's I4. Note that condition (d) requires reference to variables C possibly not contained in $\mathbf{V}$ (cf. Woodward, 2008, p. 202).

If we want to account for direct causal connection in a causal model $\langle \mathbf{V}, E, P\rangle$ by means of interventions, we have to add intervention variables to $\mathbf{V}$. In other words: We have to expand $\langle \mathbf{V}, E, P\rangle$ in a certain way. But how do we have to expand $\langle \mathbf{V}, E, P\rangle$? To answer this question, let us assume that we want to know whether X is a direct cause of Y in the unmanipulated model $\langle \mathbf{V}, E, P\rangle$. Then the manipulated model $\langle \mathbf{V}', E', P'\rangle$ will have to contain an intervention variable I_X for X w.r.t. Y and also intervention variables I_Z for all $Z \in \mathbf{V}$ different from X and Y by whose means these Z can be controlled. X is a direct cause of Y if I_X has some on-values such that we can influence Y by manipulating X with $I_X = on$ when all I_Z have taken certain on-values. On the other hand, to guarantee that X is not a direct cause of Y, we have to demonstrate that no one of Y's values can be influenced by manipulating some X-value by some intervention. For establishing such a negative causal claim, we require an intervention variable I_X by whose means we can control every X-value x. (Otherwise it could be that Y depends only on X-values that

are not correlated with I_X-values; then $I_X = on$ would have no probabilistic influence on Y, though X may be a causal parent of Y.) In addition, we require for every $Z \neq X, Y$ an intervention variable I_Z by whose means Z can be forced to take every value z. (Otherwise it could be that we can bring about only such Z-value instantiations which screen X and Y off each other; then $I_X = on$ would have no probabilistic influence on Y when Z's value is fixed by interventions, though X may be a causal parent of Y.)

In the unmanipulated model $\langle \mathbf{V}, E, P \rangle$, all intervention variables I are off. In the manipulated model $\langle \mathbf{V}', E', P' \rangle$, all intervention variables' values are realized for some but not for all individuals in the domain. This move allows us to compute probabilities for variables in $\mathbf{V}$ when $I = off$ as well as probabilities for variables in $\mathbf{V}$ for all combinations of on-value realizations of intervention variables I, while the causal structure of the unmanipulated model will be preserved in the manipulated model. (Note that we deviate here from the typical "arrow breaking" representation of interventions in the literature which assumes that in the manipulated model all individuals get manipulated.) This amounts to the following notion of an intervention expansion ("i-expansion" for short):

Definition 5 (intervention expansion) *$\langle \mathbf{V}', E', P' \rangle$ is an intervention expansion of $\langle \mathbf{V}, E, P \rangle$ w.r.t. $Y \in \mathbf{V}$ iff*
(a) $\mathbf{V}' = \mathbf{V} \dot{\cup} \mathbf{V_I}$, where $\mathbf{V_I}$ contains for every $X \in \mathbf{V}$ different from Y an intervention variable I_X w.r.t. Y (and nothing else),
(b) for all $Z_i, Z_j \in \mathbf{V}: Z_i \rightarrow Z_j$ in E' iff $Z_i \rightarrow Z_j$ in E,
(c) for every X-value x of every $X \in \mathbf{V}$ different from Y there is an on-value of the corresponding intervention variable I_X such that $P'(x|I_X = on) = 1$ and $Dep(x, I_X = on|\mathbf{z})$ holds for every instantiation $\mathbf{z}$ of every $\mathbf{Z} \subseteq \mathbf{V} \backslash \{I_X, X\}$,
(d) $P'_{\mathbf{I=off}} \uparrow \mathbf{V} = P$,
(e) $P'(\mathbf{I} = \mathbf{on}), P'(\mathbf{I} = \mathbf{off}) > 0$.

$\mathbf{I}$ in conditions (d) and (e) is the set of all newly added intervention variables I. $P'_{\mathbf{I=off}} \uparrow \mathbf{V}$ in (d) is $P'_{\mathbf{I=off}} := P'(-|\mathbf{I} = \mathbf{off})$ restricted to $\mathbf{V}$. Hence, "$P'_{\mathbf{I=off}} \uparrow \mathbf{V} = P$" means that $P'_{\mathbf{I=off}}$ coincides with P on the value space of variables in $\mathbf{V}$. Condition (a) guarantees that the i-expansion contains all the intervention variables required for testing for direct causal relationships in the sense of Woodward's (2003) definition of direct causation. The assumption that $\mathbf{V_I}$ contains only intervention variables for X w.r.t. Y is a harmless simplification. Thanks to condition (b), the manipulated model's causal structure fits to the unmanipulated model's causal structure. In particular, the i-expansion is only allowed to introduce new causal arrows going from intervention variables to variables in $\mathbf{V}$. Due to condition (c), every $X \in \mathbf{V}$ different from Y can be fully controlled by means of an intervention variable I_X for X w.r.t. Y. Condition (d) explains how the manipulated model's associated probability distribution P' fits to the unmanipulated model's distribution P. Condition (e) says that all values of intervention variables have to be realized by some individuals in the domain.

With help of this notion of an i-expansion we can now reconstruct Woodward's (2003) definition of direct causation. As already mentioned, Woodward's definition requires the existence of suitable intervention variables. Thus, we reconstruct **($\mathbf{DC}_W$)** as a partial definition whose if-condition presupposes the required intervention variables:

Definition 6 (DC) *If there exist i-expansions $\langle \mathbf{V}', E', P' \rangle$ of $\langle \mathbf{V}, E, P \rangle$ w.r.t. $Y \in \mathbf{V}$, then: $X \in \mathbf{V}$ is a direct cause of Y w.r.t. $\mathbf{V}$ iff $Dep(Y, I_X = on|\mathbf{I_Z} = \mathbf{on})$ holds in some i-expansions $\langle \mathbf{V}', E', P' \rangle$ of $\langle \mathbf{V}, E, P \rangle$ w.r.t. Y, where I_X is an intervention variable for X w.r.t. Y in $\langle \mathbf{V}', E', P' \rangle$ and $\mathbf{I_Z}$ is the set of all intervention variables in $\langle \mathbf{V}', E', P' \rangle$ different from I_X.*

(DC) mirrors Woodward's definition restricted to cases in which the required intervention variables (more precisely: the required i-expansions) exist: In case Y can be probabilistically influenced by manipulating X by means of an intervention variable I_X for X w.r.t. Y in one of these i-expansions, X is a direct cause of Y in the unmanipulated model. And vice versa: In case X is a direct cause of Y in the unmanipulated model, there will be an intervention variable I_X for X w.r.t. Y in one of these i-expansions such that Y is probabilistically sensitive to $I_X = on$.

In the next section we show that **(DC)** can account for all direct causal dependencies in a causal model if suitable i-expansions exist and CMC and Min are assumed to be satisfied.

5 OCCAM'S RAZOR, DETERMINISTIC INTERVENTIONS, AND DIRECT CAUSATION

The theory of causal Bayes nets' core axiom is the causal Markov condition (CMC) (cf. Spirtes et al., 2000, p. 29):

Definition 7 (causal Markov condition) *A causal model $\langle \mathbf{V}, E, P \rangle$ satisfies the causal Markov condition iff every $X \in \mathbf{V}$ is probabilistically independent of all its non-effects conditional on its causal parents.*

CMC is assumed to hold for causal models whose variable sets are causally sufficient. A variable set $\mathbf{V}$ is causally sufficient iff every common cause C of variables X and Y in $\mathbf{V}$ is also in $\mathbf{V}$ or takes the same value c for all individuals in the domain (cf. Spirtes et al., 2000, p. 22). From now on we implicitly assume causal sufficiency, i.e., we only consider causal models whose variable sets are causally sufficient.

A finite causal model $\langle \mathbf{V}, E, P \rangle$ satisfies the Markov condition iff P admits the following Markov factorization relative to $\langle \mathbf{V}, E \rangle$ (cf. Pearl, 2009, p. 16):

$$P(X_1, ..., X_n) = \prod_i P(X_i | Par(X_i)) \qquad (1)$$

The conditional probabilities $P(X_i|Par(X_i))$ are called X_i's parameters.

For acyclic causal models, CMC is equivalent to the d-separation criterion (Verma, 1986; Pearl, 1988, pp. 119f):

Definition 8 (d-separation criterion) *$\langle \mathbf{V}, E, P \rangle$ satisfies the d-separation criterion iff the following holds for all $X, Y \in \mathbf{V}$ and $\mathbf{Z} \subseteq \mathbf{V} \backslash \{X, Y\}$: If X and Y are d-separated by $\mathbf{Z}$ in $\langle \mathbf{V}, E \rangle$, then $Indep(X, Y|\mathbf{Z})$.*

Definition 9 (d-separation, d-connection) *$X \in \mathbf{V}$ and $Y \in \mathbf{V}$ are d-separated by $\mathbf{Z} \subseteq \mathbf{V} \backslash \{X, Y\}$ in $\langle \mathbf{V}, E \rangle$ iff X and Y are not d-connected given $\mathbf{Z}$ in $\langle \mathbf{V}, E \rangle$.*

$X \in \mathbf{V}$ and $Y \in \mathbf{V}$ are d-connected given $\mathbf{Z} \subseteq \mathbf{V} \backslash \{X, Y\}$ in $\langle \mathbf{V}, E \rangle$ iff X and Y are connected by a path π in $\langle \mathbf{V}, E \rangle$ such that no non-collider on π is in $\mathbf{Z}$, while all colliders on π are in $\mathbf{Z}$ or have an effect in $\mathbf{Z}$.

The equivalence between CMC and the d-separation criterion reveals the full content of CMC: If a causal model satisfies CMC, then every (conditional) probabilistic independence can be explained by missing (conditional) causal connections, and every (conditional) probabilistic dependence can be explained by some existing (conditional) causal connection.

In case there is a path π between X and Y in $\langle \mathbf{V}, E \rangle$ such that no non-collider on π is in $\mathbf{Z} \subseteq \mathbf{V} \backslash \{X, Y\}$ and all colliders on π are in $\mathbf{Z}$ or have an effect in $\mathbf{Z}$, π is said to be activated by $\mathbf{Z}$. We also say that X and Y are d-connected given $\mathbf{Z}$ over path π in that case. If π is not activated by $\mathbf{Z}$, π is said to be blocked by $\mathbf{Z}$. We also say that X and Y are d-separated by $\mathbf{Z}$ over path π in that case.

Occam's razor (as we understand it in this paper) dictates to prefer from all those causal structures $\langle \mathbf{V}, E \rangle$, which together with a given probability distribution P over $\mathbf{V}$ satisfy CMC, the ones which also satisfy the causal minimality condition (Min):

Definition 10 (causal minimality condition) *A causal model $\langle \mathbf{V}, E, P \rangle$ satisfying CMC satisfies the causal minimality condition iff no model $\langle \mathbf{V}, E', P \rangle$ with $E' \subset E$ also satisfies CMC (cf. Spirtes et al., 2000, p. 31).*

For acyclic causal models satisfying CMC, the following causal productivity condition (Prod) (cf. Schurz and Gebharter, forthcoming) can be seen as a reformulation of the causal minimality condition:

Definition 11 (causal productivity condition) *A causal model $\langle \mathbf{V}, E, P \rangle$ satisfies the causal productivity condition iff $Dep(X, Y|Par(Y)\backslash\{X\})$ holds for all $X, Y \in \mathbf{V}$ with $X \to Y$ in $\langle \mathbf{V}, E \rangle$.*

Theorem 1 *For every acyclic causal model $\langle \mathbf{V}, E, P \rangle$ satisfying CMC, the causal minimality condition and the causal productivity condition are equivalent.*

The equivalence of Min and Prod reveals the full content of Min: In minimal causal models, no causal arrow is superfluous, i.e., every causal arrow from X to Y is productive, meaning that it is responsible for some probabilistic dependence between X and Y (when the values of all other parents of Y are fixed).

We can now prove the following theorem:

Theorem 2 *If $\langle \mathbf{V}, E, P \rangle$ is an acyclic causal model and for every $Y \in \mathbf{V}$ there is an i-expansion $\langle \mathbf{V}', E', P' \rangle$ of $\langle \mathbf{V}, E, P \rangle$ w.r.t. Y satisfying CMC and Min, then for all $X, Y \in \mathbf{V}$ (with $X \neq Y$) the following two statements are equivalent:*
(i) $X \to Y$ in $\langle \mathbf{V}, E \rangle$.
(ii) $Dep(Y, I_X = on|\mathbf{I_Z} = \mathbf{on})$ holds in some i-expansions $\langle \mathbf{V}', E', P' \rangle$ of $\langle \mathbf{V}, E, P \rangle$ w.r.t. Y, where I_X is an intervention variable for X w.r.t. Y in $\langle \mathbf{V}', E', P' \rangle$ and $\mathbf{I_Z}$ is the set of all intervention variables in $\langle \mathbf{V}', E', P' \rangle$ different from I_X.

Theorem 2 shows that direct causation a la Woodward (2003) coincides with the graph theoretical notion of direct causation in systems $\langle \mathbf{V}, E, P \rangle$ with i-expansions w.r.t. every variable $Y \in \mathbf{V}$ satisfying CMC and Min. In particular, theorem 2 says the following: Assume we are interested in a causal model $\langle \mathbf{V}, E, P \rangle$. Assume further that for every Y in $\mathbf{V}$ there is an i-expansion $\langle \mathbf{V}', E', P' \rangle$ of $\langle \mathbf{V}, E, P \rangle$ w.r.t. Y satisfying CMC and Min. This means (among other things) that for every pair of variables $\langle X, Y \rangle$ there is at least one i-expansion with an intervention variable I_X for X w.r.t. Y and intervention variables I_Z for every $Z \in \mathbf{V}$ (different from X and Y) w.r.t. Y by whose means one can force the variables in $\mathbf{V}\backslash\{Y\}$ to take any combination of value realizations. Given this setup, theorem 2 tells us for every X and Y (with $X \neq Y$) in $\mathbf{V}$ that X is a causal parent of Y in $\langle \mathbf{V}, E \rangle$ iff $Dep(Y, I_X = on|\mathbf{I_Z} = \mathbf{on})$ holds in one of the presupposed i-expansions w.r.t. Y.

6 OCCAM'S RAZOR, STOCHASTIC INTERVENTIONS, AND DIRECT CAUSATION

In this section we generalize the main finding of sec. 5 to cases in which only stochastic interventions are available. To account for direct causal relations $X \to Y$ by means of stochastic intervention variables, two intervention vari-

ables are needed, one for X and one for Y. (For details, see below.) We define a stochastic intervention variable as follows:

Definition 12 ($\mathbf{IV}_S$) *$I_X \in \mathbf{V}$ is a stochastic intervention variable for $X \in \mathbf{V}$ w.r.t. $Y \in \mathbf{V}$ in $\langle \mathbf{V}, E, P\rangle$ iff*
(a) I_X is exogenous and there is a path $\pi : I_X \rightarrow X$ in $\langle \mathbf{V}, E\rangle$,
(b) for every on-value of I_X there is an X-value x such that $Dep(x, I_X = on|\mathbf{z})$ holds for every instantiation $\mathbf{z}$ of every $\mathbf{Z} \subseteq \mathbf{V}\backslash\{I_X, X\}$,
(c) all paths $I_X \rightarrow ... \rightarrow Y$ in $\langle \mathbf{V}, E\rangle$ have the form $I_X \rightarrow ... \rightarrow X \rightarrow ... \rightarrow Y$,
(d) I_X is independent from every variable C (in $\mathbf{V}$ or not in $\mathbf{V}$) which causes Y over a path not going through X.

The only difference between **($\mathbf{IV}_S$)** and **(IV)** is condition (b). For stochastic interventions it is not required that $I_X = on$ determines X's value to be x with probability 1. It suffices that $I_X = on$ and x are correlated conditional on every value $\mathbf{z}$ of every $\mathbf{Z} \subseteq \mathbf{V}\backslash\{I_X, X\}$. This specific constraint guarantees that X can be influenced by $I_X = on$ under all circumstances, i.e., under all kinds of conditionalization on instantiations of remainder variables in $\mathbf{V}$.

We do also have to modify our notion of an intervention expansion in case we allow for stochastic interventions. We define the following notion of a stochastic intervention expansion:

Definition 13 (stochastic intervention expansion)
$\langle \mathbf{V}', E', P'\rangle$ is a stochastic intervention expansion of $\langle \mathbf{V}, E, P\rangle$ for $X \in \mathbf{V}$ w.r.t. $Y \in \mathbf{V}$ iff
(a) $\mathbf{V}' = \mathbf{V} \dot{\cup} \mathbf{V_I}$, where $\mathbf{V_I}$ contains one stochastic intervention variable I_X for X w.r.t. Y and one stochastic intervention variable I_Y for Y w.r.t. Y which is a parent only of Y (and nothing else),
(b) for all $Z_i, Z_j \in \mathbf{V} : Z_i \rightarrow Z_j$ in E' iff $Z_i \rightarrow Z_j$ in E,
(c.1) for every X-value x there is an on-value of I_X such that $Dep(x, I_X = on|\mathbf{z})$ holds for every instantiation $\mathbf{z}$ of every $\mathbf{Z} \subseteq \mathbf{V}'\backslash\{I_X, X\}$,
(c.2) for every Y-value y, every instantiation $\mathbf{r}$ of $Par(Y)$, and every on-value of I_Y there is an on-value on^* of I_Y such that $P'(y|I_Y = on^*, \mathbf{r}) \neq P'(y|I_Y = on, \mathbf{r})$, $P'(y|I_Y = on^*, \mathbf{r}) > 0$, and $P'(y|I_Y = on^*, \mathbf{r}^*) = P'(y|I_Y = on, \mathbf{r}^*)$ holds for all $\mathbf{r}^* \in val(Par(Y))$ different from $\mathbf{r}$,
(d) $P'_{\mathbf{I}=\mathbf{off}} \uparrow \mathbf{V} = P$,
(e) $P'(\mathbf{I} = \mathbf{on}), P'(\mathbf{I} = \mathbf{off}) > 0$.

This definition differs from the definition of a (non-stochastic) i-expansion with respect to conditions (a) and (c): A stochastic i-expansion for X w.r.t. Y contains exactly two intervention variables, viz. one stochastic intervention variable I_X for X w.r.t. Y and one stochastic intervention variable I_Y for Y w.r.t. Y (which trivially satisfies conditions (c) and (d) in **($\mathbf{IV}_S$)**). While I_X may have more than one direct effect, the second intervention variable I_Y is assumed to be a causal parent only of Y. (This is required for accounting for direct causal connections; for details see (i) $\Rightarrow$ (ii) in the proof of theorem 3 in the appendix.)

The second intervention variable I_Y is required to exclude independence between I_X and Y due to a fine-tuning of Y's parameters. Such an independence can arise even if CMC and Min are satisfied, X is a causal parent of Y, and I_X and Y are each correlated with the same X-values x. For examples of this kind of non-faithfulness, see, e.g., (Neapolitan, 2004, p. 96) or (Naeger, forthcoming). In condition (c.2) we assume that every one of Y's parameters can be changed independently of all other Y-parameters (to a value $r \in]0, 1]$) by changing I_Y's on-value. This suffices to exclude non-faithful independencies between I_X and Y of the kind described above.

When not presupposing deterministic interventions, it cannot be guaranteed anymore that the value of every variable in our model of interest different from the test variable Y can be fixed by interventions. The values of a causal model's variables can, however, also be fixed by conditionalization. To account for direct causation between X and Y when only stochastic interventions are available, one has to conditionalize on a suitably chosen set $\mathbf{Z} \subseteq \mathbf{V}\backslash\{X, Y\}$ that (i) blocks all indirect causal paths between X and Y, and that (ii) fixes all X-alternative parents of Y. That $\mathbf{Z}$ blocks all indirect paths between X and Y is required to assure that dependence between $I_X = on$ and Y cannot be due to an indirect path, and fixing the values of all parents of Y different from X is required to exclude independence of $I_X = on$ and Y due to a fine-tuning of Y's X-alternative parents that may cancel the influence of $I_X = on$ on Y over a path $I_X \rightarrow X \rightarrow Y$.[2] Fortunately, every directed acyclic graph $\langle \mathbf{V}, E\rangle$ features a set $\mathbf{Z}$ satisfying requirement (i), viz. $Par(Y)\backslash\{X\}$ (cf. Schurz and Gebharter, forthcoming). Trivially, $Par(Y)\backslash\{X\}$ also satisfies requirement (ii).

With the help of **($\mathbf{IV}_S$)** and definition 13, we can now define direct causation in terms of stochastic interventions for models for which suitable stochastic i-expansions exist:

Definition 14 ($\mathbf{DC}_S$) *If there exist stochastic i-expansions $\langle \mathbf{V}', E', P'\rangle$ of $\langle \mathbf{V}, E, P\rangle$ for X w.r.t. Y, then: X is a direct cause of Y w.r.t. $\mathbf{V}$ iff $Dep(Y, I_X = on|Par(Y)\backslash\{X\}, I_Y = on)$ holds in some i-expansions $\langle \mathbf{V}', E', P'\rangle$ of $\langle \mathbf{V}, E, P\rangle$ for X w.r.t. Y, where I_X is a stochastic intervention variable for X w.r.t. Y in $\langle \mathbf{V}', E', P'\rangle$ and I_Y is a stochastic intervention variable for Y w.r.t. Y in $\langle \mathbf{V}', E', P'\rangle$.*

Now the following theorem can be proven:

[2] For details on such cases of non-faithfulness due to compensating parents see (Schurz and Gebharter, forthcoming; Pearl, 1988, p. 256).

Theorem 3 *If $\langle \mathbf{V}, E, P\rangle$ is an acyclic causal model and for every $X, Y \in \mathbf{V}$ (with $X \neq Y$) there is a stochastic i-expansion $\langle \mathbf{V}', E', P'\rangle$ of $\langle \mathbf{V}, E, P\rangle$ for X w.r.t. Y satisfying CMC and Min, then for all $X, Y \in \mathbf{V}$ (with $X \neq Y$) the following two statements are equivalent:*
(i) $X \rightarrow Y$ in $\langle \mathbf{V}, E\rangle$.
(ii) $Dep(Y, I_X = on | Par(Y)\backslash\{X\}, I_Y = on)$ holds in some i-expansions $\langle \mathbf{V}', E', P'\rangle$ of $\langle \mathbf{V}, E, P\rangle$ for X w.r.t. Y, where I_X is a stochastic intervention variable for X w.r.t. Y in $\langle \mathbf{V}', E', P'\rangle$ and I_Y is a stochastic intervention variable for Y w.r.t. Y in $\langle \mathbf{V}', E', P'\rangle$.

Theorem 3 shows that direct causation a la Woodward (2003) coincides with the graph theoretical notion of direct causation in systems $\langle \mathbf{V}, E, P\rangle$ with stochastic i-expansions for every $X \in \mathbf{V}$ w.r.t. every $Y \in \mathbf{V}$ (with $X \neq Y$) satisfying CMC and Min. In particular, theorem 3 says the following: Assume we are interested in a causal model $\langle \mathbf{V}, E, P\rangle$. Assume further that for every X, Y in $\mathbf{V}$ (with $X \neq Y$) there is a stochastic i-expansion $\langle \mathbf{V}', E', P'\rangle$ of $\langle \mathbf{V}, E, P\rangle$ for X w.r.t. Y satisfying CMC and Min. This means (among other things) that for every pair of variables $\langle X, Y\rangle$ there is at least one stochastic i-expansion featuring a stochastic intervention variable I_X for X w.r.t. Y and a stochastic intervention variable I_Y for Y w.r.t. Y. Given this setup, theorem 3 can account for every causal arrow between every X and Y (with $X \neq Y$) in $\mathbf{V}$: It says that X is a causal parent of Y in $\langle \mathbf{V}, E\rangle$ iff $Dep(Y, I_X = on | Par(Y)\backslash\{X\}, I_Y = on)$ holds in some of the presupposed stochastic i-expansions for X w.r.t. Y.

7 CONCLUSION

In this paper we investigated the consequences of assuming a certain version of Occam's razor. If one applies the razor in such a way to the theory of causal Bayes nets that it dictates to prefer only minimal causal models, one can show that Occam's razor provides a neat definition of direct causation. In particular, we demonstrated that one gets Woodward's (2003) definition of direct causation translated into causal Bayes nets terminology and restricted to contexts in which suitable i-expansions satisfying the causal Markov condition (CMC) exist. In the last section we showed how Occam's razor can be used to account for direct causal connections Woodward style even if no deterministic interventions are available. These results can be seen as a motivation of Occam's razor going beyond its merits as a methodological principle: If one wants a nice and simple interventionist definition of direct causation in the sense of Woodward (or its stochastic counterpart developed in sec. 6), then it is reasonable to apply a version of Occam's razor that suggests to eliminate non-minimal causal models.

Acknowledgements

This work was supported by DFG, research unit "Causation, Laws, Dispositions, Explanation" (FOR 1063). Our thanks go to Frederick Eberhardt and Paul Naeger for important discussions, to two anonymous referees for helpful comments on an earlier version of the paper, and to Sebastian Maaß for proofreading.

References

F. Eberhardt, and R. Scheines (2007). Interventions and causal inference. *Philosophy of Science* **74**(5):981-995.

A. Gebharter, and G. Schurz (ms). Woodward's interventionist theory of causation: Problems and proposed solutions.

C. Glymour (2004). Critical notice. *British Journal for the Philosophy of Science* **55**(4):779-790.

K. B. Korb, L. R. Hope, A. E. Nicholson, and K. Axnick (2004). Varieties of causal intervention. In C. Zhang, H. W. Guesgen, W.-K. Yeap (eds.), *Proceedings of the 8th Pacific Rim International Conference on AI 2004: Trends in Artificial Intelligence*, 322-331. Berlin: Springer.

P. Naeger (forthcoming). The causal problem of entanglement. *Synthese*.

R. Neapolitan (2004). *Learning Bayesian Networks*. Upper Saddle River, NJ: Prentice Hall.

E. P. Nyberg, and K. B. Korb (2006). Informative interventions. Technical report 2006/204, Clayton School of Information Technology, Monash University, Melbourne.

J. Pearl (1988). *Probabilistic Reasoning in Expert Systems*. San Mateo, MA: Morgan Kaufmann.

J. Pearl (2009). *Causality*. Cambridge: Cambridge University Press.

G. Schurz, and A. Gebharter (forthcoming). Causality as a theoretical concept: Explanatory warrant and empirical content of the theory of causal nets. *Synthese*.

P. Spirtes, C. Glymour, and R. Scheines (2000). *Causation, Prediction, and Search*. Cambridge, MA: MIT Press.

T. S. Verma (1986). Causal networks: Semantics and expressiveness. Technical report R-65, Cognitive Systems Laboratory, University of California, Los Angeles.

J. Woodward (2003). *Making Things Happen*. Oxford: Oxford University Press.

J. Woodward (2008). Response to Strevens. *Philosophy and Phenomenological Research* **77**(1):193-212.

J. Zhang, and P. Spirtes (2011). Intervention, determinism, and the causal minimality condition. *Synthese* **182**(3):335-347.

Appendix

The following proof of theorem 1 rests on the equivalence of CMC and the Markov factorization (1). It is, thus, restricted to finite causal structures.

Proof of theorem 1 Suppose $\langle \mathbf{V}, E, P\rangle$ with $\mathbf{V} = \{X_1, ..., X_n\}$ to be a finite acyclic causal model satisfying CMC.

Prod ⇒ Min: Assume that $\langle \mathbf{V}, E, P\rangle$ does not satisfy Min, meaning that there are $X, Y \in \mathbf{V}$ with $X \rightarrow Y$ in $\langle \mathbf{V}, E\rangle$ such that $\langle \mathbf{V}, E', P\rangle$, which results from deleting $X \rightarrow Y$ from $\langle \mathbf{V}, E\rangle$, still satisfies CMC. But then $Par(Y)\backslash\{X\}$ d-separates X and Y in $\langle \mathbf{V}, E'\rangle$, and thus, the d-separation criterion implies $Indep(X, Y|Par(Y)\backslash\{X\})$, which violates Prod.

Min ⇒ Prod: Assume that $\langle \mathbf{V}, E, P\rangle$ satisfies Min, meaning that there are no $X, Y \in \mathbf{V}$ with $X \rightarrow Y$ in $\langle \mathbf{V}, E\rangle$ such that $\langle \mathbf{V}, E', P\rangle$, which results from deleting $X \rightarrow Y$ from $\langle \mathbf{V}, E\rangle$, still satisfies CMC. The latter is the case iff (*) the parent set $Par(Y)$ of every $Y \in \mathbf{V}$ (with $Par(Y) \neq \emptyset$) is minimal in the sense that removing one of Y's parents X from $Par(Y)$ would make a difference for Y, meaning that $P(y|x, Par(Y)\backslash\{X\} = \mathbf{r}) \neq P(y|Par(Y)\backslash\{X\} = \mathbf{r})$ holds for some X-values x, some Y-values y, and some instantiations $\mathbf{r}$ of $Par(Y)\backslash\{X\}$. Otherwise P would admit the Markov factorization relative to $\langle \mathbf{V}, E\rangle$ and relative to $\langle \mathbf{V}, E'\rangle$, meaning that also $\langle \mathbf{V}, E', P\rangle$, which results from deleting $X \rightarrow Y$ from $\langle \mathbf{V}, E\rangle$, would satisfy CMC. But then $\langle \mathbf{V}, E, P\rangle$ would not be minimal, which would contradict the assumption. Now (*) entails that $Dep(X, Y|Par(Y)\backslash\{X\})$ holds for all $X, Y \in \mathbf{V}$ with $X \rightarrow Y$, i.e., that $\langle \mathbf{V}, E, P\rangle$ satisfies Prod. □

Proof of theorem 2 Assume $\langle \mathbf{V}, E, P\rangle$ is an acyclic causal model and for every $Y \in \mathbf{V}$ there is an i-expansion $\langle \mathbf{V}', E', P'\rangle$ of $\langle \mathbf{V}, E, P\rangle$ w.r.t. Y satisfying CMC and Min. Let X and Y be arbitrarily chosen elements of $\mathbf{V}$ such that $X \neq Y$.

(i) ⇒ (ii): Suppose $X \rightarrow Y$ in $\langle \mathbf{V}, E\rangle$. We assumed that there exists an i-expansion $\langle \mathbf{V}', E', P'\rangle$ of $\langle \mathbf{V}, E, P\rangle$ w.r.t. Y satisfying CMC and Min. From condition (b) of definition 5 it follows that $X \rightarrow Y$ in $\langle \mathbf{V}', E'\rangle$. Since Min is equivalent to Prod, X and Y are dependent when the values of all parents of Y different from X are fixed to certain values, meaning that there will be an X-value x and a Y-value y such that $Dep(x, y|Par(Y)\backslash\{X\} = \mathbf{r})$ holds for an instantiation $\mathbf{r}$ of $Par(Y)\backslash\{X\}$. Now there will also be a value of $\mathbf{I_Z}$ that fixes the set of all parents of Y different from X to $\mathbf{r}$. Let $\mathbf{on}$ be this $\mathbf{I_Z}$-value. Thus, also $Dep(x, y|\mathbf{I_Z} = \mathbf{on})$ and also $Dep(x, y|\mathbf{I_Z} = \mathbf{on}, \mathbf{r})$ will hold. Now let us assume that on is one of the I_X-values which are correlated with x and which force X to take value x. (The existence of such an I_X-value is guaranteed by condition (c) in definition 5.) Then we have $Dep(I_X = on, x|\mathbf{I_Z} = \mathbf{on}, \mathbf{r}) \wedge Dep(x, y|\mathbf{I_Z} = \mathbf{on}, \mathbf{r})$. From the axiom of weak union (2) (cf. Pearl, 2009, p. 11), which is probabilistically valid, we get (3) and (4) (in which $\mathbf{s} = \langle x, \mathbf{r}\rangle$ is a value realization of $Par(Y)$):

$$Indep(X, YW|Z) \Rightarrow Indep(X, Y|ZW) \tag{2}$$

$$\begin{aligned} &Indep(I_X = on, \mathbf{s} = \langle x, \mathbf{r}\rangle|\mathbf{I_Z} = \mathbf{on}) \Rightarrow \\ &Indep(I_X = on, x|\mathbf{I_Z} = \mathbf{on}, \mathbf{r}) \end{aligned} \tag{3}$$

$$\begin{aligned} &Indep(\mathbf{s} = \langle x, \mathbf{r}\rangle, y|\mathbf{I_Z} = \mathbf{on}) \Rightarrow \\ &Indep(x, y|\mathbf{I_Z} = \mathbf{on}, \mathbf{r}) \end{aligned} \tag{4}$$

With the contrapositions of (3) and (4) it now follows that $Dep(I_X = on, \mathbf{s} = \langle x, \mathbf{r}\rangle|\mathbf{I_Z} = \mathbf{on}) \wedge Dep(\mathbf{s} = \langle x, \mathbf{r}\rangle, y|\mathbf{I_Z} = \mathbf{on})$.

We now show that $Dep(I_X = on, \mathbf{s}|\mathbf{I_Z} = \mathbf{on}) \wedge Dep(\mathbf{s}, y|\mathbf{I_Z} = \mathbf{on})$ and the d-separation criterion imply $Dep(I_X = on, y|\mathbf{I_Z} = \mathbf{on})$. We define $P^*(-)$ as $P'(-|\mathbf{I_Z} = \mathbf{on})$ and proceed as follows:

$$\begin{aligned} &P^*(y|I_X = on) = \\ &\sum_i P^*(y|\mathbf{s}_i, I_X = on) \cdot P^*(\mathbf{s}_i|I_X = on) \end{aligned} \tag{5}$$

Equation (5) is probabilistically valid. Because $Par(Y)$ blocks all paths between I_X and Y, we get (6) from (5):

$$\begin{aligned} &P^*(y|I_X = on) = \\ &\sum_i P^*(y|\mathbf{s}_i) \cdot P^*(\mathbf{s}_i|I_X = on) \end{aligned} \tag{6}$$

Since $I_X = on$ forces $Par(Y)$ to take value $\mathbf{s}$ when $\mathbf{I_Z} = \mathbf{on}$, $P^*(\mathbf{s}_i|I_X = on) = 1$ in case $\mathbf{s}_i = \mathbf{s}$, and $P^*(\mathbf{s}_i|I_X = on) = 0$ otherwise. Thus, we get (7) from (6):

$$P^*(y|I_X = on) = P^*(y|\mathbf{s}) \cdot 1 \tag{7}$$

For reductio, let us assume that $Indep(I_X = on, y|\mathbf{I_Z} = \mathbf{on})$, meaning that $P^*(y|I_X = on) = P^*(y)$. But then we get (8) from (7):

$$P^*(y) = P^*(y|\mathbf{s}) \cdot 1 \tag{8}$$

Equation (8) contradicts $Dep(\mathbf{s}, y|\mathbf{I_Z} = \mathbf{on})$ above. Hence, $Dep(I_X = on, y|\mathbf{I_Z} = \mathbf{on})$ has to hold when $Dep(I_X = on, \mathbf{s}|\mathbf{I_Z} = \mathbf{on}) \wedge Dep(\mathbf{s}, y|\mathbf{I_Z} = \mathbf{on})$ holds. Therefore, $Dep(Y, I_X = on|\mathbf{I_Z} = \mathbf{on})$.

(ii) ⇒ (i): Suppose $\langle \mathbf{V}', E', P'\rangle$ is one of the presupposed i-expansions such that $Dep(Y, I_X = on|\mathbf{I_Z} = \mathbf{on})$ holds, where I_X is an intervention variable for X w.r.t. Y in $\langle \mathbf{V}', E', P'\rangle$ and $\mathbf{I_Z}$ is the set of all intervention variables in $\langle \mathbf{V}', E', P'\rangle$ different from I_X. Then the d-separation criterion implies that there must be a causal path π d-connecting I_X and Y. π cannot be a path featuring colliders, because I_X and Y would be d-separated over such

a path. π also cannot have the form $I_X \leftarrow ... - Y$. This is excluded by condition (a) in **(IV)**. So π must have the form $I_X \rightarrow ... - Y$. Since π cannot feature colliders, π must be a directed path $I_X \rightarrow ... \rightarrow Y$. Now either (A) π goes through X, or (B) π does not go through X. (B) is excluded by condition (c) in **(IV)**. Hence, (A) must be the case. If (A) is the case, then π is a directed path $I_X \rightarrow ... \rightarrow X \rightarrow ... \rightarrow Y$ going through X. Now there are two possible cases: Either (i) at least one of the paths π d-connecting I_X and Y has the form $I_X \rightarrow ... \rightarrow X \rightarrow Y$, or (ii) all paths π d-connecting I_X and Y have the form $I_X \rightarrow ... \rightarrow X \rightarrow ... \rightarrow C \rightarrow ... \rightarrow Y$.

Assume (ii) is the case, i.e., all paths π d-connecting I_X and Y have the form $I_X \rightarrow ... \rightarrow X \rightarrow ... \rightarrow C \rightarrow ... \rightarrow Y$. Let $\mathbf{r}_i$ be an individual variable ranging over $val(Par(Y))$. We define $P^*(-)$ as $P'(-|\mathbf{I_Z} = \mathbf{on})$ and proceed as follows:

$$P^*(y|I_X = on) = \sum_i P^*(y|\mathbf{r}_i, I_X = on) \cdot P^*(\mathbf{r}_i|I_X = on) \tag{9}$$

$$P^*(y) = \sum_i P^*(y|\mathbf{r}_i) \cdot P^*(\mathbf{r}_i) \tag{10}$$

Equations (9) and (10) are probabilistically valid. Since $\mathbf{I_Z} = \mathbf{on}$ forces every non-intervention variable in $\mathbf{V}'$ different from X and Y to take a certain value, $\mathbf{I_Z} = \mathbf{on}$ will also force $Par(Y)$ to take a certain value $\mathbf{r}$, meaning that $P^*(\mathbf{r}_i) = 1$ in case $\mathbf{r}_i = \mathbf{r}$, and that $P^*(\mathbf{r}_i) = 0$ otherwise. Since probabilities of 1 do not change after conditionalization, we get $P^*(\mathbf{r}_i|I_X = on) = 1$ in case $\mathbf{r}_i = \mathbf{r}$, and $P^*(\mathbf{r}_i|I_X = on) = 0$ otherwise. Thus, we get (11) from (9) and (12) from (10):

$$P^*(y|I_X = on) = P^*(y|\mathbf{r}, I_X = on) \cdot 1 \tag{11}$$

$$P^*(y) = P^*(y|\mathbf{r}) \cdot 1 \tag{12}$$

Since $Par(Y)$ blocks all paths between I_X and Y, we get $P^*(y|\mathbf{r}, I_X = on) = P^*(y|\mathbf{r})$ with the d-separation criterion, and thus, we get $P^*(y|I_X = on) = P^*(y)$ with (11) and (12). Thus, $Indep(Y, I_X = on|\mathbf{I_Z} = \mathbf{on})$ holds, which contradicts the initial assumption that $Dep(Y, I_X = on|\mathbf{I_Z} = \mathbf{on})$ holds. Therefore, (i) must be the case, i.e., there must be a path π d-connecting I_X and Y that has the form $I_X \rightarrow ... \rightarrow X \rightarrow Y$. From $\langle \mathbf{V}', E', P' \rangle$ being an i-expansion of $\langle \mathbf{V}, E, P \rangle$ it now follows that $X \rightarrow Y$ in $\langle \mathbf{V}, E \rangle$. □

Proof of theorem 3 Assume $\langle \mathbf{V}, E, P \rangle$ is an acyclic causal model and for every $X, Y \in \mathbf{V}$ (with $X \neq Y$) there is a stochastic i-expansion $\langle \mathbf{V}', E', P' \rangle$ of $\langle \mathbf{V}, E, P \rangle$ for X w.r.t. Y satisfying CMC and Min. Let X and Y be arbitrarily chosen elements of $\mathbf{V}$ such that $X \neq Y$.

(i) ⇒ (ii): Suppose $X \rightarrow Y$ in $\langle \mathbf{V}, E \rangle$. We assumed that there exists a stochastic i-expansion $\langle \mathbf{V}', E', P' \rangle$ of $\langle \mathbf{V}, E, P \rangle$ for X w.r.t. Y satisfying CMC and Min. From condition (b) of definition 13 it follows that $X \rightarrow Y$ in $\langle \mathbf{V}', E' \rangle$. Since Min is equivalent to Prod, $Dep(x, y|Par(Y)\backslash\{X\} = \mathbf{r}, I_Y = on)$ holds for some X-values x, for some Y-values y, for some of I_Y's on-values on, and for some instantiations $\mathbf{r}$ of $Par(Y)\backslash\{X\}$. Now let us assume that on is one of the I_X-values which are correlated with x conditional on $Par(Y)\backslash\{X\} = \mathbf{r}, I_Y = on$. (The existence of such an I_X-value on is guaranteed by condition (c.1) in definition 13.) Then we have $Dep(I_X = on, x|\mathbf{r}, I_Y = on) \wedge Dep(x, y|\mathbf{r}, I_Y = on)$.

We now show that $Dep(I_X = on, x|\mathbf{r}, I_Y = on) \wedge Dep(x, y|\mathbf{r}, I_Y = on)$ together with $I_X \rightarrow X \rightarrow Y$ and the d-separation criterion implies $Dep(I_X = on, y|\mathbf{r}, I_Y = on)$. We define $P^*(-)$ as $P'(-|\mathbf{r})$ and proceed as follows:

$$P^*(y|I_X = on, I_Y = on) = \sum_i P^*(y|x_i, I_X = on, I_Y = on) \cdot P^*(x_i|I_X = on, I_Y = on) \tag{13}$$

$$P^*(y|I_Y = on) = \sum_i P^*(y|x_i, I_Y = on) \cdot P^*(x_i|I_Y = on) \tag{14}$$

Equations (13) and (14) are probabilistically valid. From $I_X \rightarrow X \rightarrow Y$ and (13) we get with the d-separation criterion:

$$P^*(y|I_X = on, I_Y = on) = \sum_i P^*(y|x_i, I_Y = on) \cdot P^*(x_i|I_X = on, I_Y = on) \tag{15}$$

Since I_Y is exogenous and a causal parent only of Y, X and I_Y are d-separated by I_X, and thus, we get (16) from (15) with the d-separation criterion. Since I_Y and X are d-separated (by the empty set), we get (17) from (14) with the d-separation criterion:

$$P^*(y|I_X = on, I_Y = on) = \sum_i P^*(y|x_i, I_Y = on) \cdot P^*(x_i|I_X = on) \tag{16}$$

$$P^*(y|I_Y = on) = \sum_i P^*(y|x_i, I_Y = on) \cdot P^*(x_i) \tag{17}$$

Now either (A) $P^*(y|I_X = on, I_Y = on) \neq P^*(y|I_Y = on)$, or (B) $P^*(y|I_X = on, I_Y = on) = P^*(y|I_Y = on)$. If (A) is the case, then $Dep(Y, I_X = on|Par(Y)\backslash\{X\}, I_Y = on)$.

If (B) is the case, then $P^*(y|I_X = on, I_Y = on)$ can only equal $P^*(y|I_Y = on)$ due to a fine-tuning of $P^*(x_i|I_Y = on)$ and $P^*(x_i)$ in equations (16) and (17), respectively. We already know that X's value x and

$I_X = on$ are dependent conditional on $Par(Y)\backslash\{X\} = \mathbf{r}, I_Y = on$, meaning that $P^*(x|I_X = on, I_Y = on) \neq P^*(x|I_Y = on)$ holds. Since X and I_Y are d-separated by I_X, $P^*(x|I_X = on, I_Y = on) = P^*(x|I_X = on)$ holds. Since X and I_Y are d-separeted (by the empty set), $P^*(x|I_Y = on) = P^*(x)$ holds. It follows that $P^*(x|I_X = on) \neq P^*(x)$ holds. So (i) $P^*(x|I_X = on) > 0$ or (ii) $P^*(x) > 0$. Thanks to condition (c.2) in definition 13, every one of the conditional probabilities $P^*(y|x_i, I_Y = on)$ can be changed independently by replacing "on" in "$P^*(y|x_i, I_Y = on)$" by some I_Y-value "on^*" (with $on^* \neq on$) such that $P^*(y|x_i, I_Y = on^*) > 0$. Thus, in both cases ((i) and (ii)) it holds that $P^*(y|x, I_Y = on^*) \cdot P^*(x|I_X = on^*) \neq P^*(y|x, I_Y = on^*) \cdot P^*(x)$, while $P^*(y|x_i, I_Y = on^*) \cdot P^*(x_i|I_X = on^*) = P^*(y|x_i, I_Y = on^*) \cdot P^*(x_i)$ holds for all $x_i \neq x$. It follows that $P^*(y|I_X = on, I_Y = on^*) \neq P^*(y|I_Y = on^*)$.

(ii) ⇒ (i): Suppose $\langle \mathbf{V}', E', P' \rangle$ is one of the above assumed stochastic i-expansions for X w.r.t. Y and that $Dep(Y, I_X = on|Par(Y)\backslash\{X\}, I_Y = on)$ holds in this stochastic i-expansion. The d-separation criterion and $Dep(Y, I_X = on|Par(Y)\backslash\{X\}, I_Y = on)$ imply that I_X and Y are d-connected given $(Par(Y)\backslash\{X\}) \cup \{I_Y\}$ by a causal path $\pi : I_X - ... - Y$. π cannot have the form $I_X \leftarrow ... - Y$. This is excluded by condition (a) in **(IV$_S$)**. Thus, π must have the form $I_X \rightarrow ... - Y$. Now either (A) π goes through X, or (B) π does not go through X.

Suppose (B) is the case. Then, because of condition (c) in **(IV$_S$)**, π cannot be a directed path $I_X \rightarrow ... \rightarrow Y$. Thus, π must either (i) have the form $I_X \rightarrow ... - C \rightarrow Y$ (with a collider on π), or it (ii) must have the form $I_X \rightarrow ... - C \leftarrow Y$. If (i) is the case, then C must be in $(Par(Y)\backslash\{X\}) \cup \{I_Y\}$ (since C cannot be X). Hence, π would be blocked by $(Par(Y)\backslash\{X\}) \cup \{I_Y\}$ and, thus, would not d-connect I_X and Y given $(Par(Y)\backslash\{X\}) \cup \{I_Y\}$. Thus, (ii) must be the case. If (ii) is the case, then there has to be a collider C^* on π that either is C or that is an effect of C, and thus, also an effect of Y. But then I_X and Y can only be d-connected given $(Par(Y)\backslash\{X\}) \cup \{I_Y\}$ over π if C^* is in $(Par(Y)\backslash\{X\}) \cup \{I_Y\}$ or has an effect in $(Par(Y)\backslash\{X\}) \cup \{I_Y\}$. But this would mean that Y is a cause of Y, what is excluded by the initial assumption of acyclicity. Thus, (A) has to be the case.

If (A) is the case, then π must have the form $I_X \rightarrow ... - X - ... - Y$. If π would have the form $I_X \rightarrow ... - X - ... - C \leftarrow Y$ (where C and X are possibly identical), then there is at least one collider C^* lying on π that is an effect of Y. For I_X and Y to be d-connected given $(Par(Y)\backslash\{X\}) \cup \{I_Y\}$ over path π, $(Par(Y)\backslash\{X\}) \cup \{I_Y\}$ must activate π, meaning that C^* has to be in $(Par(Y)\backslash\{X\}) \cup \{I_Y\}$ or has to have an effect in $(Par(Y)\backslash\{X\}) \cup \{I_Y\}$. But then we would end up with a causal cycle $Y \rightarrow ... \rightarrow Y$, which would contradict the assumption of acyclicity. Hence, π must have the form $I_X \rightarrow ... - X - ... - C \rightarrow Y$ (where C and X are possibly identical). Now either (i) $C = X$ or (ii) $C \neq X$. If (ii) is the case, then $C \in (Par(Y)\backslash\{X\}) \cup \{I_Y\}$, and thus, $(Par(Y)\backslash\{X\}) \cup \{I_Y\}$ blocks π. But then I_X and Y cannot be d-connected given $(Par(Y)\backslash\{X\}) \cup \{I_Y\}$ over path π. Hence, (i) must be the case. Then π has the form $I_X \rightarrow ... - X \rightarrow Y$ and from $\langle \mathbf{V}', E', P' \rangle$ being a stochastic i-expansion of $\langle \mathbf{V}, E, P \rangle$ it follows that $X \rightarrow Y$ in $\langle \mathbf{V}, E \rangle$. □

Constructing Separators and Adjustment Sets in Ancestral Graphs

Benito van der Zander, Maciej Liśkiewicz
Theoretical Computer Science
University of Lübeck, Germany
{benito,liskiewi}@tcs.uni-luebeck.de

Johannes Textor
Theoretical Biology & Bioinformatics
Utrecht University, The Netherlands
johannes.textor@gmx.de

Abstract

Ancestral graphs (AGs) are graphical causal models that can represent uncertainty about the presence of latent confounders, and can be inferred from data. Here, we present an algorithmic framework for efficiently testing, constructing, and enumerating m-separators in AGs. Moreover, we present a new constructive criterion for covariate adjustment in directed acyclic graphs (DAGs) and maximal ancestral graphs (MAGs) that characterizes adjustment sets as m-separators in a subgraph. Jointly, these results allow to find all adjustment sets that can identify a desired causal effect with multivariate exposures and outcomes in the presence of latent confounding. Our results generalize and improve upon several existing solutions for special cases of these problems.

1 INTRODUCTION

Graphical causal models endow researchers with a language to codify assumptions about a data generating process (Pearl, 2009; Elwert, 2013). Using graphical criteria, one can asses whether the assumptions encoded in such a model allow estimation of a causal effect from observational data, which is a key issue in Epidemiology (Rothman et al., 2008), the Social Sciences (Elwert, 2013) and other fields where controlled experimentation is typically impossible. Specifically, the famous back-door criterion by Pearl (2009) can identify cases where causal effect identification is possible by standard covariate adjustment, and other methods like the front-door criterion or do-calculus can even permit identification even if the back-door criterion fails (Pearl, 2009). In current practice, however, covariate adjustment is highly preferred to such alternatives because its statistical properties are well understood, giving access to useful methodology like robust estimators and confidence intervals. In contrast, knowledge about the statistical properties of e.g. front-door estimation is still considerably lacking (VanderWeele, 2009; Glynn and Kashin, 2013)[1]. Unfortunately, the back-door criterion is not complete, i.e., it does not find all possible options for covariate adjustment that are allowed by a given graphical causal model.

In this paper, we aim to efficiently find a definitive answer for the following question: Given a causal graph $\mathcal{G}$, which covariates $\mathbf{Z}$ do we need to adjust for to estimate the causal effect of the exposures $\mathbf{X}$ on the outcomes $\mathbf{Y}$? To our knowledge, no efficient algorithm has been shown to answer this question, not even when $\mathcal{G}$ is a directed acyclic graph (DAG), though constructive solutions do exist for special cases like singleton $\mathbf{X} = \{X\}$ (Pearl, 2009), and a subclass of DAGs (Textor and Liśkiewicz, 2011). Here, we provide algorithms for adjustment sets in DAGs as well as in maximal ancestral graphs (MAGs), which extend DAGs allowing to account for unspecified latent variables. Our algorithms are guaranteed to find all valid adjustment sets for a given DAG or MAG with polynomial delay, and we also provide variants to list only those sets that minimize a user-supplied cost function or to quickly construct a simple adjustment set if one exists. Modelling multiple, possibly interrelated exposures $\mathbf{X}$ is important e.g. in case-control studies that screen several putative causes of a disease (Greenland, 1994). Likewise, the presence of unspecified latent variables often cannot be excluded in real-world settings, and the causal structure between the observed variables may not be completely known. We hope that the ability to quickly deduce from a given DAG or MAG whether and how covariate adjustment can render a causal effect identifiable will benefit researchers in such areas.

We have two main contributions. First, in Section 3, we present algorithms for verifying, constructing, and listing m-separating sets in AGs. This subsumes a number of earlier solutions for special cases of these problems, e.g.

[1]Quoting VanderWeele (2009), "Time will perhaps tell whether results like Pearl's front-door path adjustment theorem and its generalizations are actually useful for epidemiologic research or whether the results are simply of theoretical interest."

the Bayes-Ball algorithm for verification of d-separating sets (Shachter, 1998), the use of network flow calculations to find minimal d-separating sets in DAGs (Tian et al., 1998; Acid and de Campos, 2003), and an algorithm to list minimal adjustment sets for a certain subclass of DAGs (Textor and Liśkiewicz, 2011). Our verification and construction algorithms for single separators are asymptotically runtime-optimal. Although we apply our algorithms only to adjustment set construction, they are likely useful in other settings as separating sets are involved in most graphical criteria for causal effect identification. Moreover, the separators themselves constitute statistically testable implications of the causal assumptions encoded in the graph.

Second, we give a graphical criterion that characterizes adjustment sets in terms of separating sets, and is sound and complete for DAGs and MAGs without selection variables. This generalizes the sound and complete criterion for DAGs by Shpitser et al. (2010), and the sound but incomplete adjustment criterion for MAGs without selection variables by Maathuis and Colombo (2013). Our criterion exhaustively addresses adjustment set construction in the presence of latent covariates and with incomplete knowledge of causal structure if at least a MAG can be specified. We give the criterion separately for DAGs (Section 4) and MAGs (Section 5) because the same graph usually admits more adjustment options if viewed as a DAG than if viewed as a MAG.

2 PRELIMINARIES

We denote sets by bold upper case letters ($\mathbf{S}$), and sometimes abbreviate singleton sets as $\{S\} = S$. Graphs are written calligraphically ($\mathcal{G}$), and variables in upper-case (X).

Mixed graphs and paths. We consider mixed graphs $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ with nodes (vertices, variables) $\mathbf{V}$ and directed ($A \to B$), undirected ($A - B$), and bidirected ($A \leftrightarrow B$) edges $\mathbf{E}$. Nodes linked by an edge are *adjacent*. A *walk* of length n is a node sequence $V_1, \ldots, V_{n+1}$ such that there exists an edge sequence $E_1, E_2, \ldots, E_n$ for which every edge E_i connects V_i, V_{i+1}. Then V_1 is called the *start node* and V_{n+1} the *end node* of the walk. A *path* is a walk in which no node occurs more than once. Given a node set $\mathbf{X}$ and a node set $\mathbf{Y}$, a walk from $X \in \mathbf{X}$ to $Y \in \mathbf{Y}$ is called *proper* if only its start node is in $\mathbf{X}$. Given a graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ and a node set $\mathbf{V}'$, the *induced subgraph* $\mathcal{G}_{\mathbf{V}'} = (\mathbf{V}', \mathbf{E}')$ contains the edges $\mathbf{E}'$ from $\mathcal{G}$ that are adjacent only to nodes in $\mathbf{V}'$.

Ancestry. A walk of the form $V_1 \to \ldots \to V_n$ is *directed*, or *causal*. If there is a directed walk from U to V, then U is called an *ancestor* of V and V a *descendant* of U. A graph is *acyclic* if no directed walk from a node to itself is longer than 0. All directed walks in an acyclic graph are paths. A walk is *anterior* if it were directed after replacing all edges $U - V$ by $U \to V$. If there is an anterior path from U to V, then U is called an *anterior* of V. All ancestors of V are anteriors of V. Every node is its own ancestor, descendant, and anterior. For a node set $\mathbf{X}$, the set of all of its ancestors is written as $An(\mathbf{X})$. The descendant and anterior sets $De(\mathbf{X}), Ant(\mathbf{X})$ are analogously defined. Also, we denote by $Pa(\mathbf{X})$, ($Ch(\mathbf{X})$), the set of parents (children) of $\mathbf{X}$.

m-Separation. A node V on a walk w is called a *collider* if two arrowheads of w meet at V, e.g. if w contains $U \leftrightarrow V \leftarrow Q$. There can be no collider if w is shorter than 2. Two nodes U, V are called *collider connected* if there is a path between them on which all nodes except U and V are colliders. Adjacent vertices are collider connected. Two nodes U, V are called *m-connected* by a set $\mathbf{Z}$ if there is a path π between them on which every node that is a collider is in $An(\mathbf{Z})$ and every node that is not a collider is not in $\mathbf{Z}$. Then π is called an m-connecting path. The same definition can be stated simpler using walks: U, V are called m-connected by $\mathbf{Z}$ if there is a walk between them on which all colliders and only colliders are in $\mathbf{Z}$. If U, V are m-connected by the empty set, we simply say they are m-connected. If U, V are not m-connected by $\mathbf{Z}$, we say that $\mathbf{Z}$ *m-separates* them or *blocks* all paths between them. Two node sets $\mathbf{X}, \mathbf{Y}$ are m-separated by $\mathbf{Z}$ if all their nodes are pairwise m-separated by $\mathbf{Z}$. In DAGs, m-separation is equivalent to the well-known d-separation criterion (Pearl, 2009).

Ancestral graphs and DAGs. A mixed graph $\mathcal{G} = (V, E)$ is called an *ancestral graph* (AG) if the following two conditions hold: (1) For each edge $A \leftarrow B$ or $A \leftrightarrow B$, A is not an ancestor of B. (2) For each edge $A - B$, there are no edges $A \leftarrow C$, $A \leftrightarrow C$, $B \leftarrow C$ or $B \leftrightarrow C$. There can be at most one edge between two nodes in an AG (Richardson and Spirtes, 2002). Syntactically, all DAGs are AGs and all AGs containing only directed edges are DAGs. An AG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ is a *maximal ancestral graph* (MAG) if every non-adjacent pair of nodes U, V can be m-separated by some $\mathbf{Z} \subseteq \mathbf{V} \setminus \{U, V\}$. Every AG $\mathcal{G}$ can be turned into a MAG $\mathcal{M}$ by adding bidirected edges between node pairs that cannot be m-separated (Richardson and Spirtes, 2002).

3 ALGORITHMS FOR M-SEPARATION

In this section, we compile an algorithmic framework for solving a host of problems related to verification, construction, and enumeration of m-separating sets in AGs. The problems are defined in Fig. 1, which also shows the asymptotic runtime of their solutions. Throughout, n stands for the number of nodes and m for the number of edges in a graph. All of these problems except ListSep can be solved by rather straightforward modifications of existing algorithms (Acid and de Campos, 1996; Shachter, 1998; Tian et al., 1998; Textor and Liśkiewicz, 2011).

Pseudocodes of these algorithms are shown for reference and implementation in the Appendix of this paper, as are proof details omitted from the main text.

An important tool for solving similar problems for d-separation is *moralization*, by which d-separation can be reduced to a vertex cut in an undirected graph. This reduction allows to solve problems like FINDMINSEP using standard network flow algorithms (Acid and de Campos, 1996). Moralization can be generalized to AGs in the following manner.

Definition 3.1 (Moralization of AGs (Richardson and Spirtes, 2002))**.** *Given an AG $\mathcal{G}$, the augmented graph $(\mathcal{G})^a$ is an undirected graph with the same node set as $\mathcal{G}$ such that $X - Y$ is an edge in $(\mathcal{G})^a$ if and only if X and Y are collider connected in $\mathcal{G}$.*

Theorem 3.2 (Reduction of m-Separation to vertex cuts (Richardson and Spirtes, 2002))**.** *Given an AG $\mathcal{G}$ and three node sets $\mathbf{X}$,$\mathbf{Y}$ and $\mathbf{Z}$, $\mathbf{Z}$ m-separates $\mathbf{X}$ and $\mathbf{Y}$ if and only if $\mathbf{Z}$ is an $\mathbf{X}$-$\mathbf{Y}$ node cut in $(\mathcal{G}_{Ant(\mathbf{X}\cup\mathbf{Y}\cup\mathbf{Z})})^a$.*

A direct implementation of Definition 3.1 would lead to a suboptimal algorithm. Therefore, we first give an asymptotically optimal (linear time in output size) moralization algorithm for AGs. We then solve TESTMINSEP, FINDMINSEP, FINDMINCOSTSEP and LISTMINSEP by generalizing existing correctness proofs of the moralization approach for d-separation (Tian et al., 1998).

Not all our solutions are based on moralization, however. Moralization takes time $O(n^2)$, and TESTSEP and FINDSEP can be solved faster, i.e. in asymptotically optimal time $O(n + m)$.

Lemma 3.3 (Efficient AG moralization)**.** *Given an AG $\mathcal{G}$, the augmented graph $(\mathcal{G})^a$ can be computed in time $O(n^2)$.*

Proof. The algorithm proceeds in four steps. (1) Start by setting $(\mathcal{G})^a$ to $\mathcal{G}$ replacing all edges by undirected ones. (2) Identify all connected components in $\mathcal{G}$ with respect to bidirected edges (two nodes are in the same such component if they are connected by a path consisting only of bidirected edges). Nodes without adjacent bidirected edges form singleton components. (3) For each pair U, V of nodes from the same component, add the edge $U - V$ to $(\mathcal{G})^a$ if it did not exist already. (4) For each component, identify all its parents (nodes U with an edge $U \to V$ where U is in the component) and link them all by undirected edges in $(\mathcal{G})^a$. Now two nodes are adjacent in $(\mathcal{G})^a$ if and only if they are collider connected in $\mathcal{G}$. All four steps can be performed in time $O(n^2)$. □

Lemma 3.4. *Let $\mathbf{X}, \mathbf{Y}, \mathbf{I}, \mathbf{R}$ be sets of nodes with $\mathbf{I} \subseteq \mathbf{R}$, $\mathbf{R} \cap (\mathbf{X} \cup \mathbf{Y}) = \emptyset$. If there exists an m-separator $\mathbf{Z}_0$, with $\mathbf{I} \subseteq \mathbf{Z}_0 \subseteq \mathbf{R}$ then $\mathbf{Z} = Ant(\mathbf{X}\cup\mathbf{Y}\cup\mathbf{I})\cap\mathbf{R}$ is an m-separator.*

Corollary 3.5 (Ancestry of minimal separators)**.** *Given an AG $\mathcal{G}$, and three sets $\mathbf{X}, \mathbf{Y}, \mathbf{I}$, every minimal set $\mathbf{Z}$ over all m-separators containing $\mathbf{I}$ is a subset of $Ant(\mathbf{X} \cup \mathbf{Y} \cup \mathbf{I})$.*

Proof. Assume there is a minimal separator $\mathbf{Z}$ with $\mathbf{Z} \not\subseteq Ant(\mathbf{X} \cup \mathbf{Y} \cup \mathbf{I})$. According to Lemma 3.4 we have that $\mathbf{Z}' = Ant(\mathbf{X} \cup \mathbf{Y} \cup \mathbf{I}) \cap \mathbf{Z}$ is a separator with $\mathbf{I} \subseteq \mathbf{Z}'$. But $\mathbf{Z}' \subseteq Ant(\mathbf{X} \cup \mathbf{Y} \cup \mathbf{I})$ and $\mathbf{Z}' \subseteq \mathbf{Z}$, so $\mathbf{Z} \neq \mathbf{Z}'$ and $\mathbf{Z}$ is not a minimal separator. □

Corollary 3.5 applies to minimum-cost separators as well because every minimum-cost separator must be minimal. Now we can solve FINDMINCOSTSEP and FINDMINSIZESEP by using weighted min-cut, which takes time $O(n^3)$ using practical algorithms, and LISTMINSEP by using Takata's algorithm to enumerate minimal vertex cuts with delay $O(n^3)$ (Takata, 2010).

However, for FINDMINSEP and TESTMINSEP, we can do better than using standard vertex cuts.

Proposition 3.6. *The task* FINDMINSEP *can be solved in time $O(n^2)$.*

Proof. Two algorithms are given in the appendix, one with runtime $O(nm)$ (Algorithm 8) and one with runtime $O(n^2)$ (Algorithm 9). □

Corollary 3.7. *The task* TESTMINSEP *can be solved in time $O(n^2)$.*

Proof. First verify whether $\mathbf{Z}$ is an m-separator using moralization. If not, return "no". Otherwise, set $S = \mathbf{Z}$ and solve FINDMINSEP. Return "yes" if the output is $\mathbf{Z}$ and "no", otherwise. □

Moralization can in the worst case quadratically increase the size of a graph. Therefore, in some cases, it may be preferable to avoid moralization if the task at hand is rather simple, as are the two tasks considered below.

Proposition 3.8. *The task* FINDSEP *can be solved in time $O(n + m)$.*

Proof. This follows directly from Lemma 3.4, and the fact that the set $Ant(\mathbf{X} \cup \mathbf{Y} \cup \mathbf{I}) \cap \mathbf{R}$ can be found in linear time from the MAG without moralization. Note that unlike in DAGs, two non-adjacent nodes cannot always be m-separated in ancestral graphs. □

By modifying the Bayes-Ball algorithm (Shachter, 1998) appropriately, we get the following.

Proposition 3.9. *The task* TESTSEP *can be solved in time $O(n + m)$.*

Lastly, we consider the problem of listing *all* m-separators. Here is an algorithm to solve that problem with polynomial delay.

Verification: For given $\mathbf{X}, \mathbf{Y}$ and $\mathbf{Z}$ decide if ...		
TESTSEP	$\mathbf{Z}$ m-separates $\mathbf{X}, \mathbf{Y}$	$O(n+m)$
TESTMINSEP	$\mathbf{Z}$ m-separates $\mathbf{X}, \mathbf{Y}$ but no $\mathbf{Z}' \subsetneq \mathbf{Z}$ does	$O(n^2)$
Construction: For given $\mathbf{X}, \mathbf{Y}$ and auxiliary $\mathbf{I}, \mathbf{R}$, output ...		
FINDSEP	an m-separator $\mathbf{Z}$ with $\mathbf{I} \subseteq \mathbf{Z} \subseteq \mathbf{R}$	$O(n+m)$
FINDMINSEP	a minimal m-separator $\mathbf{Z}$ with $\mathbf{I} \subseteq \mathbf{Z} \subseteq \mathbf{R}$	$O(n^2)$
FINDMINCOSTSEP	a minimum-cost m-separator $\mathbf{Z}$ with $\mathbf{I} \subseteq \mathbf{Z} \subseteq \mathbf{R}$	$O(n^3)$
Enumeration: For given $\mathbf{X}, \mathbf{Y}, \mathbf{I}, \mathbf{R}$ enumerate all ...		
LISTSEP	m-separators $\mathbf{Z}$ with $\mathbf{I} \subseteq \mathbf{Z} \subseteq \mathbf{R}$	$O(n(n+m))$ delay
LISTMINSEP	minimal m-separators $\mathbf{Z}$ with $\mathbf{I} \subseteq \mathbf{Z} \subseteq \mathbf{R}$	$O(n^3)$ delay

Table 1: Definitions of algorithmic tasks related to m-separation. Throughout, $\mathbf{X}, \mathbf{Y}, \mathbf{R}$ are pairwise disjoint node sets, $\mathbf{Z}$ is disjoint with $\mathbf{X}, \mathbf{Y}$ which are nonempty, and $\mathbf{I}, \mathbf{R}, \mathbf{Z}$ can be empty. By a minimal m-separator $\mathbf{Z}$, with $\mathbf{I} \subseteq \mathbf{Z} \subseteq \mathbf{R}$, we mean a set such that no proper subset $\mathbf{Z}'$ of $\mathbf{Z}$, with $\mathbf{I} \subseteq \mathbf{Z}'$, m-separates the pair $\mathbf{X}$ and $\mathbf{Y}$. Analogously, we define a minimal and a minimum-cost m-separator. The construction algorithms will output $\perp$ if no set fulfilling the listed condition exists. Delay complexity for e.g. LISTMINSEP refers to the time needed to output one solution when there can be exponentially many solutions (see Takata (2010)).

function LISTSEP($\mathcal{G}, \mathbf{X}, \mathbf{Y}, \mathbf{I}, \mathbf{R}$)
  if FINDSEP($\mathcal{G}, \mathbf{X}, \mathbf{Y}, \mathbf{I}, \mathbf{R}$) $\neq \perp$ **then**
    if $\mathbf{I} = \mathbf{R}$ **then** Output $\mathbf{I}$
    else
      $V \leftarrow$ an arbitrary node of $\mathbf{R} \setminus \mathbf{I}$
      LISTSEP($\mathcal{G}, \mathbf{X}, \mathbf{Y}, \mathbf{I} \cup \{V\}, \mathbf{R}$)
      LISTSEP($\mathcal{G}, \mathbf{X}, \mathbf{Y}, \mathbf{I}, \mathbf{R} \setminus \{V\}$)

Figure 1: ListSep

Proposition 3.10. *The task* LISTSEP *can be solved with polynomial delay* $O(n(n+m))$.

Proof. Algorithm LISTSEP performs backtracking to enumerate all $\mathbf{Z}$ with $\mathbf{I} \subseteq \mathbf{Z} \subseteq \mathbf{R}$ aborting branches that will not find a valid separator. Since every leaf will output a separator, the tree height is at most n and the existence check needs $O(n+m)$, the delay time is $O(n(n+m))$. The algorithm generates every separator exactly once: if initially $\mathbf{I} \subseteq \mathbf{R}$, with $V \in \mathbf{R} \setminus \mathbf{I}$, then the first recursive call returns all separators $\mathbf{Z}$ with $V \in \mathbf{Z}$ and the second call returns all $\mathbf{Z}'$ with $V \notin \mathbf{Z}'$. Thus the generated separators are pairwise disjoint. This is a modification of the enumeration algorithm for minimal vertex separators (Takata, 2010). □

4 ADJUSTMENT IN DAGS

In this section, we leverage the algorithmic framework of the last section together with a new constructive, sound and complete criterion for covariate adjustment in DAGs to solve all problems listed in Table 1 for adjustment sets instead of m-separators in the same asymptotic time. First, however, we need to introduce some more notation pertaining to the causal interpretation DAGs.

Do-operator and adjustment sets. A DAG $\mathcal{G}$ encodes the factorization of joint distribution p for the set of variables $\mathbf{V} = \{X_1, \ldots, X_n\}$ as $p(\mathbf{v}) = \prod_{j=1}^{n} p(x_j|pa_j)$, where pa_j denotes a particular realization of the parent variables of X_j in $\mathcal{G}$. When interpreted causally, an edge $X_i \to X_j$ is taken to represent a direct causal effect of X_i on X_j. For disjoint $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{V}$, the *(total) causal effect* of $\mathbf{X}$ on $\mathbf{Y}$ is $p(\mathbf{y}|do(\mathbf{x}))$ where $do(\mathbf{x})$ represents an intervention that sets $\mathbf{X} = \mathbf{x}$. In a DAG, this intervention corresponds to removing all edges into $\mathbf{X}$, disconnecting $\mathbf{X}$ from its parents. We denote the resulting graph as $\mathcal{G}_{\overline{\mathbf{X}}}$. Given DAG $\mathcal{G}$ and a joint probability density p for $\mathbf{V}$ the post-intervention distribution can be expressed in a truncated factorization formula:

$$p(\mathbf{v}|do(\mathbf{x})) = \begin{cases} \prod_{X_j \in \mathbf{V} \setminus \mathbf{X}} p(x_j|pa_j) & \text{for } \mathbf{V} \text{ consistent with } \mathbf{x} \\ 0 & \text{otherwise.} \end{cases}$$

Definition 4.1 (Adjustment (Pearl, 2009))**.** *Given a DAG* $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ *and pairwise disjoint* $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$, $\mathbf{Z}$ *is called* covariate adjustment for estimating the causal effect of $\mathbf{X}$ on $\mathbf{Y}$, *or simply* adjustment, *if for every distribution* p *consistent with* $\mathcal{G}$ *we have* $p(\mathbf{y} \mid do(\mathbf{x})) = \sum_{\mathbf{z}} p(\mathbf{y} \mid \mathbf{x}, \mathbf{z}) p(\mathbf{z})$.

Definition 4.2 (Adjustment criterion (Shpitser et al., 2010; Shpitser, 2012))**.** *Let* $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ *be a DAG, and* $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$ *be pairwise disjoint subsets of variables. The set* $\mathbf{Z}$ *satisfies the adjustment criterion relative to* $(\mathbf{X}, \mathbf{Y})$ *in* $\mathcal{G}$ *if*

(*a*) *no element in* $\mathbf{Z}$ *is a descendant in* $\mathcal{G}$ *of any* $W \in \mathbf{V} \setminus \mathbf{X}$ *which lies on a proper causal path from* $\mathbf{X}$ *to* $\mathbf{Y}$ *and*

(*b*) *all proper non-causal paths in* $\mathcal{G}$ *from* $\mathbf{X}$ *to* $\mathbf{Y}$ *are blocked by* $\mathbf{Z}$.

Remark 4.3. *In (Shpitser et al., 2010; Shpitser, 2012) the criterion is stated in a slightly different way, namely using in the condition* (*a*) $\mathcal{G}_{\overline{\mathbf{X}}}$ *instead of* $\mathcal{G}$. *However, the two statements are equivalent.*

Proof. First note that if $\mathbf{Z}$ satisfies the condition (*a*) then $\mathbf{Z}$ satisfies (*a*) with $\mathcal{G}_{\overline{\mathbf{X}}}$ instead of $\mathcal{G}$, too. Since condi-

tions (b) in Definition 4.2 and in (Shpitser et al., 2010; Shpitser, 2012) are identical, the adjustment criterion above implies the criterion of Shpitser et al.

Now assume $\mathbf{Z}$ satisfies the condition (a) with $\mathcal{G}_{\overline{\mathbf{X}}}$ instead of $\mathcal{G}$ and the condition (b). We show that $\mathbf{Z}$ then satisfies the condition (a), or there must exist some $W \in \mathbf{V} \setminus \mathbf{X}$, which lies on a proper causal path from $\mathbf{X}$ to $\mathbf{Y}$, and a causal path from W to $\mathbf{Z}$ which intersects $\mathbf{X}$.

Let $W \to \ldots \to Y$ denote the suffix of the path from $\mathbf{X}$ to $\mathbf{Y}$ starting in W. Note that this path can consist only of the vertex W. Additionally, for the causal path from W to $\mathbf{Z}$, let $W \to \ldots \to X$ be its shortest prefix which intersects $\mathbf{X}$. Then, from the condition (a), with $\mathcal{G}_{\overline{\mathbf{X}}}$ instead of $\mathcal{G}$, we know that no vertex of $W \to \ldots \to X$ belongs to $\mathbf{Z}$. This leads to a contradiction with the condition (b) since $X \leftarrow \ldots \leftarrow W \to \ldots \to Y$ is a proper non-causal path in $\mathcal{G}$ from $\mathbf{X}$ to $\mathbf{Y}$ that is not blocked by $\mathbf{Z}$. □

Analogously to $\mathcal{G}_{\overline{\mathbf{X}}}$, by $\mathcal{G}_{\underline{\mathbf{X}}}$ we denote a DAG obtained from $\mathcal{G}$ by removing all edges leaving $\mathbf{X}$.

4.1 CONSTRUCTIVE BACK-DOOR CRITERION

Definition 4.4 (Proper back-door graph)**.** *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a DAG, and $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{V}$ be pairwise disjoint subsets of variables. The* proper back-door graph*, denoted as $\mathcal{G}^{pbd}_{\mathbf{XY}}$, is obtained from $\mathcal{G}$ by removing the first edge of every proper causal path form $\mathbf{X}$ to $\mathbf{Y}$.*

Note the difference between the back-door graph $\mathcal{G}_{\underline{\mathbf{X}}}$ and the proper back-door graph $\mathcal{G}^{pbd}_{\mathbf{XY}}$: in $\mathcal{G}_{\underline{\mathbf{X}}}$ *all* edges leaving $\mathbf{X}$ are removed while in $\mathcal{G}^{pbd}_{\mathbf{XY}}$ only those that lie on a proper causal path. However, to construct $\mathcal{G}^{pbd}_{\mathbf{XY}}$ still only elementary operations are sufficient. Indeed, we remove all edges $X \to D$ in $\mathbf{E}$ such that $X \in \mathbf{X}$ and D is in the subset, which we call $PCP(\mathbf{X}, \mathbf{Y})$, obtained as follows:

$$PCP(\mathbf{X}, \mathbf{Y}) = (De_{\overline{\mathbf{X}}}(\mathbf{X}) \setminus \mathbf{X}) \cap An_{\underline{\mathbf{X}}}(\mathbf{Y}) \qquad (1)$$

where $De_{\overline{\mathbf{X}}}(\mathbf{W})$ denotes descendants of $\mathbf{W}$ in $\mathcal{G}_{\overline{\mathbf{X}}}$. $An_{\underline{\mathbf{X}}}(\mathbf{W})$ is defined analogously for $\mathcal{G}_{\underline{\mathbf{X}}}$. Hence, the proper back-door graph can be constructed from $\mathcal{G}$ in linear time $O(m + n)$.

Now we propose the following adjustment criterion. For short, we will denote the set $De(PCP(\mathbf{X}, \mathbf{Y}))$ as $Dpcp(\mathbf{X}, \mathbf{Y})$.

Definition 4.5 (Constructive back-door criterion (CBC))**.** *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a DAG, and let $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$ be pairwise disjoint subsets of variables. The set $\mathbf{Z}$ satisfies the* constructive back-door criterion *relative to $(\mathbf{X}, \mathbf{Y})$ in $\mathcal{G}$ if*

(a) $\mathbf{Z} \subseteq \mathbf{V} \setminus Dpcp(\mathbf{X}, \mathbf{Y})$ *and*

(b) $\mathbf{Z}$ *d-separates* $\mathbf{X}$ *and* $\mathbf{Y}$ *in the proper back-door graph* $\mathcal{G}^{pbd}_{\mathbf{XY}}$.

Theorem 4.6. *The constructive back-door criterion is equivalent to the adjustment criterion.*

Proof. First observe that the conditions (a) of both criteria are identical. Assume conditions (a) and (b) of the adjustment criterion hold. We show that (b) of the constructive back-door criterion follows. Let π be any proper path from $\mathbf{X}$ to $\mathbf{Y}$ in $\mathcal{G}^{pbd}_{\mathbf{XY}}$. Because $\mathcal{G}^{pbd}_{\mathbf{XY}}$ does not contain causal paths from $\mathbf{X}$ to $\mathbf{Y}$, π is not causal and has to be blocked by $\mathbf{Z}$ in $\mathcal{G}$ by the assumption. Since removing edges cannot open paths, π is blocked by $\mathbf{Z}$ in $\mathcal{G}^{pbd}_{\mathbf{XY}}$ as well.

Now we show that (a) and (b) of the constructive back-door criterion together imply (b) of the adjustment criterion. If that were not the case, then there could exist a proper non-causal path π from $\mathbf{X}$ to $\mathbf{Y}$ that is blocked in $\mathcal{G}^{pbd}_{\mathbf{XY}}$ but open in $\mathcal{G}$. There can be two reasons why π is blocked in $\mathcal{G}^{pbd}_{\mathbf{XY}}$: (1) The path starts with an edge $X \to D$ that does not exist in $\mathcal{G}^{pbd}_{\mathbf{XY}}$. Then we have $D \in PCP(\mathbf{X}, \mathbf{Y})$. For π to be non-causal, it would have to contain a collider $C \in An(\mathbf{Z}) \cap De(D) \subseteq An(\mathbf{Z}) \cap Dpcp(\mathbf{X}, \mathbf{Y})$. But because of ($a$), $An(\mathbf{Z}) \cap Dpcp(\mathbf{X}, \mathbf{Y})$ is empty. (2) A collider C on π is an ancestor of $\mathbf{Z}$ in $\mathcal{G}$, but not in $\mathcal{G}^{pbd}_{\mathbf{XY}}$. Then there must be a directed path from C to $\mathbf{Z}$ via an edge $X \to D$ with $D \in An(\mathbf{Z}) \cap PCP(\mathbf{X}, \mathbf{Y})$, contradicting (a). □

4.2 ADJUSTING FOR MULTIPLE EXPOSURES

For a singleton set $\mathbf{X} = \{X\}$ of exposures we know that if a set of variables $\mathbf{Y}$ is disjoint from $\{X\} \cup Pa(X)$ then one obtains easily an adjustment set with respect to X and $\mathbf{Y}$ as $\mathbf{Z} = Pa(X)$ (Pearl, 2009, Theorem 3.2.2). The situation changes drastically if the effect of multiple exposures is estimated. Theorem 3.2.5 in Pearl (2009) claims that the expression for $P(\mathbf{y} \mid do(\mathbf{x}))$ is obtained by adjusting for $Pa(\mathbf{X})$ if $\mathbf{Y}$ is disjoint from $\mathbf{X} \cup Pa(\mathbf{X})$, but, as the DAG in Fig. 2 shows, this is not true: the set $\mathbf{Z} = Pa(X_1, X_2) = \{Z_2\}$ is not an adjustment set according to $\{X_1, X_2\}$ and Y. In this case one can identify the causal effect by adjusting for $\mathbf{Z} = \{Z_1, Z_2\}$ only. Indeed, for more than one exposure, no adjustment set may exist at all even without latent covariates and even though $\mathbf{Y} \cap (\mathbf{X} \cup Pa(\mathbf{X})) = \emptyset$, e.g. in the DAG $X_1 \overset{\frown}{\to} X_2 \leftarrow Z \leftarrow Y$.

Using our criterion, we can construct a simple adjustment set explicitly if one exists. For a DAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ we define the set

$$Adj(\mathbf{X}, \mathbf{Y}) = An(\mathbf{X} \cup \mathbf{Y}) \setminus (\mathbf{X} \cup \mathbf{Y} \cup Dpcp(\mathbf{X}, \mathbf{Y})).$$

Theorem 4.7. *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a DAG and let $\mathbf{X}, \mathbf{Y} \subseteq V$ be distinct node sets. Then the following statements are equivalent:*

1. *There exists an adjustment in $\mathcal{G}$ w.r.t. $\mathbf{X}$ and $\mathbf{Y}$.*

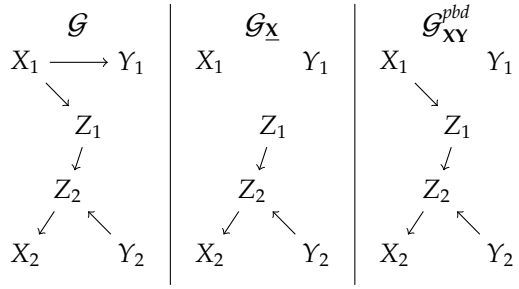


Figure 2: A DAG where for $\mathbf{X} = \{X_1, X_2\}$ and $\mathbf{Y} = \{Y_1, Y_2\}$, $\mathbf{Z} = \{Z_1, Z_2\}$ is a valid and minimal adjustment, but no set fulfills the back-door criterion (Pearl, 2009), and the parents of $\mathbf{X}$ are not a valid adjustment set either.

2. *$Adj(\mathbf{X}, \mathbf{Y})$ is an adjustment w.r.t. $\mathbf{X}$ and $\mathbf{Y}$.*

3. *$Adj(\mathbf{X}, \mathbf{Y})$ d-separates $\mathbf{X}$ and $\mathbf{Y}$ in the proper back-door graph $\mathcal{G}^{pbd}_{\mathbf{XY}}$.*

Proof. The implication (*3*) $\Rightarrow$ (*2*) follows directly from the criterion Def. 4.5 and the definition of $Adj(\mathbf{X}, \mathbf{Y})$. Since the implication (*2*) $\Rightarrow$ (*1*) is obvious, it remains to prove (*1*) $\Rightarrow$ (*3*).

Assume there exists an adjustment set $\mathbf{Z}_0$ w.r.t. $\mathbf{X}$ and $\mathbf{Y}$. From Theorem 4.6 we know that $\mathbf{Z}_0 \cap Dpcp(\mathbf{X}, \mathbf{Y}) = \emptyset$ and that $\mathbf{Z}_0$ d-separates $\mathbf{X}$ and $\mathbf{Y}$ in $\mathcal{G}^{pbd}_{\mathbf{XY}}$. Our task is to show that $Adj(\mathbf{X}, \mathbf{Y})$ d-separates $\mathbf{X}$ and $\mathbf{Y}$ in $\mathcal{G}^{pbd}_{\mathbf{XY}}$. This follows from Lemma 3.4 used for the proper back-door graph $\mathcal{G}^{pbd}_{\mathbf{XY}}$ if we take $\mathbf{I} = \emptyset$, $\mathbf{R} = \mathbf{V} \setminus (\mathbf{X} \cup \mathbf{Y} \cup Dpcp(\mathbf{X}, \mathbf{Y}))$. □

From Equation 1 and the definition $Dpcp(\mathbf{X}, \mathbf{Y}) = De(PCP(\mathbf{X}, \mathbf{Y}))$ we then obtain immediately:

Corollary 4.8. *Given two distinct sets $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{V}$, $Adj(\mathbf{X}, \mathbf{Y})$ can be found in $O(n + m)$ time.*

4.3 TESTING, COMPUTING, AND ENUMERATING ADJUSTMENT SETS

Using our criterion, every algorithm for m-separating sets $\mathbf{Z}$ between $\mathbf{X}$ and $\mathbf{Y}$ can be used for adjustment sets with respect to $\mathbf{X}$ and $\mathbf{Y}$, by requiring that $\mathbf{Z}$ not contain any node in $Dpcp(\mathbf{X}, \mathbf{Y})$. This allows solving all problems listed in Table 1 for adjustment sets in DAGs instead of m-separators. Below, we name those problems analogously as for m-separation, e.g. the problem to decide whether $\mathbf{Z}$ is an adjustment set w.r.t. $\mathbf{X}, \mathbf{Y}$ is named TESTADJ in analogy to TESTSEP.

TESTADJ can be solved by testing if $\mathbf{Z} \cap Dpcp(\mathbf{X}, \mathbf{Y}) = \emptyset$ and $\mathbf{Z}$ is a d-separator in the proper back-door graph $\mathcal{G}^{pbd}_{\mathbf{XY}}$. Since $\mathcal{G}^{pbd}_{\mathbf{XY}}$ can be constructed from $\mathcal{G}$ in linear time, the total time complexity of this algorithm is $O(n + m)$.

TESTMINADJ can be solved with an algorithm that iteratively removes nodes from $\mathbf{Z}$ and tests if the resulting set remains an adjustment set w.r.t. $\mathbf{X}$ and $\mathbf{Y}$. This can be done in time $O(n(n + m))$. Alternatively, one can construct the proper back-door graph $\mathcal{G}^{pbd}_{\mathbf{XY}}$ from $\mathcal{G}$ and test if $\mathbf{Z}$ is a minimal d-separator, with $\mathbf{Z} \subseteq \mathbf{V} \setminus Dpcp(\mathbf{X}, \mathbf{Y})$ between $\mathbf{X}$ and $\mathbf{Y}$. This can be computed in time $O(n^2)$. The correctness of these algorithms follows from the proposition below, which is a generalization of the result in Tian et al. (1998).

Proposition 4.9. *If no single node Z can be removed from an adjustment set $\mathbf{Z}$ such that the resulting set $\mathbf{Z}' = \mathbf{Z} \setminus Z$ is no longer an adjustment set, then $\mathbf{Z}$ is minimal.*

The remaining problems like FINDADJ, FINDMINADJ etc. can be solved using corresponding algorithms for finding, resp. listing m-separations applied for proper back-door graphs. Since the proper back-door graph can be constructed in linear time the time complexities to solve the problems above are as listed in Table 1.

5 ADJUSTMENT IN MAGS

We now generalize the results from the previous section to MAGs. Two examples may illustrate why this generalization is not trivial. First, take $\mathcal{G} = X \to Y$. If $\mathcal{G}$ is interpreted as a DAG, then the empty set is valid for adjustment. If $\mathcal{G}$ is however taken as a MAG, then there exists no adjustment set as $\mathcal{G}$ represents among others the DAG $U \overset{\to}{\to} X \overset{\to}{\to} Y$ where U is an unobserved confounder. Second, take $\mathcal{G} = A \to X \to Y$. In that case, the empty set is an adjustment set regardless of whether $\mathcal{G}$ is interpreted as a DAG or a MAG. The reasons will become clear as we move on. First, let us recall the semantics of a MAG. The following definition can easily be given for AGs in general, but we do not need this generality for our purpose.

Definition 5.1 (DAG representation by MAGs (Richardson and Spirtes, 2002))**.** *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a DAG, and let $\mathbf{S}, \mathbf{L} \subseteq \mathbf{V}$. The MAG $\mathcal{M} = \mathcal{G}[^{\mathbf{L}}_{\mathbf{S}}]$ is a graph with nodes $\mathbf{V} \setminus (\mathbf{S} \cup \mathbf{L})$ and defined as follows. (1) Two nodes U and V are adjacent in $\mathcal{G}[^{\mathbf{L}}_{\mathbf{S}}]$ if they cannot be m-separated by any $\mathbf{Z}$ with $\mathbf{S} \subseteq \mathbf{Z} \subseteq \mathbf{V} \setminus \mathbf{L}$ in $\mathcal{G}$. (2) The edge between U and V is*

$U - V$ *if* $U \in An(\mathbf{S} \cup V)$ *and* $V \in An(\mathbf{S} \cup U)$*;*

$U \to V$ *if* $U \in An(\mathbf{S} \cup V)$ *and* $V \notin An(\mathbf{S} \cup U)$*;*

$U \leftrightarrow V$ *if* $U \notin An(\mathbf{S} \cup V)$ *and* $V \notin An(\mathbf{S} \cup U)$*.*

We call $\mathbf{L}$ latent *variables and* $\mathbf{S}$ selection *variables. We say there is* selection bias *if* $\mathbf{S} \neq \emptyset$*.*

Hence, every MAG represents an infinite set of underlying DAGs that all share the same ancestral relationships. For a given MAG $\mathcal{M}$, we can construct a represented DAG $\mathcal{G}$ by

replacing every edge $X - Y$ by a path $X \to S \leftarrow Y$, and every edge $X \leftrightarrow Y$ by $X \leftarrow L \to Y$, where S and L are new nodes; then $\mathcal{M} = \mathcal{G}[^{\mathbf{L}}_{\mathbf{S}}$ where $\mathbf{S}$ and $\mathbf{L}$ are all new nodes. $\mathcal{G}$ is called the *canonical DAG* of $\mathcal{M}$ (Richardson and Spirtes, 2002), which we write as $C(\mathcal{M})$.

Lemma 5.2 (Preservation of separating sets (Richardson and Spirtes, 2002))**.** *$\mathbf{Z}$ m-separates $\mathbf{X}, \mathbf{Y}$ in $\mathcal{G}[^{\mathbf{L}}_{\mathbf{S}}$ if and only if $\mathbf{Z} \cup \mathbf{S}$ m-separates $\mathbf{X}, \mathbf{Y}$ in $\mathcal{G}$.*

We now extend the concept of adjustment to MAGs in the usual way (Maathuis and Colombo, 2013).

Definition 5.3 (Adjustment in MAGs)**.** *Given a MAG $\mathcal{M} = (\mathbf{V}, \mathbf{E})$ and two variable sets $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{V}$, $\mathbf{Z} \subseteq \mathbf{V}$ is an adjustment set for $\mathbf{X}, \mathbf{Y}$ in $\mathcal{M}$ if for every probability distribution $p(\mathbf{v}')$ consistent with a DAG $\mathcal{G} = (\mathbf{V}', \mathbf{E}')$ for which $\mathcal{G}[^{\mathbf{L}}_{\mathbf{S}} = \mathcal{M}$ for some $\mathbf{S}, \mathbf{L} \subseteq \mathbf{V}' \setminus \mathbf{V}$, we have*

$$p(\mathbf{y} \mid do(\mathbf{x})) = \sum_{\mathbf{z}} p(\mathbf{y} \mid \mathbf{x}, \mathbf{z}, \mathbf{s}) p(\mathbf{z} \mid \mathbf{s}) \,. \tag{2}$$

Selection bias (i.e., $\mathbf{S} \neq \emptyset$) substantially complicates adjustment, and in fact nonparametric causal inference in general (Zhang, 2008)[2]. Due to these limitations, we restrict ourselves to the case $\mathbf{S} = \emptyset$ in the rest of this section. Note however that recovery from selection bias is sometimes possible with additional population data, and graphical conditions exist to identify such cases (Barenboim et al., 2014).

5.1 ADJUSTMENT AMENABILITY

In this section we first identify a class of MAGs in which adjustment is impossible because of causal ambiguities – e.g., the simple MAG $X \to Y$ falls into this class, but the larger MAG $A \to X \to Y$ does not.

Definition 5.4 (Visible edge (Zhang, 2008))**.** *Given a MAG $\mathcal{M} = (\mathbf{V}, \mathbf{E})$, an edge $X \to Y \in \mathbf{E}$ is called* visible *if in all DAGs $\mathcal{G} = (\mathbf{V}', \mathbf{E}')$ with $\mathcal{G}[^{\mathbf{L}}_{\mathbf{S}} = \mathcal{M}$ for some $\mathbf{S}, \mathbf{L} \subseteq \mathbf{V}'$, all d-connected walks between X and Y in $\mathcal{G}$ that contain only nodes of $\mathbf{S} \cup \mathbf{L} \cup X \cup Y$ are directed paths.*

Intuitively, an invisible directed edge $X \to Y$ means that there may still hidden confounding factors between X and Y, which is guaranteed not to be the case if the edge is visible.

Lemma 5.5 (Graphical conditions for edge visibility (Zhang, 2008))**.** *In a MAG $\mathcal{M} = (\mathbf{V}, \mathbf{E})$, an edge $X \to D$ is visible if and only if there is a node A not adjacent to D where (1) $A \to X \in \mathbf{E}$ or $A \leftrightarrow X \in \mathbf{E}$, or (2) there is a collider path $A \leftrightarrow V_1 \leftrightarrow \ldots \leftrightarrow V_n \leftrightarrow X$ or $A \to V_1 \leftrightarrow \ldots \leftrightarrow V_n \leftrightarrow X$ where all V_i are parents of D.*

[2] A counterexample is the graph $A \leftarrow X \to Y$, where we can safely assume that A is the ancestor of a selection variable. A sufficient and necessary condition for adjustment under selection bias is $\mathbf{Y} \perp\!\!\!\perp \mathbf{S} \mid \mathbf{X}$ (Barenboim et al., 2014), which is so restrictive that most statisticians would probably not even speak of "selection bias" anymore in such a case.

Definition 5.6. *We call a MAG $\mathcal{M} = (\mathbf{V}, \mathbf{E})$* adjustment amenable *w.r.t. $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{V}$ if all proper causal paths from $\mathbf{X}$ to $\mathbf{Y}$ start with a visible directed edge.*

Lemma 5.7. *If a MAG $\mathcal{M} = (\mathbf{V}, \mathbf{E})$ is not adjustment amenable w.r.t. $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{V}$ then there exists no adjustment set $\mathbf{W}$ for $\mathbf{X}, \mathbf{Y}$ in $\mathcal{M}$.*

Proof. If the first edge $X \to D$ on some causal path to $\mathbf{Y}$ in $\mathcal{M}$ is not visible, then there exists a consistent DAG $\mathcal{G}$ where there is a non-causal path between X and $\mathbf{Y}$ via V that could only be blocked in $\mathcal{M}$ by conditioning on D or some of its descendants. But such conditioning would violate the adjustment criterion in $\mathcal{G}$. □

5.2 ADJUSTMENT CRITERION FOR MAGS

We now show that DAG adjustment criterion generalizes to adjustment amenable MAGs. The adjustment criterion and the constructive back-door criterion are defined like their DAG counterparts (Definitions 4.2 and 4.4), replacing d-with m-separation for the latter.

Theorem 5.8. *Given an adjustment amenable MAG $\mathcal{M} = (\mathbf{V}, \mathbf{E})$ and three disjoint node sets $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$, the following statements are equivalent:*

(i) $\mathbf{Z}$ *is an adjustment relative to* $\mathbf{X}, \mathbf{Y}$ *in* $\mathcal{M}$.

(ii) $\mathbf{Z}$ *fulfills the adjustment criterion (AC) w.r.t.* $(\mathbf{X}, \mathbf{Y})$ *in* $\mathcal{M}$.

(iii) $\mathbf{Z}$ *fulfills the constructive backdoor criterion (CBC) w.r.t.* $(\mathbf{X}, \mathbf{Y})$ *in* $\mathcal{M}$.

Proof. The equivalence of (ii) and (iii) is established by observing that the proof of Theorem 4.6 generalizes to m-separation. Below we establish equivalence of (i) and (ii).

$\neg(ii) \Rightarrow \neg(i)$: If $\mathbf{Z}$ violates the adjustment criterion in $\mathcal{M}$, it does so in the canonical DAG $C(\mathcal{M})$, and thus is not an adjustment in $\mathcal{M}$.

$\neg(i) \Rightarrow \neg(ii)$: Let $\mathcal{G}$ be a DAG with $\mathcal{G}[^{\mathbf{L}}_{\emptyset} = \mathcal{M}$ in which $\mathbf{Z}$ violates the AC. We show that (*a*) if $\mathbf{Z} \cap Dpcp(\mathbf{X}, \mathbf{Y}) \neq \emptyset$ in $\mathcal{G}$ then $\mathbf{Z} \cap Dpcp(\mathbf{X}, \mathbf{Y}) \neq \emptyset$ in $\mathcal{M}$ as well, or there exists a proper non-causal path in $\mathcal{M}$ that cannot be m-separated; and (*b*) if $\mathbf{Z} \cap Dpcp(\mathbf{X}, \mathbf{Y}) = \emptyset$ in $\mathcal{G}$ and $\mathbf{Z}$ d-connects a proper non-causal path in $\mathcal{G}$, then it m-connects a proper non-causal path in $\mathcal{M}$.

(*a*) Suppose that in $\mathcal{G}$, $\mathbf{Z}$ contains a node Z in $Dpcp(\mathbf{X}, \mathbf{Y})$, and let $\mathbf{W} = \mathrm{PCP}(\mathbf{X}, \mathbf{Y}) \cap An(Z)$. If $\mathcal{M}$ still contains at least one node $W_1 \in \mathbf{W}$, then W_1 lies on a proper causal path in $\mathcal{M}$ and Z is a descendant of W_1 in $\mathcal{M}$. Otherwise, $\mathcal{M}$

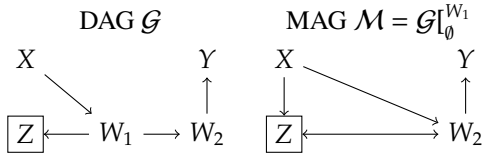


Figure 3: Illustration of the case in the proof of Theorem 5.8 where Z descends from W_1 which in a DAG $\mathcal{G}$ is on a proper causal path from X to Y, but is not a descendant of a node on a proper causal path from X to Y in the MAG $\mathcal{M}$ after marginalizing W_1. In such cases, conditioning on Z will m-connect X and Y in $\mathcal{M}$ via a proper non-causal path.

must contain a node $W_2 \in \mathrm{PCP}_{\mathcal{G}}(\mathbf{X}, \mathbf{Y}) \setminus An(Z)$ (possibly $W_2 \in \mathbf{Y}$) such that $W_2 \leftrightarrow A$, $X \rightarrow W_2$, and $X \rightarrow A$ are edges in $\mathcal{M}$, where $A \in An(Z)$ (possibly $A = Z$; see Fig. 3). Then $\mathcal{M}$ contains an m-connected proper non-causal path $X \rightarrow A \leftrightarrow W \rightarrow W_2 \rightarrow \ldots \rightarrow Y$.

(b) Suppose that in $\mathcal{G}$, $\mathbf{Z} \cap Dpcp(\mathbf{X}, \mathbf{Y}) = \emptyset$, and there exists an open proper non-causal path from $\mathbf{X}$ to $\mathbf{Y}$. Then there must then also be a proper non-causal *walk* $w_{\mathcal{G}}$ from some $X \in \mathbf{X}$ to some $Y \in \mathbf{Y}$ (Lemma A.1), which is d-connected by $\mathbf{Z}$ in $\mathcal{G}$. Let $w_{\mathcal{M}}$ denote the subsequence of $w_{\mathcal{G}}$ formed by nodes in $\mathcal{M}$, which includes all colliders on $w_{\mathcal{G}}$. The sequence $w_{\mathcal{M}}$ is a path in $\mathcal{M}$, but is not necessarily m-connected by $\mathbf{Z}$; all colliders on $w_{\mathcal{M}}$ are in $\mathbf{Z}$ because every non-$\mathbf{Z}$ must be a parent of at least one of its neighbours, but there can subsequences $U, Z_1, \ldots, Z_k, V$ on $w_{\mathcal{M}}$ where all $Z_i \in \mathbf{Z}$ but some of the Z_i are not colliders on $w_{\mathcal{M}}$. However, then we can form from $w_{\mathcal{M}}$ an m-connected walk by bypassing some sequences of $\mathbf{Z}$-nodes (Lemma A.9). Let $w'_{\mathcal{M}}$ be the resulting walk.

If $w'_{\mathcal{M}}$ is a proper non-causal walk, then there must also exist a proper non-causal path in $\mathcal{M}$ (Lemma A.1), violating the AC. It therefore remains to show that $w'_{\mathcal{M}}$ is not a proper causal path. This must be the case if $w_{\mathcal{G}}$ does not contain colliders, because then the first edge of $w_{\mathcal{M}} = w'_{\mathcal{M}}$ cannot be a visible directed edge out of X. Otherwise, the only way for $w'_{\mathcal{M}}$ to be proper causal is if all $\mathbf{Z}$-nodes in $w_{\mathcal{M}}$ have been bypassed in $w'_{\mathcal{M}}$ by edges pointing away from $\mathbf{X}$. In that case, one can show by several case distinctions that the first edge $X \rightarrow D$ of $w'_{\mathcal{M}}$, where $D \notin \mathbf{Z}$, cannot be visible (see Figure 4 for an example of such a case).

For simplicity, assume that $\mathcal{M}$ contains a subpath $A \rightarrow X \rightarrow D$ where A is not adjacent to D; the other cases of edge visibility like $A \leftrightarrow X \rightarrow D$ (Lemma 5.5). are treated analogously. In $\mathcal{G}$, there are inducing paths (possibly several) π_{AX} from A to X and π_{XD} from X to D w.r.t $\emptyset, \mathbf{L}$; π_{AX} must have an arrowhead at X. We distinguish several cases on the shape of π_{XD}. (1) A path π_{XD} has an arrowhead at X as well. Then A, D are adjacent (Lemma A.13), a contradiction. (2) No inducing path π_{XD} has an arrowhead at X. Then $w_{\mathcal{G}}$ must start with an

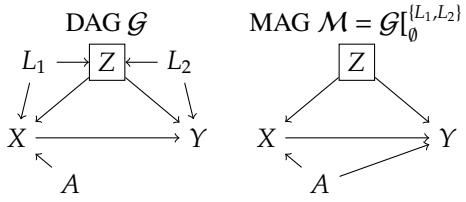


Figure 4: Case (b) in the proof of Theorem 5.8: A proper non-causal path $w_{\mathcal{G}} = X \leftarrow L_1 \rightarrow Z \leftarrow L_s \rightarrow Y$ in a DAG is d-connected by $\mathbf{Z}$, but the corresponding proper non-casual path $w_{\mathcal{M}} = X \leftarrow Z \rightarrow Y$ is not m-connected in the MAG, and its m-connected subpath $w'_{\mathcal{M}} = X \rightarrow Y$ is proper causal. However, this also renders the edge $X \rightarrow Y$ invisible, because otherwise A could be m-separated from Y by $\mathbf{U} = \{X, Z\}$ in $\mathcal{M}$ but not in $\mathcal{G}$.

arrow out of X, and must contain a collider $Z \in De(X)$ because $w_{\mathcal{G}}$ is not causal. (a) $Z \in De(D)$. This contradicts $\mathbf{Z} \cap Dpcp(\mathbf{X}, \mathbf{Y}) = \emptyset$. So (b) $Z \notin De(D)$. Then by construction of $w'_{\mathcal{M}}$ (Lemma A.9), $w_{\mathcal{M}}$ must start with an inducing $\mathbf{Z}$-trail $X \rightarrow Z, Z_1, \ldots, Z_n, D$, which is also an inducing path from X to D in $\mathcal{G}$ w.r.t. $\emptyset, \mathbf{L}$. Then $Z, Z_1, \ldots, Z_n, D$ must also be an inducing path in $\mathcal{G}$ w.r.t. $\emptyset, \mathbf{L}$ because $An(X) \subseteq An(Z)$. Hence Z and D are adjacent. We distinguish cases on the path $X \rightarrow Z, D$ in $\mathcal{M}$. (i) If $X \rightarrow Z \rightarrow D$, then Z lies on a proper causal path, contradicting $\mathbf{Z} \cap Dpcp(\mathbf{X}, \mathbf{Y}) = \emptyset$. (ii) If $X \rightarrow Z \leftrightarrow D$, or $X \rightarrow Z \leftarrow D$, then we get an m-connected proper non-causal walk along Z and D. □

5.3 ADJUSTMENT SET CONSTRUCTION

In the previous section, we have already shown that the CBC is equivalent to the AC for MAGs as well; hence, adjustment sets for a given MAG $\mathcal{M}$ can be found by forming the proper back-door graph $\mathcal{M}^{pbd}_{\mathbf{XY}}$ and then applying the algorithms from the previous section. In principle, care must be taken when removing edges from MAGs as the result might not be a MAG; however, this is not the case when removing only directed edges.

Lemma 5.9 (Closure of maximality under removal of directed edges)**.** *Given a MAG $\mathcal{M}$, every graph $\mathcal{M}'$ formed by removing only directed edges from $\mathcal{M}$ is also a MAG.*

Proof. Suppose the converse, i.e. $\mathcal{M}$ is no longer a MAG after removal of some edge $X \rightarrow D$. Then X and D cannot be m-separated even after the edge is removed because X and D are collider connected via a path whose nodes are all ancestors of X or D (Richardson and Spirtes, 2002). The last edge on this path must be $C \leftrightarrow D$ or $C \leftarrow D$, hence $C \notin An(D)$, and thus we must have $C \in An(X)$. But then we get $C \in An(D)$ in $\mathcal{M}$ via the edge $X \rightarrow V$, a contradiction. □

Corollary 5.10. *For every MAG $\mathcal{M}$, the proper back-door graph $\mathcal{M}_{\mathbf{XY}}^{pbd}$ is also a MAG.*

For MAGs that are not adjustment amenable, the CBC might falsely indicate that an adjustment set exists even though that set may not be valid for some represented graph. Fortunately, adjustment amenability is easily tested using the graphical criteria of Lemma 5.5. For each child D of $\mathbf{X}$ in $PCP(\mathbf{X}, \mathbf{Y})$, we can test the visibility of all edges $\mathbf{X} \rightarrow D$ simultaneously using depth first search. This means that we can check all potentially problematic edges in time $O(n + m)$. If all tests pass, we are licensed to apply the CBC, as shown above. Hence, we can solve all algorithmic tasks in Table 1 for MAGs in the same way as for DAGs after an $O(k(n + m))$ check of adjustment amenability, where $k \leq |Ch(\mathbf{X})|$.

6 DISCUSSION

We have compiled efficient algorithms for solving several tasks related to m-separators in ancestral graphs, and applied those together with a new, constructive adjustment criterion to provide a complete and informative answer to the question when, and how, a desired causal effect can be estimated by covariate adjustment. Our results fully generalize to MAGs in the absence of selection bias. One may argue that the MAG result is more useful for exploratory applications (inferring a graph from data) than confirmatory ones (drawing a graph based on theory), as researchers will prefer drawing DAGs instead of MAGs due to the easier causal interpretation of the former. Nevertheless, in such settings the results can provide a means to construct more "robust" adjustment sets: If there are several options for covariate adjustment in a DAG, then one can by interpreting the same graph as a MAG possibly generate an adjustment set that is provably valid for a much larger class of DAGs. This might partially address the typical criticism that complete knowledge of the causal structure is unrealistic.

Our adjustment criterion generalizes the work of Shpitser et al. (2010) to MAGs and therefore now completely characterizes when causal effects are estimable by covariate adjustment in the presence of unmeasured confounders with multivariate exposures and outcomes. This also generalizes recent work by Maathuis and Colombo (2013) who provide a criterion which, for DAGs and MAGs without selection bias, is stronger than the back-door criterion but weaker than ours. They moreover show their criterion to hold also for CPDAGs and PAGs, which represent equivalence classes of DAGs and MAGs as they are constructed by causal discovery algorithms. It is possible that the constructive back-door criterion could be generalized further to those cases, which we leave for future work.

References

Silvia Acid and Luis M. de Campos. An algorithm for finding minimum d-separating sets in belief networks. In *Proceedings of UAI 1996*, pages 3–10, 1996.

Silvia Acid and Luis M. de Campos. Searching for bayesian network structures in the space of restricted acyclic partially directed graphs. *Journal of Artificial Intelligence Research (JAIR)*, 18:445–490, 2003.

Elias Barenboim, Jin Tian, and Judea Pearl. Recovering from selection bias in causal and statistical inference. In *Proceedings of AAAI-14*, 2014.

Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms, Second Edition*. The MIT Press, 2nd edition, September 2001. ISBN 0262032937.

Felix Elwert. *Graphical Causal Models*, pages 245–273. Handbooks of Sociology and Social Research. Springer, 2013.

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

Sander Greenland. Hierarchical regression for epidemiologic analyses of multiple exposures. *Environmental Health Perspectives*, 102 Suppl 8:33–39, Nov 1994.

Marloes H. Maathuis and Diego Colombo. A generalized backdoor criterion. arXiv:1307.5636, 2013.

Judea Pearl. *Causality*. Cambridge University Press, 2009. ISBN 0-521-77362-8.

Thomas Richardson and Peter Spirtes. Ancestral graph markov models. *Annals of Statistics*, 30:927–1223, 2002.

Kenneth J. Rothman, Sander Greenland, and Timothy L. Lash. *Modern Epidemiology*. Wolters Kluwer, 2008. ISBN 0781755646.

Ross D. Shachter. Bayes-ball: The rational pastime. In *Proceedings of UAI 1998*, pages 480–487, 1998.

Ilya Shpitser. Appendum to on the validity of covariate adjustment for estimating causal effects, 2012. unpublished manuscript.

Ilya Shpitser, Tyler VanderWeele, and James Robins. On the validity of covariate adjustment for estimating causal effects. In *Proceedings of UAI 2010*, pages 527–536. AUAI Press, 2010.

Ken Takata. Space-optimal, backtracking algorithms to list the minimal vertex separators of a graph. *Discrete Applied Mathematics*, 158:1660–1667, 2010.

Johannes Textor and Maciej Liśkiewicz. Adjustment criteria in causal diagrams: An algorithmic perspective. In *Proceedings of UAI*, pages 681–688, 2011.

Jin Tian, Azaria Paz, and Judea Pearl. Finding minimal d-separators. Technical Report R-254, University of California, Los Angeles, 1998. URL ftp.cs.ucla.edu/pub/stat_ser/r254.pdf.

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

Jiji Zhang. Causal reasoning with ancestral graphs. *Journal of Machine Learning Research*, 9:1437–1474, 2008.

A APPENDIX

A.1 AUXILIARY LEMMAS AND PROOFS

In this section, we prove Lemma 3.4 and several auxiliary Lemmas that are necessary for the proof of Theorem 5.8.

Proof of Lemma 3.4. Let us consider a proper walk $w = X, V_1, \dots, V_n, Y$ with $X \in \mathbf{X}, Y \in \mathbf{Y}$. If w does not contain a collider, all nodes V_i are in $Ant(\mathbf{X} \cup \mathbf{Y})$ and the walk is blocked by $\mathbf{Z}$, unless $\{V_1, \dots, V_n\} \cap \mathbf{R} = \emptyset$ in which case the walk is not blocked by $\mathbf{Z}_0$ either. If the walk contains colliders $\mathbf{C}$, it is blocked, unless $\mathbf{C} \subseteq \mathbf{Z} \subseteq \mathbf{R}$. Then all nodes V_i are in $Ant(\mathbf{X} \cup \mathbf{Y} \cup \mathbf{I})$ and the walk is blocked, unless $\{V_1, \dots, V_n\} \cap \mathbf{R} = \mathbf{C}$. Since $\mathbf{C} \subseteq \mathbf{Z}$ is a set of anteriors, there exists a shortest (possible containing 0 edges) path $\pi_j = V_j \to \dots \to W_j$ for each $V_j \in \mathbf{C}$ with $W_j \in \mathbf{X} \cup \mathbf{Y} \cup \mathbf{I}$ (it cannot contain an undirected edge, since there is an arrow pointing to V_j). Let $\pi'_j = V_j \to \dots \to W'_j$ be the shortest subpath of π_j that is not blocked by $\mathbf{Z}_0$. Let w' be the walk w after replacing each V_j by the walk $V_j \to \dots \to W'_j \leftarrow \dots \leftarrow V_j$. If any of the W_j is in $\mathbf{X} \cup \mathbf{Y}$ we truncate the walk, such that we get the shortest walk between nodes of $\mathbf{X}$ and $\mathbf{Y}$. Since π'_j is not blocked, w' contains no colliders except w'_j and all other nodes of w' are not in $\mathbf{R}$, w' is not blocked and $\mathbf{Z}_0$ is not a separator. □

Lemma A.1. *Given a DAG $\mathcal{G}$ and sets $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$ satisfying $\mathbf{Z} \cap Dpcp(\mathbf{X}, \mathbf{Y}) = \emptyset$, $\mathbf{Z}$ m-connects a proper non-causal path between $\mathbf{X}$ and $\mathbf{Y}$ if and only if it m-connects a proper non-causal walk between $\mathbf{X}$ and $\mathbf{Y}$.*

Proof. $\Leftarrow$: Let w be the m-connected proper non-causal walk. It can be transformed to an m-connected path π by removing loops of nodes that are visited multiple times. Since no nodes have been added, π remains proper, and the first edges of π and w are the same. So if w does not start with a $\to$ edge, π is non-causal. If w starts with an edge $X \to D$, there exists a collider with a descendant in $\mathbf{Z}$ which is in $De(D)$. So π has to be non-causal, or it would contradict $\mathbf{Z} \cap Dpcp(\mathbf{X}, \mathbf{Y}) = \emptyset$.

$\Rightarrow$: Let π be an m-connected proper non-causal path. It can be changed to an m-connected walk w by inserting $C_i \to \dots \to Z_i \leftarrow \dots \leftarrow C_i$ for every collider C_i on π and a corresponding $Z_i \in \mathbf{Z}$. Since no edges are removed from π, w is non-causal, but not necessarily proper, since the inserted walks might contain nodes of $\mathbf{X}$. However, in that case, w can be truncated to a proper walk w' starting at the last node of $\mathbf{X}$ on w. Then w' is non-causal, since it contains the subpath $\mathbf{X} \leftarrow \dots \leftarrow C_i$. □

In all of the below, $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ is a DAG, $\mathbf{Z}, \mathbf{L} \subseteq \mathbf{V}$ are disjoint, and $\mathcal{M} = \mathcal{G}[^{\mathbf{L}}_{\emptyset}]$.

Definition A.2 (Inducing path (Richardson and Spirtes, 2002))**.** *A path $\pi = V_1, \dots, V_{n+1}$ is called* inducing *with respect to $\mathbf{Z}, \mathbf{L}$ if all non-colliders on π except V_1 and V_{n+1} are in $\mathbf{L}$, and all colliders on π are in $An(\{V_1, V_{n+1}\} \cup \mathbf{Z})$.*

Every inducing path w.r.t. $\mathbf{Z}, \mathbf{L}$ is m-connected by $\mathbf{Z}$.

Lemma A.3 (Richardson and Spirtes (2002))**.** *If there is an inducing path w from $U \in \mathbf{V}$ to $V \in \mathbf{V}$ with respect to $\mathbf{Z}, \mathbf{L}$, then there exists no set $\mathbf{Z}'$ with $\mathbf{Z} \subseteq \mathbf{Z}' \subseteq (\mathbf{V} \setminus \mathbf{L})$ such that $\mathbf{Z}'$ d-separates U and V in $\mathcal{G}$ or m-separates U and V in $\mathcal{G}[^{\mathbf{L}}_{\emptyset}]$.*

Proof. This is Theorem 4.2, cases (v) and (vi), in Richardson and Spirtes (2002). □

Lemma A.4. *Two nodes U, V are adjacent in $\mathcal{G}[^{\mathbf{L}}_{\emptyset}]$ if and only if $\mathcal{G}$ contains an inducing path π between U and V with respect to $\emptyset, \mathbf{L}$. Moreover, the edge between U, V in $\mathcal{G}[^{\mathbf{L}}_{\emptyset}]$ can only have an arrowhead at U (V) if all such π have an arrowhead at U (V) in $\mathcal{G}$.*

Proof. The first part on adjacency is proved in (Richardson and Spirtes, 2002). For the second part on arrowheads, suppose π does not have an arrowhead at U, then π starts with an edge $U \to D$. Hence $D \notin An(U)$, so $D \in An(V)$ because π is an inducing path and therefore also $U \in An(V)$. Hence, the edge between U and V in $\mathcal{G}[^{\mathbf{L}}_{\emptyset}]$ must be $U \to V$. The argument for V is identical. □

Lemma A.5. *Suppose Z_0, Z_1, Z_2 is a path in $\mathcal{G}[^{\mathbf{L}}_{\emptyset}]$ on which Z_1 is a non-collider. Suppose an inducing path π_{01} from Z_0 to Z_1 w.r.t. $\emptyset, \mathbf{L}$ in $\mathcal{G}$ has an arrowhead at Z_1, and an inducing path π_{12} from Z_1 to Z_2 w.r.t. $\emptyset, \mathbf{L}$ has an arrowhead at Z_1. Then the walk $w_{012} = \pi_{01}\pi_{12}$ can be truncated to an inducing path from Z_0 to Z_2 w.r.t. $\emptyset, \mathbf{L}$ in $\mathcal{G}$.*

Proof. The walk w_{012} does not contain more non-colliders than those on π_{01} or π_{12}, so they must all be in $\mathbf{L}$. It remains to show that the colliders on w_{012} are in $An(Z_0 \cup Z_2)$. Because Z_1 is not a collider on Z_0, Z_1, Z_2, at least one of the edges Z_0, Z_1 and Z_1, Z_2 must be a directed edge pointing away from Z_1. Assume without loss of generality that $Z_0 \leftarrow Z_1$ is that edge. Then all colliders on π_{01} are in $An(Z_0 \cup Z_1) = An(Z_0) \subseteq An(Z_0 \cup Z_2)$, and all colliders on π_{12} are in $An(Z_1 \cup Z_2) \subseteq An(Z_0 \cup Z_2)$. Z_1 itself is a collider on w_{012} and is also in $An(Z_0)$. Hence, the walk w_{012}

is d-connected, and can be truncated to an inducing path that starts with the first arrow of π_{01} and ends with the last arrow of π_{12}. □

Definition A.6 (Inducing $\mathbf{Z}$-trail)**.** *Let $\pi = V_1, \dots, V_{n+1}$ be a path in $\mathcal{G}[^{\mathbf{L}}_{\emptyset}$ such that $V_2, \dots, V_n \in \mathbf{Z}$, $V_1, V_{n+1} \notin \mathbf{Z}$, and for each $i \in \{1, \dots, n\}$, there is an inducing path w.r.t. $\emptyset, \mathbf{L}$ linking V_i, V_{i+1} that has an arrowhead at V_i (V_{i+1}) if $V_i \in \mathbf{Z}$ ($V_{i+1} \in \mathbf{Z}$). Then π is called an inducing $\mathbf{Z}$-trail.*

Lemma A.7. *Let $\pi = V_1, \dots, V_{n+1}$ be an inducing $\mathbf{Z}$-trail, and let π' be a subsequence of π formed by removing one node V_i of π such that $V_i \in \mathbf{Z}$ is a non-collider on π. Then π' is an inducing $\mathbf{Z}$-trail.*

Proof. According to Lemma A.5, if V_i is a non-collider on π, then V_{i-1} and V_{i+1} are linked by an inducing path π that contains an arrowhead at V_{i-1} (V_{i+1}) if $V_{i-1} \in \mathbf{Z}$ ($V_{i+1} \in \mathbf{Z}$). Therefore, V_{i-1} and V_{i+1} are themselves adjacent, π' is a path, and is a $\mathbf{Z}$-trail. □

Corollary A.8. *Every inducing $\mathbf{Z}$-trail $\pi = V_1, \dots, V_{n+1}$ has a subpath π' that is m-connected by $\mathbf{Z}$.*

Proof. Transform π into π' by replacing non-collider nodes in $\mathbf{Z}$ by the direct edge linking their neighbours until no such node exists anymore. By inductively applying Lemma A.7, we see that π' is also an inducing $\mathbf{Z}$-trail, and every node in $\mathbf{Z}$ is a collider because otherwise we would have continued transforming. So π' must be m-connected by $\mathbf{Z}$. □

Lemma A.9. *Let $w_{\mathcal{G}}$ be a walk from X to Y in $\mathcal{G}$, $X, Y \notin \mathbf{L}$, that is d-connected by $\mathbf{Z}$. Let $w_{\mathcal{M}} = V_1, \dots, V_{n+1}$ be the subsequence of $w_{\mathcal{G}}$ consisting only of the nodes in $\mathcal{M} = \mathcal{G}[^{\mathbf{L}}_{\emptyset}$. Then $\mathbf{Z}$ m-connects X and Y in $\mathcal{M}$ via a path along a subsequence $w'_{\mathcal{M}}$ formed from $w_{\mathcal{M}}$ by removing some nodes in $\mathbf{Z}$ (possibly $w'_{\mathcal{M}} = w_{\mathcal{M}}$).*

Proof. First, truncate from $w_{\mathcal{M}}$ all subwalks between nodes in $\mathbf{Z}$ that occur more than once. Now consider all subsequences $V_1, \dots, V_{n+1}$, $n > 1$, of $w_{\mathcal{M}}$ where $V_2, \dots, V_n \in \mathbf{Z}$, $V_1, V_{n+1} \notin \mathbf{Z}$, which now are all paths in $w_{\mathcal{M}}$. On those subsequences, every V_i must be adjacent in $\mathcal{G}$ to V_{i+1} via a path containing no colliders, and all non-endpoints on that path must be in $\mathbf{L}$. So there are inducing paths w.r.t. $\emptyset, \mathbf{L}$ between all V_i, V_{i+1}, which have arrowheads at V_i (V_{i+1}) if $V_i \in \mathbf{Z}$ ($V_{i+1} \in \mathbf{Z}$). So $V_1, \dots, V_{n+1}$ is an inducing $\mathbf{Z}$-trail, and has a subpath which m-connects V_1, V_{n+1} given $\mathbf{Z}$. Transform $w_{\mathcal{M}}$ to $w'_{\mathcal{M}}$ by replacing all inducing $\mathbf{Z}$-trails by their m-connected subpaths. According to Lemma A.4, non-colliders on $w_{\mathcal{M}}$ cannot be colliders on $w'_{\mathcal{M}}$, as bypassing inducing paths can remove but not create arrowheads. Moreover, all nodes in $\mathbf{Z}$ on $w'_{\mathcal{M}}$ are colliders. Hence $w'_{\mathcal{M}}$ is m-connected by $\mathbf{Z}$. □

Corollary A.10. *Each edge on $w'_{\mathcal{M}}$ as defined above corresponds to an inducing path w.r.t $\emptyset, \mathbf{L}$ in $\mathcal{G}$ along nodes on $w_{\mathcal{G}}$.*

Lemma A.11. *Suppose there exists an inducing path π_{01} from Z_0 to Z_1 w.r.t. $\mathbf{S}, \mathbf{L}$ with an arrowhead at Z_1 and an inducing path from Z_1 to Z_2 w.r.t. $\mathbf{S}', \mathbf{L}$ with an arrowhead at Z_1. Then the walk $w_{012} = \pi_{01}\pi_{12}$ can be truncated to an inducing path from Z_0 to Z_2 w.r.t. $\mathbf{S} \cup \mathbf{S}' \cup \{Z_1\}, \mathbf{L}$ in $\mathcal{G}$.*

Proof. The walk w_{012} does not contain more non-colliders than those on π_{01} or π_{12}, so they must all be in $\mathbf{L}$. All colliders on $\pi_{0,1}$ and $\pi_{1,2}$ as well as Z_1 are in $An(Z_0, Z_1, Z_2, \mathbf{S}, \mathbf{S}')$, and therefore also all colliders of w_{012}.

Hence, the walk w_{012} is d-connected, and can be truncated to an inducing path that starts with the first arrow of π_{01} and ends with the last arrow of π_{12}. □

Lemma A.12. *Suppose $Z_0, Z_1, \dots, Z_{k+1}$ is a path in $\mathcal{G}[^{\mathbf{L}}_{\emptyset}$ with an arrowhead at Z_{k+1} on which all $Z_1, \dots, Z_k$ are colliders. Then there exists an inducing path from Z_0 to Z_{k+1} w.r.t. $\{Z_1, \dots, Z_k\}, \mathbf{L}$ with an arrowhead at Z_{k+1}.*

Proof. Because all Z_i, Z_{i+1} are adjacent and all $Z_1, \dots, Z_k$ are colliders there exist inducing paths $\pi_{i,i+1}$ w.r.t. $\emptyset, \mathbf{L}$ from Z_i to Z_{i+1} that have arrowheads at $Z_1, \dots, Z_k$ (Lemma A.4). The claim follows by repeatedly applying Lemma A.11 to the $\pi_{i,i+1}$'s. □

Lemma A.13. *Suppose $A \to V_1 \leftrightarrow \dots \leftrightarrow V_k \leftrightarrow X \to D$ or $A \leftrightarrow V_1 \leftrightarrow \dots \leftrightarrow V_k \leftrightarrow X \to D$ is a path in $\mathcal{G}[^{\mathbf{L}}_{\emptyset}$ (possibly $k = 0$), each V_i is a parent of D and there exists an inducing path π_{XD} from X to D w.r.t $\emptyset, \mathbf{L}$ that has arrowheads on both ends. Then A and D cannot be m-separated in $\mathcal{G}[^{\mathbf{L}}_{\emptyset}$.*

Proof. Assume the path is $A \to V_1 \leftrightarrow \dots \leftrightarrow V_k \leftrightarrow X \to D$. The case where the path starts with $A \leftrightarrow V_1$ can be handled identically, since the first arrowhead does not affect m-separation.

Assume A and D can be m-separated in $\mathcal{G}[^{\mathbf{L}}_{\emptyset}$, and let $\mathbf{Z}$ be such a separator. If V_1 is not in $\mathbf{Z}$ then the path $A \to V_1 \to D$ is not blocked, so $V_1 \in \mathbf{Z}$. Inductively it follows, if V_i is not in $\mathbf{Z}$, but all $\forall j < i : V_j \in \mathbf{Z}$ then the path $A \to V_1 \leftrightarrow \dots \leftrightarrow V_{i-1} \leftrightarrow V_i \to D$ is not blocked, so $V_i \in \mathbf{Z}$ for all i.

There exist an inducing path π_{AX} from A to X with an arrowhead at X w.r.t. to $\{V_1, \dots, V_k\}, \mathbf{L}$ (Lemma A.12) which can be combined with π_{XD} to an inducing path from A to D w.r.t. to $\{V_1, \dots, V_k, X\}, \mathbf{L}$ (Lemma A.11).

Hence no m-separator of A, D can contain $\{X, V_1, \dots, V_k\}$ (Lemma A.3). Then there cannot exist an m-separator, because every separator must include $V_1, \dots, V_k$ and the path

$A \to V_1 \leftrightarrow V_2 \leftrightarrow \ldots \leftrightarrow V_k \leftrightarrow X \to D$ is open without $X \in \mathbf{Z}$. □

A.2 ALGORITHMS

This section contains algorithm pseudocodes and parts of their correctness proofs that were omitted from the main text for space reasons.

A.2.1 TESTING

For a given ancestral graph $\mathcal{G}$ the problem TESTSEP can be solved with a modified Bayes-Ball algorithm in time $O(n+m)$. In the algorithm every bi-directed edge $A \leftrightarrow B$ is considered as a pair of edges $A \leftarrow \cdot \rightarrow B$ and an undirected edge $A - B$ as a directed edge pointing to the currently visited node.

function TESTSEP($\mathcal{G}, \mathbf{X}, \mathbf{Y}, \mathbf{Z}$)
  Run Bayes-Ball from $\mathbf{X}$
  return ($\mathbf{Y}$ not reachable)

Figure 5: TestSep

The problem TESTMINSEP can be solved using Algorithm 6 TESTMINSEP in $O(|\mathbf{E}^m_{An}|) = O(n^2)$ time. Alternatively, the problem can be solved with an algorithm that iteratively removes from $\mathbf{Z}$ nodes and tests if the resulting set remains an m-separator. This can be done in time $O(n(n+m))$. The correctness of the algorithms for TESTMINSEP can be shown by generalizing the results presented in (Tian et al., 1998) for m-separation. 6 TESTMINSEP, runs in $O(|\mathbf{E}^m_{An}|)$ because R_x and R_y can be computed with an ordinary search that aborts when a node in $\mathbf{Z}$ is reached.

function TESTMINSEP($\mathcal{G}, \mathbf{X}, \mathbf{Y}, \mathbf{Z}$)
  if $\mathbf{Z} \setminus Ant(\mathbf{X} \cup \mathbf{Y}) \neq \emptyset$ **then return** false
  if not TESTSEP($\mathcal{G}, \mathbf{X}, \mathbf{Y}, \mathbf{Z}$) **then**
    return false
  $\mathcal{G}'^a \leftarrow \mathcal{G}^a_{Ant(\mathbf{X}\cup\mathbf{Y})}$
  $R_x \leftarrow \{Z \in \mathbf{Z} \mid \exists$ path $X - Z$ in $\mathcal{G}'^a$
      not intersecting $\mathbf{Z} \setminus \{Z\}\}$
  if $Z \notin R_x$ **then return** false
  $R_y \leftarrow \{Z \in \mathbf{Z} \mid \exists$ path $Y - Z$ in $\mathcal{G}'^a$
      not intersecting $\mathbf{Z} \setminus \{Z\}\}$
  if $Z \notin R_y$ **then return** false
  return true

Figure 6: TestMinSep

A.2.2 FINDING AN M-SEPARATOR

The problem can be solved using Algorithm 7 FINDSEP in $O(n+m)$ time. The correctness follows directly from Lemma 3.4.

function FINDSEP($\mathcal{G}, \mathbf{X}, \mathbf{Y}, \mathbf{I}, \mathbf{R}$)
  $\mathbf{R}' \leftarrow \mathbf{R} \setminus (\mathbf{X} \cup \mathbf{Y})$
  $\mathbf{Z} \leftarrow Ant(\mathbf{X}, \mathbf{Y}, \mathbf{I}) \cap \mathbf{R}'$
  if TESTSEP($\mathcal{G}, \mathbf{X}, \mathbf{Y}, \mathbf{Z}$) **then**
    return $\mathbf{Z}$
  else
    return $\perp$

Figure 7: FindSep

A.2.3 FINDING A MINIMAL M-SEPARATOR

For a given AG $\mathcal{G}$ the problem FINDMINSEP can be solved with algorithm 8 FINDMINSEPNAIVE in $O(|Ant(\mathbf{X} \cup \mathbf{Y})||E_{An}|) = O(n(n+m))$ or algorithm 9 FINDMINSEPMORAL in $O(|\mathbf{E}^m_{An}|) = O(n^2)$ time.

function FINDMINSEPNAIVE($\mathcal{G}, \mathbf{X}, \mathbf{Y}, \mathbf{I}, \mathbf{R}$)
  $\mathcal{G}' \leftarrow \mathcal{G}_{Ant(\mathbf{X}\cup\mathbf{Y}\cup\mathbf{I})}$
  $\mathbf{Z} \leftarrow \mathbf{R} \cap Ant(\mathbf{X} \cup \mathbf{Y} \cup \mathbf{I})$
  if not TESTSEP($\mathcal{G}', \mathbf{X}, \mathbf{Y}, \mathbf{Z}$) **then**
    return $\perp$
  for all U in $\mathbf{Z} \setminus \mathbf{I}$ **do**
    if TESTSEP($\mathcal{G}', \mathbf{X}, \mathbf{Y}, \mathbf{Z} \setminus \{U\}$) **then**
      $\mathbf{Z} \leftarrow \mathbf{Z} \setminus \{U\}$
  return $\mathbf{Z}$

Figure 8: FindMinSepNaive

Algorithm 8 FINDMINSEPNAIVE depends on an implicit moral graph and the fact that in an undirected graph every node that cannot be removed from a separating set has to be in separating subsets, and runs in $O(|Ant(\mathbf{X} \cup \mathbf{Y})||\mathbf{E}_{An}|)$.

function FINDMINSEPMORAL($\mathcal{G}, \mathbf{X}, \mathbf{Y}, \mathbf{I}, \mathbf{R}$)
  $\mathcal{G}' \leftarrow \mathcal{G}_{Ant(\mathbf{X}\cup\mathbf{Y}\cup\mathbf{I})}$
  $\mathcal{G}'^a \leftarrow \mathcal{G}^a_{Ant(\mathbf{X}\cup\mathbf{Y}\cup\mathbf{I})}$
  $\mathbf{Z}' \leftarrow \mathbf{R} \cap Ant(\mathbf{X} \cup \mathbf{Y})$
  Remove from $\mathcal{G}'^a$ all nodes of $\mathbf{I}$
  if not TESTSEP($\mathcal{G}', \mathbf{X}, \mathbf{Y}, \mathbf{Z}$) **then**
    return $\perp$
  Run BFS from $\mathbf{X}$. Whenever a node in $\mathbf{Z}'$ is met, mark it, if it is not already marked and do not continue along the path. When BFS stops, let $\mathbf{Z}''$ be the set of all marked nodes. Remove all markings
  Run BFS from $\mathbf{Y}$. Whenever a node in $\mathbf{Z}''$ is met, mark it, if it is not already marked and do not continue along the path. When BFS stops, let $\mathbf{Z}$ be the set of all marked nodes.
  return $\mathbf{Z} \cup \mathbf{I}$

Figure 9: FindMinSepMoral

Algorithm 9 FINDMINSEPMORAL begins with the separating set $\mathbf{R} \cap Ant(\mathbf{X} \cup \mathbf{Y})$ and finds a subset satisfying the conditions tested by algorithm 6 TESTMINSEP, in $O(|\mathbf{E}^m_{An}|)$.

A.2.4 FINDING A MINIMUM COST M-SEPARATOR

The problem MINCOSTSEP can be solved with algorithm 10 FINDMINCOSTSEP in $O(n^3)$.

function FINDMINCOSTSEP($\mathcal{G}, \mathbf{X}, \mathbf{Y}, \mathbf{I}, \mathbf{R}, w$)
  $\mathcal{G}' \leftarrow \mathcal{G}_{Ant(\mathbf{X}\cup\mathbf{Y}\cup\mathbf{I})}$
  $\mathcal{G}'^a \leftarrow \mathcal{G}^a_{Ant(\mathbf{X}\cup\mathbf{Y}\cup\mathbf{I})}$
  Add a node X^m connected to all nodes in $\mathbf{X}$, and a node Y^m connected to all nodes in $\mathbf{Y}$.
  Assign infinite cost to all nodes in $\mathbf{X} \cup \mathbf{Y} \cup (\mathbf{V} \setminus \mathbf{R})$ and cost $w(Z)$ to every other node Z.
  Remove all nodes of $\mathbf{I}$ from $\mathcal{G}'^a$.
  Change the graph to a flow network as described in Cormen et al. (2001) and return a minimum cutset $\mathbf{Z}$.

Figure 10: FindMinCostSep

The correctness without $\mathbf{I}$ follows from the fact that a minimum set is a minimal set and the minimal cut found in the ancestor moral graph is therefore the minimal separating set. The handling of $\mathbf{I}$ is shown in Acid and de Campos (1996).

A.2.5 ENUMERATING ALL MINIMAL M-SEPARATORS

The problem LISTMINSEP can be solved with algorithm 11 LISTMINSEP with $O(n^3)$ delay between every outputted $\mathbf{Z}$.

function LISTMINSEP($\mathcal{G}, \mathbf{X}, \mathbf{Y}, \mathbf{I}, \mathbf{R}$)
  $\mathcal{G}' \leftarrow \mathcal{G}_{Ant(\mathbf{X}\cup\mathbf{Y}\cup\mathbf{I})}$
  $\mathcal{G}'^a \leftarrow \mathcal{G}^a_{Ant(\mathbf{X}\cup\mathbf{Y}\cup\mathbf{I})}$
  Add a node X^m connected to all $\mathbf{X}$ nodes.
  Add a node Y^m connected to all $\mathbf{Y}$ nodes.
  Remove all nodes of $\mathbf{I}$.
  Remove all nodes of $\mathbf{V} \setminus \mathbf{R}$, but insert additional edges connecting the neighbours.
  of all removed nodes.
  Use the algorithm in Takata (2010) to list all sets separating X^m and Y^m.

Figure 11: ListMinSep

The correctness is shown by Textor and Liśkiewicz (2011) for adjustment sets and generalizes directly to m-separators, because after moralization, both problems are equivalent to enumerating vertex cuts of an undirected graph. The handling of $\mathbf{I}$ is shown by Acid and de Campos (1996).

A.2.6 TESTING FOR ADJUSTMENT AMENABILITY

Let $N(V)$ denote all nodes adjacent to V, and $Sp(V)$ denote all spouses of V, i.e., nodes W such that $W \leftrightarrow V \in \mathbf{E}$. The adjustment amenability of a graph $\mathcal{G}$ w.r.t sets $\mathbf{X}, \mathbf{Y}$ can be tested with the following algorithm:

function TESTADJUSTMENTAMENABILITY($\mathcal{G}, \mathbf{X}, \mathbf{Y}$)
  for all D in $Ch(\mathbf{X}) \cap PCP(\mathbf{X}, \mathbf{Y})$ **do**
    $\mathbf{C} \leftarrow \emptyset$
    $\mathbf{A} \leftarrow \emptyset$
    function CHECK(V)
      if $\mathbf{C}[V]$ **then return** $\mathbf{A}[V]$
      $\mathbf{C}[V] \leftarrow$ true
      $\mathbf{A}[V] \leftarrow ((Pa(V) \cup Sp(V)) \setminus N(D) \neq \emptyset)$
      for all $W \in Sp(V) \cap Pa(D)$ **do**
        if CHECK(W) **then** $\mathbf{A}[V] \leftarrow$ true
      return $\mathbf{A}[V]$
    for all X in $\mathbf{X} \cap Pa(D)$ **do**
      if $\neg$CHECK(X) **then**
        return false

Figure 12: TestAdjustmentAmenability

The algorithm checks for every edge $X \rightarrow D$ on a proper causal path to $\mathbf{Y}$ whether it satisfies the amenability conditions of Lemma 5.5 by searching a collider path through the parents of D to a node Z not connected to D; note that condition (1) of Lemma 5.5 is identical to condition (2) with an empty collider path. Since CHECK performs a depth-first-search by checking every node only once and then continuing to its neighbors, each iteration of the outer for-loop in the algorithm runs in linear time $O(n + m)$. Therefore, the entire algorithm runs in $O(k(n + m))$ where $k \leq |Ch(\mathbf{X})|$.

Propensity Score Matching for Causal Inference with Relational Data

David Arbour **Katerina Marazopoulou** **Dan Garant** **David Jensen**
School of Computer Science
University of Massachusetts Amherst
Amherst, MA 01003
{darbour, kmarazo, dgarant, jensen}@cs.umass.edu

Abstract

Propensity score matching (PSM) is a widely used method for performing causal inference with observational data. PSM requires fully specifying the set of confounding variables of treatment and outcome. In the case of relational data, this set may include non-intuitive relational variables, i.e., variables derived from the relational structure of the data. In this work, we provide an automated method to derive these relational variables based on the relational structure and a set of naive confounders. This automatic construction includes two unusual classes of variables: relational degree and entity identifiers. We provide experimental evidence that demonstrates the utility of these variables in accounting for certain latent confounders. Finally, through a set of synthetic experiments, we show that our method improves the performance of PSM for causal inference with relational data.

1 INTRODUCTION

Propensity score matching (PSM) [Rosenbaum and Rubin, 1983] is a widely used tool for determining causal effects from observational data. Propensity scores summarize the effects of a potentially large number of confounding variables by creating a *predictive* model of treatment. The computation of a propensity score requires specifying a set of potentially confounding variables. This task is relatively straightforward for propositional (i.i.d.) data. However, many causal analyses consider data in which treatment, outcome, and potential confounders can arise from the interactions among multiple types of interrelated entities. Propensity score matching becomes substantially more challenging in such relational data.

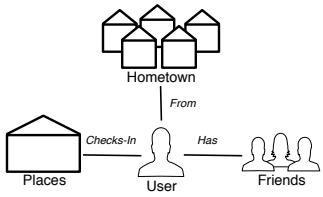


Figure 1: Example of relational data: users are friends with other users, each user comes from a hometown, and users check-in at places.

To illustrate this, consider the example domain shown in Figure 1, depicting a plausible relational domain. Foursquare is an example of a real system that could produce this sort of data. Suppose a researcher is interested in using data from this domain to assess whether smoking causes a user to gain weight. One approach would be to construct a propensity score model with user attributes that the researcher believes could be causes of whether a user smokes and the user's weight, such as alcohol consumption and ethnicity:

$$[User].Smokes \sim [User].Drinks + [User].Ethnicity.$$

While this accounts for attributes associated with the user, it fails to account for possible confounders derived from relational variables. For example, it is plausible that the alcohol consumption of a user's friends is a common cause of $[User].Weight$ and $[User].Smokes$. To account for these effects, the corresponding relational variables should be included in the propensity score model.

It is not difficult to envision more complicated relational variables having an effect. In fact, as previous work has shown [Maier et al., 2013b], the number of relational variables can be arbitrarily large depending on how many entity and relationship types exist in the network, the size of the network, and the length of the longest path (the largest degree of separation) in the network where direct dependence exists.

An additional level of complexity introduced by rela-

tional data is that relational structures may result in multiple instances of a given variable. For example, a user with multiple friends could be influenced by the drinking behaviour of each of those friends. Typically, an aggregation function, such as `mean`, is used to combine this set of values into a single value. Properly conditioning on a relational variable entails choosing the correct set of aggregation functions to represent the distribution of values contained in the set. For example, in order to condition on a relational variable, it may be necessary to condition on multiple aspects of the distribution of those values, such as the `mean` and the standard deviation (`stdev`).

To address these issues, we introduce relational propensity score matching (RPSM), a method that applies propensity score matching to relational domains. RPSM leverages the framework of relational models [Getoor and Taskar, 2007, Maier et al., 2013b] to automatically construct the set of possible relational confounders given a simpler specification of the assumed dependency structure. RPSM also identifies opportunities to use relational degree variables and entity identifiers, which, as we show empirically, can reduce the bias arising from latent relational confounders. We evaluate RPSM via a set of synthetic experiments using the relational structure of a real-world relational domain, Foursquare.

2 BACKGROUND

In this section we provide a brief overview of matching methods and propensity scores. We then introduce the relational concepts necessary to formalize RPSM.

2.1 MATCHING

In the framework of potential outcomes [Rubin, 1974], estimating the causal effect of treatment T on variable Y is formalized as a comparison of potential outcomes. More formally, let T_i be a binary treatment variable for unit i and let Y_i be the outcome variable for unit i, where $i \in \{1, \ldots, n\}$. $Y_i(T_i = 0)$ denotes the value of Y_i that would be observed if no treatment was applied to unit i. Similarly, $Y_i(T_i = 1)$ is the value of Y_i that would be observed if unit i had received treatment. The causal effect of T on Y is estimated by comparing the difference $Y_i(T_i = 1) - Y_i(T_i = 0)$ across all units i.

In practice, a specific unit either receives treatment or not. Therefore, for a given value of i we never know both $Y_i(T = 1)$ and $Y_i(T = 0)$. Experimental studies often randomly assign units to treatment and control groups, so that the expected distribution of the covariates in these groups is identical. In observational studies, where randomization is not possible, *matching* can be used to pair similar samples from the treated and the control groups. Matching can be generally defined as a method that aims to approximate random assignment by equating the distribution of covariates in the treated and control group [Stuart, 2010].

Matching requires a measure quantifying how similar two individuals are. This is achieved by (1) selecting a set of features to be used in the computation of similarity, and (2) choosing a similarity function to apply on those features (for example Mahalanobis distance, propensity score, etc.). Once a similarity measure has been chosen, individuals are matched based on this measure. There are multiple methods for performing matching (see Stuart [2010] and Ho et al. [2007] for a survey of matching methods). In this paper, we employ full matching [Hansen and Klopfer, 2006], which creates a collection of matched sets (the size of the collection is chosen automatically). Each matched set contains at least one treated and one control unit. Full matching has been shown to be optimal with respect to similarity within matched sets [Rosenbaum, 1991].

Matching methods make the assumption of *ignorable treatment assignment*, i.e., treatment assignment is independent of the outcome given the observed covariates. This assumption guides the selection of appropriate covariates for the computation of similarity.

2.2 PROPENSITY SCORE

The propensity score [Rosenbaum and Rubin, 1983] is the probability of receiving treatment, given the observed covariates $\mathbf{X}_i$

$$e_i(\mathbf{X}_i) = P(T_i = 1|\mathbf{X}_i).$$

Propensity scores are a form of dimensionality reduction that projects the original covariates down to a single value which preserves distance with respect to the likelihood of treatment. Matching can then be performed on the propensity score, as opposed to the covariates directly. The prevailing explanation for why propensity scores are appropriate for matching is that they are balancing scores (given the value of the propensity score, the treatment and control groups have the same distribution of covariates), and they preserve ignorability of treatment assignment (if treatment assignment is ignorable given the covariates, then treatment assignment is also ignorable given the propensity score) [Stuart, 2010].

Any method that models the conditional probability of a binary variable given a set of predictors can be used to estimate a propensity score. In this work, we employ logistic regression, a widely used method for obtaining a propensity score. However, other models (such as boosted trees, support vector machines,

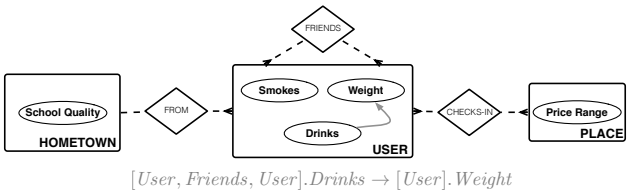


Figure 2: Relational model for the Foursquare domain. The underlying relational schema (ER diagram) is shown in black. The attributes on the entities are fictional. The relational dependency is shown in gray.

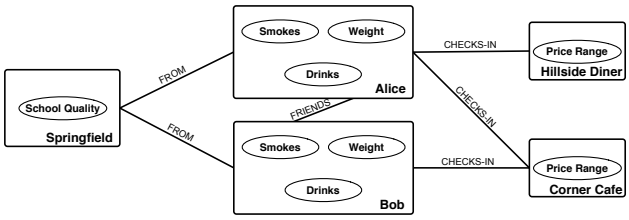


Figure 3: Example relational skeleton for the Foursquare domain. This could be a small fragment of a (potentially) larger skeleton.

and neural networks) have been explored in the literature [Westreich et al., 2010, McCaffrey et al., 2004, Lee et al., 2010].

A key advantage of propensity scores is their robustness to model misspecification [Drake, 1993], i.e., including irrelevant variables[1] in the calculation of the propensity score. Because the propensity score model is built upon a *predictive* rather than causal model of treatment, many of the issues that arise with traditional regression modeling, such as multicollinearity, are no longer a threat to validity. Further, in contrast to matching directly on the covariates, propensity scores can down-weight or disregard variables that are not associated with treatment and have been erroneously included in the propensity model. However, as Pearl [2009] has observed, common effects of the treatment and outcome must not be included in the propensity score model. In general, the set of d-connecting paths between treatment and outcome needs to be considered. The propensity score model must include a (not necessarily minimal) separating set of treatment and outcome. One approach to eliminating variables that are potential common effects of treatment and outcome is the injunction of Rosenbaum and Rubin [1983] to restrict the set of covariates to pretreatment variables (variables whose values are measured prior to treatment).

2.3 RELATIONAL CONCEPTS

Propositional representations, such as Bayesian networks, describe domains with a single entity type. However, many real-world systems involve multiple types of entities that interact with each other. Data produced by such systems are called *relational* or *network data*. In this section, we introduce the basic relational concepts, following the notation and terminology of Maier et al. [2013b].

A *relational schema* $\mathcal{S} = (\mathcal{E}, \mathcal{R}, \mathcal{A}, card)$ specifies the set of entity, relationship, and attribute classes of a domain. It includes a cardinality function that imposes constraints on the number of times an entity instance can participate in a relationship. A relational schema can be graphically represented with an Entity-Relationship (ER) diagram. Figure 2 shows the ER diagram for the Foursquare domain. In this example, there are three entity classes (*User*, *Place*, *Hometown*), and three relationship classes, (*Friends*, *ChecksIn*, *From*). The entity class *User* has three attributes: *Smokes*, *Weight*, and *Drinks*. The cardinality constraints are depicted using crow's feet notation. For example, the cardinality of the *From* relationship is one-to-many, indicating that one user has one hometown, but many users can be from the same hometown.

A *relational skeleton* is a partial instantiation of a relational schema that specifies the set of entity and relationship instances that exist in the domain. Figure 3 depicts an example relational skeleton for the Foursquare domain. The network consists of two *User* instances, Alice and Bob, who are friends with each other and come from the same hometown. There are two *Place* instances, Hillside Diner and Corner Cafe.

Given a relational schema, one can specify *relational paths*, which intuitively correspond to possible ways of traversing the schema (see Maier et al. [2013b] for a formal definition). For the schema shown in Figure 2, possible paths include $[User, Friends, User]$ (a person's friends), and $[User, Friends, User, From, Hometown]$ (the hometowns of a person's friends). *Relational variables* consist of a relational path and an attribute that can be reached through that path. For example, the relational variable $[User, Friends, User].Drinks$ corresponds to the alcohol consumption of a person's friends. Probabilistic dependencies can be defined between relational variables. In this work, we consider dependencies where the path of the outcome relational variable is a single item. In this case, the path of the treatment relational variable describes how dependence is induced. For example, the *relational dependency*

$$[User, Friends, User].Drinks \rightarrow [User].Weight$$

[1] Variables that are marginally independent of treatment or outcome.

states that the alcohol consumption of a user's friends affects that user's weight.

A *relational model* $\mathcal{M} = (\mathcal{S}, \mathcal{D}, \Theta)$ is a collection of relational dependencies $\mathcal{D}$ defined over a relational schema along with their parameterizations Θ (a conditional probability distribution for each attribute given its parents). The structure of a relational model can be depicted by superimposing the dependencies on the ER diagram of the relational schema, as shown in Figure 2, and labeling each arrow with the dependency it corresponds to. If labels are omitted, the resulting graphical representation is known as a *class-dependency graph*.

Recent work by Maier et al. [2013b] provides a framework that enables reasoning about d-separation in relational models. Toward that end, they introduce *abstract ground graphs* (AGGs), a graphical structure that captures relational dependencies and can be used to answer relational d-separation queries. Abstract ground graphs are defined from a given perspective, the base item of the analysis, and include nodes that correspond to relational variables. For practical applications, the size of abstract ground graphs is limited by a (domain dependent) hop-threshold, which constrains the length of relational paths that will be considered. Intuitively, the hop-threshold corresponds to the relational "distance" of a cause from its effect.

2.4 NEW TYPES OF VARIABLES

In this section we present the new types of variables that are enabled by relational domains: (1) Relational variables (a way of defining a larger number of potential confounders) and aggregation; (2) Degree variables (a type of confounder not available without relational data); (3) Entity identifiers (which enable blocking, a way to account for latent confounders only available within relational data). Those types of variables are used in the calculation of relational propensity scores and are referred to as *relational covariates*.

2.4.1 Aggregation Functions

A fundamental characteristic of relational data is the heterogeneity of the underlying relational structure. For example, a person can have many friends, different people have different sets of friends, and those sets can overlap to varying degrees. This implies that when constructing relational variables for a specific individual, the construction process will often return a set of values rather than a single value. For instance, the relational variable "friends' age" for a person consists of a set of values containing the age of each one of that person's friends. In the field of statistical relational learning, aggregation functions are commonly used to summarize the values of related instances into a single value, representative of the distribution. Common aggregation functions include `mean`, `stdev`, `mode`, `count`, `sum`, `min`, `max`, and `median`. Researchers have also defined more complex aggregation methods [Perlich and Provost, 2006].

Figure 4: Relational schema that depicts a hierarchy. A state has many towns, but each town is in one state, and many people are from the same town, but each person is from one town.

2.4.2 Degree Variables

Other work has pointed out that variation in the size of the set of values for a relational variable can strongly affect the distribution of the observed values of many aggregation functions [Jensen et al., 2003]. Jensen et al. call the size of this set the "degree" and it is equivalent, in the terminology of Maier et al. [2013b], to the size of the terminal set of a relational path. To account for the effects of degree on aggregated values, RPSM includes degree variables in the calculation of propensity scores.

2.4.3 Entity Identifiers

Blocking designs are widely used in experimental studies to account for latent confounders [Fisher, 1935]. Rattigan et al. [2011] formalized *relational blocking* as an operator that can be used to infer causal dependence in observational data expressed in a relational representation. By blocking on the identifier of an entity, relational blocking accounts for the effect of latent variables associated with that entity. Blocking is uniquely available for relational data. Moreover, since blocking on an entity appears to avoid inducing dependence due to colliders on that entity, blocking may partially alleviate a key threat to validity noted by Pearl [2009].

In this work, we incorporate relational blocking with propensity scores by including entity identifiers as covariates in the calculation of propensity scores. We restrict the use of blocking to hierarchies, i.e., parts of the relational schema that are connected through a series of many-to-one relationships. An example hierarchy is shown in Figure 4. In this case, blocking on the identifier for towns (i.e., grouping users based on their hometown) accounts for the effect of latent variables associated with *Hometown*, and for the effect of latent variables associated with the *State* within which

each town is located. More generally, blocking on the identifier of an entity in a hierarchy accounts for the effect of latent confounders that reside in that entity and in entities that appear higher up in the hierarchy.

3 RELATIONAL PROPENSITY SCORE MATCHING

We consider the following problem: given an entity E and two attributes on that entity, treatment $[E].T$ and outcome $[E].O$, we seek to decide between $[E].T \rightarrow [E].O$ and $[E].T \not\rightarrow [E].O$. For notational convenience, we restrict our attention to cases where the treatment and outcome are on the same entity. In practice, RPSM can be applied to any treatment and outcome lying on entities that are connected through one-to-one relationships. We assume that the relational skeleton has been given *a priori*, i.e., all entity and relationship instances have been fully and correctly specified. Additionally, we assume that the effects of all latent variables can be accounted for by using relational blocking (in other words, latent variables exists only on paths that can be blocked on).

Relational propensity score matching (RPSM) provides an automatic method for constructing the set of aggregated relational variables, degree variables and entity identifiers (i.e., the *relational covariates*) to perform propensity score matching on relational data. The procedure for RPSM is described in Algorithm 1. RPSM takes as input a data-set $\mathcal{X}$, a relational schema, the treatment and outcome attributes, a set of possible confounding attributes, a set of aggregation functions, and a hop-threshold h. The algorithm constructs the set of relational covariates based on the confounding attributes, the aggregation functions, and hop-threshold (line 2, discussed below in detail). The propensity score of the *treatment* given the *covariates* is then computed (line 3) and matching is performed based on the propensity score (line 4).

The construction of relational covariates is presented in Algorithm 2. The algorithm first constructs all potential relational variables for the confounding attributes from the given perspective, up to the specified hop-threshold (line 1).[2] This is the set of *relational confounders*. Then, for each relational confounder, it creates the appropriate relational covariates by applying the given aggregation functions (lines 7-8). A degree variable is then added for the paths of the relational confounders (line 9). Finally, the algorithm identifies parts of the schema that form a hierarchy and adds identifier variables for the schema item lowest in the hierarchy to perform blocking (lines 10-14). Relational covariates that were constructed from relational variables that are now determined by the blocking path are removed from the list of covariates (line 13).

Algorithm 1: `RPSM` ($\mathcal{X}$, *schema*, *treatment*, *outcome*, *confoundingAttrs*, *aggrFunctions*, h)

```
1 perspective ← item class of treatment, outcome
2 covariates ← GetRelationalCovariates (schema,
  perspective, confoundingAttrs, aggrFunctions, h)
3 propensityScore ← Calculate propensity score for
  treatment ~ covariates using X
4 matches ← Match (propensityScore, treatment, X)
5 return matches
```

Algorithm 2: `GetRelationalCovariates` (*schema*, *perspective*, *confoundingAttrs*, *aggrFunctions*, h)

```
1  relationalConfounders ← relational variables with
   attributes in confoundingAttrs from perspective
   perspective up to hop-threshold h
2  relCovariates ← ∅
3  for P.X in relationalConfounders do
4      if P == [perspective] then
5          relCovariates ← relCovariates ∪ P.X
6      else
7          for agg in aggrFunctions do
8              relCovariates ← relCovariates∪agg(P.X)
9          relCovariates ← relCovariates ∪ degree(P)
10 for P.X in relationalConfounders do
11     if P is valid blocking choice for perspective then
12         controlled ← relational variables that P
           controls for
13         relCovariates ← relCovariates \ controlled
14         relCovariates ← relCovariates ∪ P.id
15 return relCovariates
```

Example 3.1. Consider our earlier scenario of assessing the effect of smoking on a user's weight. The treatment is *User.Smokes* and the outcome is *User.Weight* (the perspective of the analysis is the *User* entity class). If *Drinks* is given as a possible confounding attribute and the hop-threshold is 4, the algorithm will add the following relational variables to the set of relational confounders:

$$[User].Drinks$$
$$[User, Friends, User].Drinks$$
$$[User, Friends, User, Friends, User].Drinks$$
$$[User, ChecksIn, Place, ChecksIn, User].Drinks$$
$$[User, From, Hometown, From, User].Drinks$$

The next step is to create relational covariates based on the above relational variables. First, relational variables that only involve the *User* entity, in this case $[User].Drinks$, are added to the set of relational co-

[2] The algorithm can be trivially extended to exclude certain relational paths. For example, if the user has domain knowledge that would exclude specific relational paths or relational variables from the list of potential confounders.

variates. Because these covariates are propositional, aggregation functions are not applied.

The aggregation functions are then applied to relational variables that cross the boundaries of the *User* entity. If the set of aggregation functions is {`mean`}, the algorithm will add the following to the set of relational covariates:

$\texttt{mean}\big([User, Friends, User].Drinks\big)$,
$\texttt{mean}\big([User, Friends, User, Friends, User].Drinks\big)$,
$\texttt{mean}\big([User, ChecksIn, Place, ChecksIn, User].Drinks\big)$,
$\texttt{mean}\big([User, From, Hometown, From, User].Drinks\big)$

The set of relational covariates is augmented by including the degree of the relational paths that involve more than one entity classes:

$\texttt{degree}\big([User, Friends, User]\big)$,
$\texttt{degree}\big([User, Friends, User, Friends, User]\big)$,
$\texttt{degree}\big([User, ChecksIn, Place, ChecksIn, User]\big)$,
$\texttt{degree}\big([User, From, Hometown, From, User]\big)$

Finally, *id* variables are added to the relational paths. In this case, there exists a hierarchy expressed by the relational path $[User, From, Hometown]$. Therefore, the algorithm adds the following relational covariate:

$$[User, From, Hometown].id$$

In practice, the hop-threshold should be chosen on a case by case basis, using expert knowledge of the application domain. The choice of aggregation functions can be guided by an analysis of each variable's marginal distribution from the perspective of the treatment and outcome.

4 SYNTHETIC EXPERIMENTS

To evaluate the performance of RPSM we examine the following hypotheses:

1. Propensity score matching models that are limited to simplistic relational attributes ($h = 2$) fail to fully account for confounding network effects ($h = 4$) (Section 4.1).
2. Traditional aggregates for relational data, such as `mean`, when used in isolation do not sufficiently condition on the *distribution* of confounding relational variables (Section 4.2).
3. The inclusion of identifiers for entities that lie along valid blocking paths accounts for latent confounders on those entities as well on entities connected to them. That is, including entity identifiers in the propensity model performs an implicit causal blocking design (Section 4.3).

For all experiments we used the structure derived from a sample of a real-world network, Foursquare [Gao et al., 2012], augmented with synthetic attributes on the entities. This allows for controlling the dependencies between attributes as well as the marginal and conditional distributions, while leveraging relationships from a real-network. The relational schema for the Foursquare network is shown in Figure 2. The relational skeleton consists of 9,599 users, 47,164 friendships, 182,968 locations where users "checked-in" via the mobile application, 1,360,123 check-ins, and the users' hometowns. Aggregate statistics for the network are shown in Table 1.

Table 1: Descriptive statistics for the Foursquare relational skeleton used in the synthetic experiments.

Aggregate	Friends	Check-Ins
mean	9.45	120.09
median	5	73
min	1	1
max	3674	2477

For our experiments we generated data from multiple models to test each hypothesis individually. In all experiments, the treatment is $[User].Smokes$ and the outcome $[User].Weight$. Each model was parameterized as follows: The value of the treatment was drawn from a logistic model parametrized using coefficients drawn from $\mathcal{U}(-2, 2)$ and interaction terms increasing in degree from 1 (no interaction) to 10 (up to 10 interacting covariates, not necessarily distinct, per term). We refer to this varying degree as "covariate complexity". The value of outcome was drawn from a linear model with coefficients drawn from $\mathcal{U}(-2, 2)$ and an error distribution drawn from $\mathcal{N}(0, 1)$. Marginal distributions for each variable were drawn from $\mathcal{N}(\mu, \sigma)$, with μ and σ sampled for each variable individually from $\mathcal{U}(0, 5)$ and $\mathcal{U}(1, 3)$, respectively.

We used logistic regression to calculate the propensity score and then performed full matching using the `optmatch` package [Hansen and Klopfer, 2006]. A linear model was applied using treatment and matching assignment as covariates and outcome as the response variable to assess statistical significance, with an α value of 0.01 for determining dependence. In this setting, we would expect a low error rate for linear log-odds functions (covariate complexity is 1), given the perfect correspondence between the generating models and the estimation methods when the set of covariates is correctly specified (no interaction terms). Adding interaction terms renders the models progressively less appropriate. We report Type I and Type II errors. Type I error corresponds to cases where a valid causal dependence exists between treatment and outcome and RPSM incorrectly concludes that there exists no such dependence. Type II error corresponds

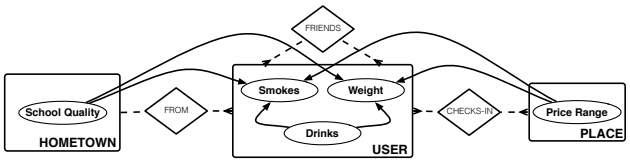


Figure 5: Class-dependency graph for the models used to evaluate the effect of using relational variables with longer hop-thresholds as covariates.

to cases where RPSM incorrectly concludes that there exists a dependency between treatment and outcome.

4.1 SIMPLE NETWORK DEPENDENCIES

We examine the first hypothesis, whether propensity score models limited to simplistic relational attributes fail to fully account for confounding network effects, by generating data from two models. Both models have the same class-dependency graph, shown in Figure 5, but differ in the length of the longest true dependency.

In the first model (World2), the true relational confounders are at most two hops away from the treatment and outcome entity. This corresponds to dependencies that can be read directly from the class dependency graph, e.g., the places a user checks in to. The set of true relational confounders for the model is:

$$[User].Drinks$$
$$[User, From, Hometown].SchoolQuality$$
$$[User, ChecksIn, Place].PriceRange$$

In the second model (World4), the set of true confounders is extended to include relational variables up to four hops away, e.g., other users that check in to the same places as a user. The set of confounders includes all of the confounders of the first model as well as:

$$[User, Friends, User].Drinks$$
$$[User, ChecksIn, Place, ChecksIn, User].Drinks$$
$$[User, From, Hometown, From, User].Drinks$$
$$[User, Friends, User, Friends, User].Drinks$$
$$[User, Friends, User, ChecksIn, Place].PriceRange$$
$$[User, Friends, User, From, Hometown].SchoolQuality$$

Using the above procedure we ran 100 trials. For each trial we considered two cases, one in which treatment and outcome are conditionally independent and one in which there is a direct effect between them. We then compared two methods for creating the relational covariates for propensity score matching:

1. RPSM using `mean`, `stdev`, `max`, `min` as aggregation functions and $h = 2$ without blocking or degree variables (RPSM2)

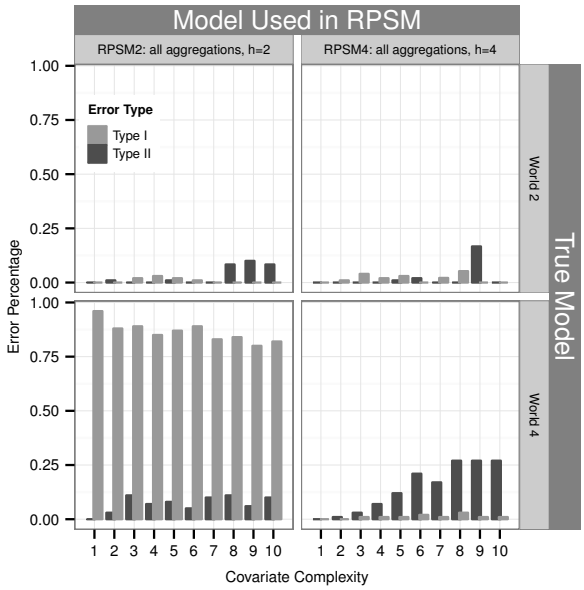


Figure 6: Percentage of Type I and II error when RPSM2 and RPSM4 are applied to data generated by World2 and World4 models with increasing covariate complexity, averaged over 100 trials.

2. RPSM using `mean`, `stdev`, `max`, `min` as aggregation functions and $h = 4$ without blocking or degree variables (RPSM4)

The results are shown in Figure 6. Along the diagonal the RPSM model is consistent with the world configuration. When models are over-specified, for instance RPSM4 in the World2 configuration, RPSM4 achieves comparable performance to RPSM2. However, when models are underspecified, for instance RPSM2 in the World 4 configuration, a spurious effect is inferred between treatment and outcome in the conditionally independent case. These results also demonstrate a case in which RPSM can successfully tolerate large numbers of irrelevant covariates.

4.2 COMPLEX NETWORK DEPENDENCIES

In this section, we examine the second hypothesis regarding the effect of using complex aggregation function in the construction of relational covariates. We generated data from models with the same class-dependency graph as in Section 4.1. We used World2 and World4, as before, and two simplified models which consider *only* `mean` as an aggregate, with hop-thresholds of 2 (World2-) and 4 (World4-). We then used the RPSM2 and RPSM4 methods for constructing relational covariates and two simpler propensity score models that only include `mean` as an aggregate

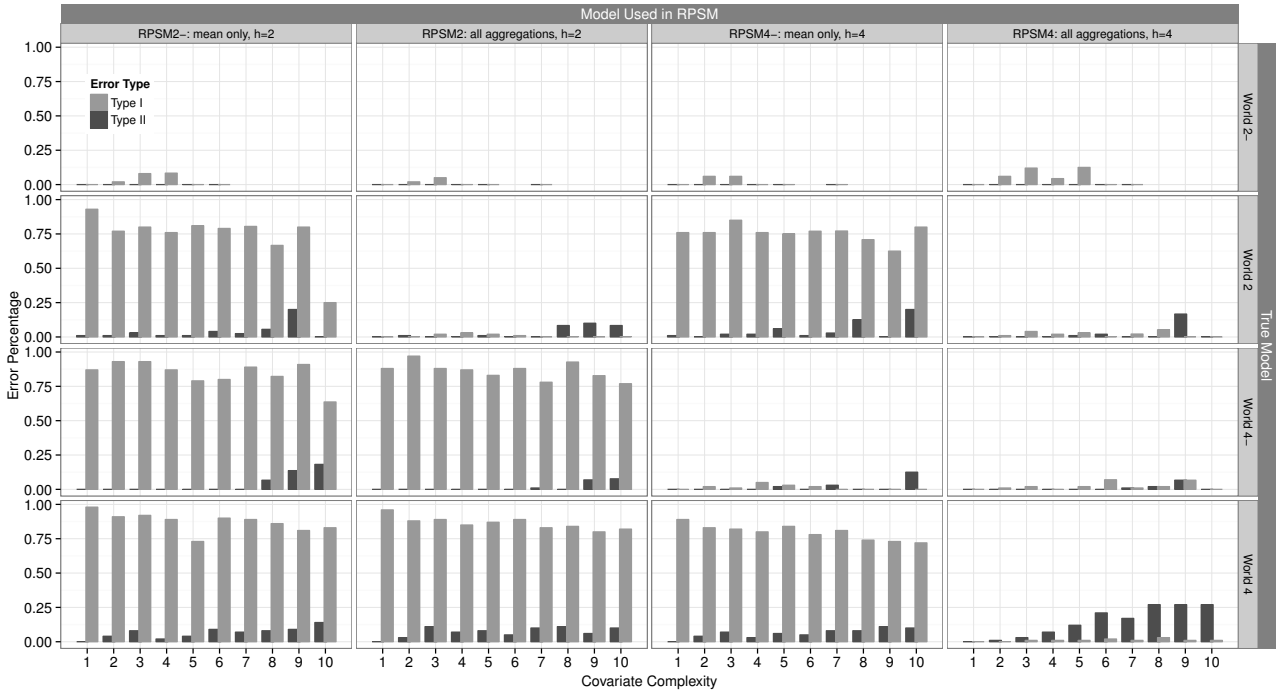


Figure 7: Type I and Type II error when RPSM2-, RPSM2, RPSM4-, and RPSM4 are applied to data generated by World2-, World2, World4-, and World4 models with increasing covariate complexity, averaged over 100 trials.

(RPSM2- with $h=2$ and RPSM4- with $h=4$).

The results are shown in Figure 7. Along the main diagonal, the assumptions of the RPSM model are consistent with the true world configuration. In cases where the employed model uses `mean` as the sole aggregation function but distributional dependencies are more complex, RPSM commits a large number of Type I errors. However, the over-specified models (e.g., RPSM4 in World2) maintain accuracy levels that are consistent with the most efficient RPSM configuration.

4.3 ENTITY IDENTIFIERS

The final experiment examines the third hypothesis regarding the effect of including entity identifiers in the relational propensity score model. We generated data from a model similar to that of Figure 5, with an additional latent confounder on the *Hometown* entity. We then created relational covariates using four strategies:

1. Use all observed variables and hop-threshold of 2 (RPSM2) and 4 (RPSM4).
2. Use degree variables and entity-identifiers for all eligible blocking paths with either $h = 2$ or $h = 4$ (RPSM2+ and RPSM4+ respectively).

The results are shown in Figure 8. RPSM2 and RPSM4 perform poorly, because of the bias induced by unconditioned confounders. RPSM2+ performs well when true relational dependencies are limited to $h = 2$. RPSM4+ performs well in all cases. This is an indication that including the entity identifiers in the propensity model performs blocking, producing effects similar to the explicit conditioning performed by Rattigan et al. [2011]. This also strengthens the connection between relational blocking and a conjecture made by Perlich and Provost [2006] that the inclusion of identifier variables in a non-causal setting can be used to create a relational fixed or random effects model. Given these results, the ability to automatically identify and utilize entity identifiers provides a strong argument for using RPSM as opposed to a propositional approach. While blocking accounts for a relatively small subset of all possible confounders, it provides a substantial improvement over the alternative of assuming no latent confounders.

5 RELATED WORK

Multi-level propensity score models [Hong and Raudenbush, 2006, Li et al., 2013] provide a method for accounting for group or cluster level effects. This corresponds to a one-to-many relationship in a relational schema. RPSM can be seen as an extension of the multi-level setting, capturing not only one-to-many group level effects, but also many-to-many effects. There has also been significant progress in un-

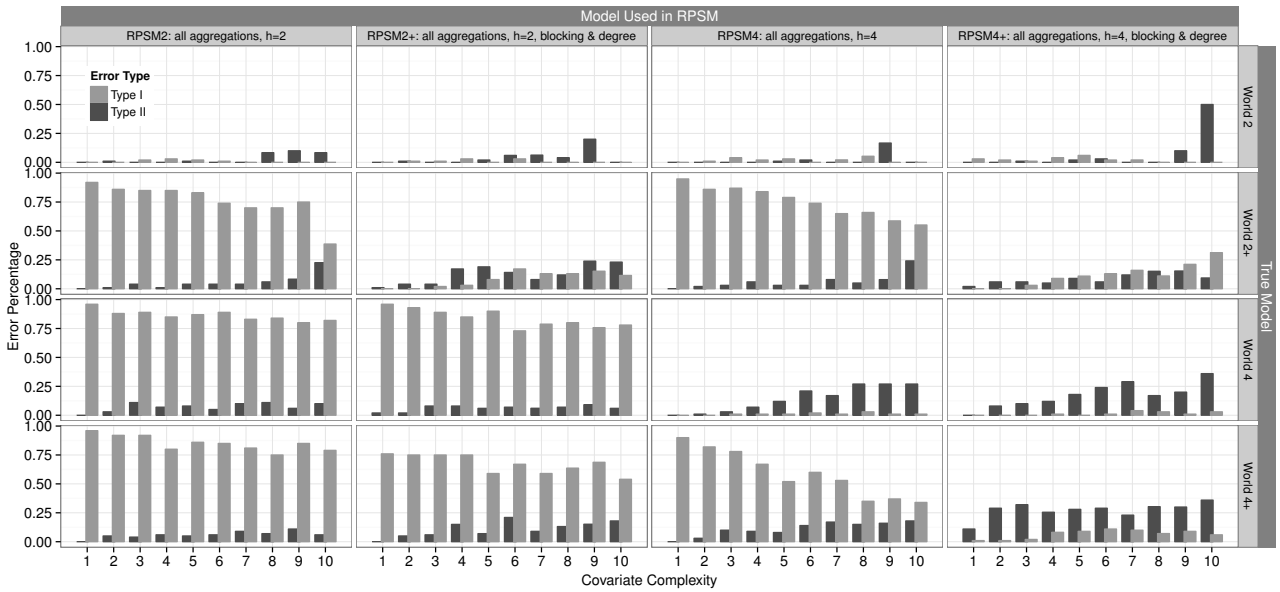


Figure 8: Type I and Type II error when RPSM2, RPSM2+, RPSM4, and RPSM4+ are applied to data generated by World2, World2+, World4, and World4+ with increasing covariate complexity, averaged over 100 trials.

derstanding the consequences of the stable unit treatment value assumption (SUTVA)[3] for matching and propensity models in the fields of statistics, epidemiology and econometrics [Hudgens and Halloran, 2008, Tchetgen and VanderWeele, 2012, Manski, 2013]. This work does not address SUTVA violations, but extensions to that setting are a focus of future work.

Perlich and Provost [2006] introduced relational fixed and random effects models using identifier attributes as features in the ACORA framework. RPSM differs in two important aspects. First, the aim of the aforementioned work is predictive, rather than causal. Second, RPSM incorporates degree variables and provides an algorithm for deciding *which* relational variables should be included, rather than assuming the correct set of relational variables and aggregating.

In the area of relational causal discovery, Maier et al. [2013a] introduced a constraint-based algorithm, RCD, that leverages relational *d*-separation [Maier et al., 2013b] to learn causal models from relational data. RCD learns a joint causal model of a relational domain and abstracts away the mechanics of performing individual tests of conditional independence, while RPSM focuses on evaluating a single causal dependence and the conditioning mechanism.

[3]SUTVA states that the outcome of an individual is independent of the treatment status of other individuals.

6 FUTURE WORK

We plan on examining RPSM further, using more complex synthetic data and real-world data. An interesting avenue for future research is extending RPSM to the case where the treatment or outcome lies along a one-to-many relational path (e.g., the effect of a treatment performed on an individual on an aggregate attribute of the individual's friends). There are also a number of methods for performing matching without a propensity score, such as matching on the full set of covariates [Stuart, 2010], coarsened exact matching [Iacus et al., 2012], and entropy balancing [Hainmueller, 2012]. Extending these methods to the relational setting would allow practitioners flexibility in terms of the set of assumptions required for a given causal analysis.

7 CONCLUSIONS

Propensity score matching provides a powerful and robust method for causal inference on propositional data. However, naively applying PSM to relational data ignores both new challenges and opportunities presented by this richer type of data. RPSM automatically constructs the set of relational covariates to be used in the propensity score model given a set of confounding attributes, a set of aggregation functions, and a hop threshold. Further, it exploits the relational structure by identifying degree variables and entity identifiers, which can account for latent relational confounders. We evaluate its efficacy via synthetic experiments that leverage a real-world relational skeleton.

References

C. Drake. Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics*, 49(4):1231–1236, 1993.

R. A. Fisher. *The Design of Experiments*. Oliver and Boyd, Edinburgh, 1935.

H. Gao, J. Tang, and H. Liu. gSCorr: Modeling geo-social correlations for new check-ins on location-based social networks. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pages 1582–1586. ACM, 2012.

L. Getoor and B. Taskar. *Introduction to statistical relational learning*. MIT press, 2007.

J. Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46, 2012.

B. B. Hansen and S. O. Klopfer. Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics*, 15(3):609–627, 2006.

D. E. Ho, K. Imai, G. King, and E. A. Stuart. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15(3):199–236, 2007.

G. Hong and S. W. Raudenbush. Evaluating kindergarten retention policy. *Journal of the American Statistical Association*, 101(475), 2006.

M. G. Hudgens and M. E. Halloran. Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482), 2008.

S. M. Iacus, G. King, and G. Porro. Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, 20(1):1–24, 2012.

D. D. Jensen, J. Neville, and M. Hay. Avoiding bias when aggregating relational data with degree disparity. In *Proceedings of the Twentieth International Conference on Machine Learning*, pages 274–281. AAAI Press, 2003.

B. K. Lee, J. Lessler, and E. A. Stuart. Improving propensity score weighting using machine learning. *Statistics in Medicine*, 29(3):337–346, 2010.

F. Li, A. M. Zaslavsky, and M. B. Landrum. Propensity score weighting with multilevel data. *Statistics in Medicine*, 32(19):3373–3387, 2013.

M. Maier, K. Marazopoulou, D. Arbour, and D. Jensen. A sound and complete algorithm for learning causal models from relational data. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pages 371–380, 2013a.

M. Maier, K. Marazopoulou, and D. Jensen. Reasoning about independence in probabilistic models of relational data. *arXiv preprint arXiv:1302.4381*, 2013b.

C. F. Manski. Identification of treatment response with social interactions. *The Econometrics Journal*, 16(1):S1–S23, 2013.

D. F. McCaffrey, G. Ridgeway, and A. R. Morral. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, 9(4):403, 2004.

J. Pearl. Remarks on the method of propensity score. *Statistics in Medicine*, 28(9):1415–1416, 2009.

C. Perlich and F. Provost. Distribution-based aggregation for relational learning with identifier attributes. *Machine Learning*, 62(1-2):65–105, February 2006.

M. J. Rattigan, M. Maier, and D. Jensen. Relational blocking for causal discovery. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, pages 145–151, 2011.

P. R. Rosenbaum. A characterization of optimal designs for observational studies. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(3): 597–610, 1991.

P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

E. A. Stuart. Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25 (1):1–21, 2010.

E. J. T. Tchetgen and T. J. VanderWeele. On causal inference in the presence of interference. *Statistical Methods in Medical Research*, 21(1):55–75, 2012.

D. Westreich, J. Lessler, and M. J. Funk. Propensity score estimation: Neural networks, support vector machines, decision trees (cart), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology*, 63(8):826–833, 2010.

Type-II Errors of Independence Tests Can Lead to Arbitrarily Large Errors in Estimated Causal Effects: An Illustrative Example

Nicholas Cornia & Joris M. Mooij
Informatics Institute
University of Amsterdam, The Netherlands
{n.cornia,j.m.mooij}@uva.nl

Abstract

Estimating the strength of causal effects from observational data is a common problem in scientific research. A popular approach is based on exploiting observed conditional independences between variables. It is well-known that this approach relies on the assumption of faithfulness. In our opinion, a more important practical limitation of this approach is that it relies on the ability to distinguish independences from (arbitrarily weak) dependences. We present a simple analysis, based on purely algebraic and geometrical arguments, of how the estimation of the causal effect strength, based on conditional independence tests and background knowledge, can have an arbitrarily large error due to the uncontrollable type II error of a single conditional independence test. The scenario we are studying here is related to the LCD algorithm by Cooper [1] and to the instrumental variable setting that is popular in epidemiology and econometry. It is one of the simplest settings in which causal discovery and prediction methods based on conditional independences arrive at non-trivial conclusions, yet for which the lack of uniform consistency can result in arbitrarily large prediction errors.

Introduction

Inferring causation from observational data is a common problem in several fields, such as biology and economics. To deal with the presence of unmeasured confounders of observed random variables the so-called *instrumental variable* technique [2] has found applications in genetics [3], epidemiology [4, 5] and economics [6]. Given two observable random variables possibly influenced by a hidden confounder, an *instrumental variable* is a third observed variable which is assumed to be independent of the confounder. In practice it is difficult to decide whether the instrumental variable definition is satisfied, and the method has aroused some skepticism [7]. In this paper, we study a setting that is similar in spirit to the instrumental variable model, but where all conditional independence assumptions are directly testable on the observed data. A similar scenario was first studied by Cooper [1] and independently rediscovered in the context of genome biology by Chen *et al.* [8].

An important assumption in causal discovery methods based on conditional independences is *faithfulness*, which means that the observed joint distribution does not contain any additional (conditional) independences beyond those induced by the causal structure. Usually, faithfulness is justified by the assumption that unfaithful distributions are a set of Lebesgue measure zero in the set of the model parameters. By showing that one can create a sequence of faithful distributions which converges to an unfaithful one, Robins *et al.* proved the lack of uniform consistency of causal discovery algorithms [9]. Zhang and Spirtes [10] then introduced the "Strong Faithfulness" assumption to recover the uniform consistency of causal discovery. Using geometric and combinatorial arguments, Uhler *et al.* [11] addressed the question of how restrictive the Strong Faithfulness assumption is in terms of the volume of distributions that do not satisfy this assumption. Even for a modest number of nodes and for sparse graphs, the "not strongly faithful" regions can be surprisingly large, and Uhler *et al.* argue that this result should discourage the use of large scale causal algorithms based on conditional independence tests, such as the PC and FCI algorithms [12].

In this work, we analyse in the context of the LCD setting how an error in a *single* conditional independence test may already lead to arbitrarily large errors in predicted causal effect strengths, even when

the faithfulness assumption is *not* violated. Our results may not be surprising for those familiar with the work of [9], but we believe that the analysis we present here may be easier to understand to those without a background in statistics, as we separate statistical issues (the possibility of type II errors in the conditional independence test from a finite sample) from a rather straightforward analysis of the problem in the population setting. We use an algebraic approach, showing how causal prediction may lead to wrong predictions already in the simple context of linear structural equation models with a multivariate Gaussian distribution.

In Section 1, we begin with a brief description of the problem setting in a formal way, giving the definitions of the causal effect, instrumental variable, LCD algorithm and the toy model we present. We consider three observed random variables (X_1, X_2, X_3), which is the minimal number such that a non-trivial conditional independence test can be obtained. In Section 2, we show how an (arbitrarily weak) conditional dependence that goes undetected can influence our estimation of the causal effect of X_2 on X_3 from the observed covariance matrix, when a confounder between X_2 and X_3 is almost off-set by a direct effect from X_1 to X_3. In fact, we show that this phenomenon can lead to an arbitrarily large error in the estimated causal effect as the noise variance of X_2 approaches zero. We finish with conclusions in Section 3.

1 Problem setting

1.1 LCD algorithm

The model we are interested in arises from the work of Cooper [1], who proposed the "LCD" algorithm for causal discovery in observational databases and the more recent paper of Chen et al.[8], who proposed the "Trigger" algorithm to infer transcriptional regulatory networks among genes. Throughout this section we will assume:

- Acyclicity;
- No Selection Bias.

Definition 1.1. (LCD setting) Given three random variables X_1, X_2, X_3 such that the following statistical properties and prior assumptions are satisfied:
Statistical dependences:

- $X_1 \not\perp\!\!\!\perp X_2$
- $X_2 \not\perp\!\!\!\perp X_3$
- $X_1 \perp\!\!\!\perp X_3 | X_2$

Prior assumptions:

- $An(X_1) \cap \{X_2, X_3\} = \emptyset$
- Faithfulness

where $An(X)$ is the set of the causal ancestors of X (which includes X itself), so this condition means that we assume that X_1 is not caused by the other observed variables X_2, X_3.

Cooper [1] proved that:

Theorem 1.1. *Under the assumptions in Definition 1.1, the causal structure must be a subgraph of:*

$$X_1 \overset{\hookrightarrow}{\rightarrow} X_2 \rightarrow X_3$$

Here, the directed arrows indicate a direct causal relationship and the bidirected edge denotes an unobserved confounder. □

Our primary interest is to predict $p(X_3|do(X_2))$, the distribution of X_3 after an intervention on X_2. In general, this quantity may differ from $p(X_3|X_2)$, the conditional distribution of X_3 given X_2 [13]. In the linear-Gaussian case, the quantity

$$\frac{\partial \mathbb{E}(X_3|do(X_2))}{\partial X_2}$$

measures the causal effect of X_2 on X_3.

It is easy to show that in the LCD setting, these quantities are equal:

Corollary 1.1. *Under the LCD assumptions in Definition 1.1,*

$$p(X_3|do(X_2)) = p(X_3|X_2).$$

Therefore, in the linear-Gaussian case, the quantity

$$\frac{\partial \mathbb{E}(X_3|do(X_2))}{\partial X_2} = \frac{\partial \mathbb{E}(X_3|X_2)}{\partial X_2} = \frac{\mathbb{C}\text{ov}(X_3, X_2)}{\mathbb{V}\text{ar}(X_2)} \quad (1)$$

is a valid estimator for the causal effect of X_2 on X_3.

1.2 Relationship with instrumental variables

The other model relevant for our discussion is the so called *instrumental variable* model. Following Pearl [13], we define:

Definition 1.2. (Instrumental Variable setting) Given three random variables X_1, X_2, X_3, we call X_1 an *instrumental variable* if the following conditions are satisfied:
Statistical dependences:

- $X_1 \not\perp\!\!\!\perp X_2$

Prior assumptions:

- $X_1 \perp\!\!\!\perp X_3 | do(X_2)$
- Faithfulness

The second assumption says that X_1 and X_3 are independent after an intervention on the variable X_2. In terms of the causal graph, this means that all the unblocked paths between X_1 and X_3 contain an arrow that points to X_2.

Unfortunately the instrumental variable property cannot be directly tested from observed data. The causal graph for the IV setting is a subgraph of:

$$X_1 \overset{\leftrightarrow}{\rightarrow} X_2 \overset{\leftrightarrow}{\rightarrow} X_3$$

So, a possible confounder between X_2 and X_3 is allowed, in contrast with the LCD setting. Note that the LCD setting is a special case of the IV model.

Lemma 1.1. *Under the IV assumptions in Definition 1.2 and for the linear-Gaussian case, the quantity*

$$\frac{\mathbb{C}\mathrm{ov}(X_1, X_3)}{\mathbb{C}\mathrm{ov}(X_1, X_2)}$$

is a valid estimator for the causal effect of X_2 on X_3.

1.3 Type II errors in LCD

In practice, the confidence on the result of the conditional independence test $X_1 \perp\!\!\!\perp X_3 | X_2$ in the LCD setting depends on the sample size. Indeed, it could be hard to distinguish a weak conditional dependence

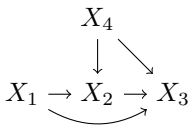
$$X_1 \not\perp\!\!\!\perp X_3 | X_2$$

from a conditional independence using a sample of finite size. Here we study the question of what happens to our prediction of the causal effect of X_2 on X_3 if the conditional independence test encounters a type II error (i.e., erroneously accepts the null hypothesis of independence).

Note that a type I error (i.e., erroneously rejecting the null hypothesis of independence) in the tests $X_1 \not\perp\!\!\!\perp X_2$ and $X_2 \not\perp\!\!\!\perp X_3$ will not be as dangerous as a type II error in the conditional independence test. Indeed, the probability of a type I error can be made arbitrarily small by tuning the significance level appropriately. In addition, a type I error would let the LCD algorithm reject a valid triple, i.e., lower the recall instead of leading to wrong predictions.

For these reasons we study the model described in the following definition, which allows the presence of a hidden confounder X_4, and a direct effect from X_1 on X_3 (not mediated via X_2). We assume that these additional features result in a possible weak conditional dependence between X_1 and X_3 given X_2. For simplicity we consider only the linear-Gaussian case. We also assume no confounders between X_1 and X_2, or between X_1 and X_3, or between X_1, X_2, X_3. This simplification will not influence the final result of the paper, because we will prove how unboundedness of the causal effect estimation error is already achieved for this special case.

Definition 1.3. We assume that the "true" causal model has the following causal graph:



which is one of the possible causal structures that is compatible with the following conditions:
Statistical dependences:

- $X_1 \not\perp\!\!\!\perp X_2$
- $X_2 \not\perp\!\!\!\perp X_3$
- **A weak conditional dependence**

$$X_1 \not\perp\!\!\!\perp X_3 | X_2$$

Prior assumptions:

- Faithfulness
- $An(X_1) \cap \{X_2, X_3\} = \emptyset$

The observed random variables are X_1, X_2, X_3 while X_4 is a hidden confounder, assumed to be independent from X_1.

The joint distribution of the observed variables is assumed to be a multivariate Gaussian distribution with covariance matrix Σ and zero mean vector. We also assume that the structural equations of the model are linear. Then

$$X = AX + E, \tag{2}$$

where

$$X = \big(X_1, \ldots, X_4\big)^T$$

is the vector of the extended system,

$$E = \big(E_1, \ldots, E_4\big)^T$$

is the vector of the independent noise terms, such that

$$E \sim \mathcal{N}\big(0, \Delta\big) \quad \Delta = diag\big(\delta_i^2\big),$$

and $A = (\alpha_{ij}) \in \mathcal{M}_4\big(\mathbb{R}\big)$ is (up to a permutation of indices) a real upper triangular matrix in the space $\mathcal{M}_4(\mathbb{R})$ of real 4×4 matrices that defines the causal strengths between the random variables of the system.

Remark 1.1. In [14], an implicit representation for the confounder X_4 is used, by using non-zero covariance between the noise variables E_2, E_3. It can be shown that for our purposes, the two representations are equivalent and yield the same conclusions.

In the Gaussian case, a conditional independence is equivalent to a vanishing partial correlation:

Lemma 1.2. *Given a set of three random variables (X_1, X_2, X_3) with a multivariate Gaussian distribution the conditional independence*

$$X_1 \perp\!\!\!\perp X_3 \mid X_2$$

is equivalent to a vanishing partial correlation

$$\rho_{13\cdot 2} = \frac{\rho_{13} - \rho_{12}\rho_{23}}{\sqrt{\big(1-\rho_{12}^2\big)\big(1-\rho_{23}^2\big)}} = 0 \tag{3}$$

where ρ_{ij} is the correlation coefficient of X_i and X_j.

In the model described in Definition 1.3,

$$\frac{\partial \mathbb{E}(X_3|do(X_2))}{\partial X_2} = \alpha_{23}. \tag{4}$$

In contrast with the LCD model in Definition 1.1, the equality (1) no longer holds. We are interested in the error in the estimation of the effect of X_2 on X_3 that would be due to a type II error of the conditional independence test in the LCD algorithm. The next section is dedicated to the analysis of the difference between the true value (4) and the estimated one in (1):

$$\big|\mathbb{E}\big(X_3|X_2\big) - \mathbb{E}\big(X_3|do(X_2)\big)\big| = \big|g\big(A,\Sigma\big)\big||X_2|,$$

where the "causal effect estimation error" is given by:

$$g\big(A,\Sigma\big) = \frac{\Sigma_{32}}{\Sigma_{22}} - \alpha_{23}. \tag{5}$$

2 Estimation of the causal effect error from the observed covariance matrix

The following proposition gives a set of equations for the observed covariance matrix Σ, given the model parameters $\big(A,\Delta\big)$ and the linear structural equation model (2).

Proposition 2.1. *The mapping $\Phi : (A,\Delta) \mapsto \Sigma$ that maps model parameters (A,Δ) to the observed covariance matrix Σ according to the model in Definition 1.3 is given by:*

$$\Sigma_{11} = \delta_1^2 \tag{6}$$

$$\Sigma_{12} = \alpha_{12}\delta_1^2 \tag{7}$$

$$\Sigma_{13} = \big(\alpha_{13} + \alpha_{23}\alpha_{12}\big)\delta_1^2 \tag{8}$$

$$\begin{aligned}\Sigma_{11}\Sigma_{23} = {}& \Sigma_{12}\Sigma_{13} \\ & + \Sigma_{11}\big(\delta_2^2\alpha_{23} + \delta_4^2\alpha_{42}\big(\alpha_{43} + \alpha_{23}\alpha_{42}\big)\big)\end{aligned} \tag{9}$$

$$\Sigma_{11}\Sigma_{22} = \Sigma_{12}^2 + \Sigma_{11}\big(\delta_2^2 + \delta_4^2\alpha_{42}^2\big) \tag{10}$$

$$\begin{aligned}\Sigma_{11}\Sigma_{33} = {}& \Sigma_{13}^2 \\ & + \Sigma_{11}\big(\delta_2^2\alpha_{23}^2 + \delta_3^2 + \delta_4^2\big(\alpha_{43} + \alpha_{23}\alpha_{42}\big)^2\big).\end{aligned} \tag{11}$$

Proof. It is possible to express the covariance matrix $\bar{\Sigma}$ of the joint distribution of $X_1, \ldots, X_4$ in terms of the model parameters as follows:

$$\bar{\Sigma} = \big(I - A\big)^{-T}\Delta\big(I - A\big)^{-1}.$$

The individual components in (6)–(11) can now be obtained by straightforward algebraic calculations. □

Remark 2.1. (Instrumental variable estimator) From equation (8) it follows immediately that for $\alpha_{13} = 0$, we have

$$\alpha_{23} = \frac{\Sigma_{13}}{\Sigma_{12}},$$

which corresponds to the usual causal effect estimator in the instrumental variable setting [3].

The lemma we present now reflects the fact that we are always free to choose the scale for the unobserved confounder X_4:

Lemma 2.1. *The equations of proposition 2.1 are invariant under the following transformation*

$$\bar{\alpha}_{4j} = \sqrt{\delta_4^2}\alpha_{4j}, \qquad \bar{\delta}_4^2 = 1$$

for $j \in \{2,3\}$.

Proof. This invariance follows from the fact that α_{42} and α_{43} always appear in a homogeneous polynomial of degree 2, and they are always coupled with a δ_4^2 term. □

Without loss of generality we can assume from now on that $\delta_4^2 = 1$.

Remark 2.2. (Geometrical Interpretation) From a geometrical point of view the joint system of equations for the observed covariance matrix defines a manifold $\mathcal{M}_\Sigma$ in the space of the model parameters $\mathcal{M}_4(\mathbb{R}) \times D_{\delta^2}$, where $\mathcal{M}_4(\mathbb{R})$ is the space of the possible causal strengths α_{ij} and

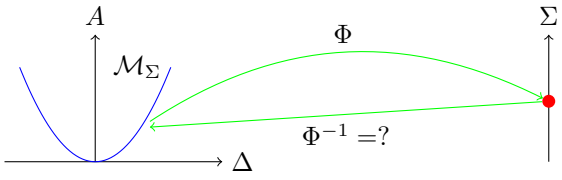
$$D_{\delta^2} = \prod_{i=1}^{3}[0, \Sigma_{ii}]$$

is the compact hypercube of the noise variances. Note that we have used the symmetry $\bar{\Sigma}_{44} = \delta_4^2 = 1$ and that

$$\delta_i^2 \leq \Sigma_{ii}$$

from equations (6), (10) and (11). Note that the map $\Phi : (A,\Delta) \mapsto \Sigma$ is not injective. This means that given an observed covariance matrix Σ, it is not possible to identify the model parameters in a unique way.

Indeed, the number of equations is six, while the number of model parameters is eight. Geometrically, this means that the manifold $\mathcal{M}_\Sigma$ does not reduce to a single point in the space of model parameters. Nevertheless it is still an interesting question whether the function g is a bounded function on $\mathcal{M}_\Sigma$ or not, i.e., whether we can give any guarantees on the estimated causal effect. Indeed, for the instrumental variable case with binary variables, such bounds can be derived (see, e.g., [13]).



The following Theorem and its Corollary are the main results of this paper. We will prove that there still remain degrees of freedom in the noise variances δ_2^2, δ_3^2 and the signs s_1, s_2, given the observed covariance matrix Σ, that will lead to an unbouded causal effect estimation error $g(A, \Sigma)$.

Theorem 2.1. *Given the causal model in Definition 1.3, there exists a map*

$$\Psi : \mathcal{M}_3(\mathbb{R}) \times D(\Sigma) \times \{-1, +1\}^2 \to \mathcal{M}_4(\mathbb{R}) \quad (12)$$

such that for all (A, Δ)*:*

$$\Psi(\Phi(A, \Delta), \delta_2^2, \delta_3^2, s_1, s_2) = A. \quad (13)$$

Here $D(\Sigma) = [0, m/\Sigma_{11}] \times [0, \det\Sigma/m] \subset \mathbb{R}^2$ *is the rectangle where the noise variances of* X_2 *and* X_3 *live, with* m *defined below in (19). The map* Ψ *gives explicit solutions for the causal strengths* α_{ij}*, given the observed covariance matrix* Σ*, the noise variances* δ_2^2, δ_3^2 *and signs* $s_i = \pm 1$*. The components of* Ψ *are given by:*

$$\alpha_{12} = \frac{\Sigma_{12}}{\Sigma_{11}} \quad (14)$$

$$\alpha_{42} = s_1\sqrt{\frac{m}{\Sigma_{11}} - \delta_2^2} \quad (15)$$

$$\alpha_{43} = s_2\frac{\sqrt{\det\Sigma - m\delta_3^2}}{\sqrt{\delta_2^2\Sigma_{11}}} \quad (16)$$

$$\alpha_{13} = s_1 s_2\frac{\Sigma_{12}\sqrt{\det\Sigma - m\delta_3^2}\sqrt{m - \Sigma_{11}\delta_2^2}}{m\sqrt{\delta_2^2\Sigma_{11}}} + \frac{\vartheta}{m}, \quad (17)$$

and the most important one for our purpose:

$$\alpha_{23} = \frac{\gamma}{m} - s_1 s_2\frac{\sqrt{\det\Sigma - m\delta_3^2}\sqrt{m - \Sigma_{11}\delta_2^2}}{m\sqrt{\delta_2^2}}. \quad (18)$$

Here,

$$\begin{aligned} m &= \Sigma_{11}\Sigma_{22} - \Sigma_{12}^2 > 0 \quad (19)\\ \eta &= \Sigma_{11}\Sigma_{33} - \Sigma_{13}^2 > 0\\ \omega &= \Sigma_{22}\Sigma_{33} - \Sigma_{23}^2 > 0\\ \vartheta &= \Sigma_{13}\Sigma_{22} - \Sigma_{12}\Sigma_{23}\\ \gamma &= \Sigma_{11}\Sigma_{23} - \Sigma_{12}\Sigma_{13}. \end{aligned}$$

Proof. The proof proceeds by explicitly solving the system of equations (6)–(11). Some useful identities are:

$$\begin{aligned} \alpha_{13} &= \frac{\Sigma_{12}\alpha_{42}\alpha_{43}}{m} + \frac{\vartheta}{m},\\ \alpha_{42}\alpha_{43} &= \frac{\gamma - \alpha_{23}m}{\Sigma_{11}},\\ \rho_{13\cdot 2} &= \frac{\vartheta}{\sqrt{\omega m}},\\ \eta m - \gamma^2 &= \Sigma_{11}\det\Sigma. \end{aligned}$$

The signs in the equations are a consequence of the second degree polynomial equations. □

Corollary 2.1. *It is possible to express the error in the estimated causal effect as*

$$g\big(\Psi(\Sigma, \delta_2^2, \delta_3^2, s_1, s_2), \Sigma\big) = \frac{\vartheta\Sigma_{12}}{m\Sigma_{22}} + s_1 s_2\frac{\sqrt{\det\Sigma - m\delta_3^2}\sqrt{m - \Sigma_{11}\delta_2^2}}{m\sqrt{\delta_2^2}}. \quad (20)$$

By optimizing over δ_3^2 *we get:*

$$\alpha_{23} \in [b_-, b_+] \subset \mathbb{R},$$

with

$$b_\pm(\delta_2^2) = \frac{\gamma}{m} \pm \frac{\sqrt{\det\Sigma}\sqrt{m - \Sigma_{11}\delta_2^2}}{m\sqrt{\delta_2^2}}. \quad (21)$$

The length of the interval $[b_-, b_+]$ *is a function of* (Σ, δ_2^2) *and satisfies*

$$\frac{\partial|b_+ - b_-|}{\partial\delta_2^2} < 0.$$

Proof. Equation (20) follows from (18) and:

$$\frac{\Sigma_{23}}{\Sigma_{22}} = \frac{\gamma}{m} + \frac{\vartheta\Sigma_{12}}{m\Sigma_{22}}.$$

From equation (11), combined with the results of Theorem 2.1, we can obtain the following inequality, using also the fact that $\delta_3^2\Sigma_{11} > 0$:

$$m\alpha_{23}^2 - 2\gamma\alpha_{23} + \eta - \Sigma_{11}\alpha_{43}^2 \geq 0.$$

The two solutions of the inequality define the interval $[b_-, b_+]$. Its length is a decreasing function of δ_2^2. □

Unfortunately, the causal effect strength α_{23} in equation (18) is unbounded. This means that for all the choices of the observed covariance matrix Σ that are in accordance with the model assumptions in Definition 1.3, the set of model parameters $(A, \Delta) \in \mathcal{M}_\Sigma$ that would explain Σ leads to an unbounded error g.

Indeed, a singularity is reached in the hyperplane $\delta_2^2 = 0$, which corresponds to making the random variable X_2 deterministic with respect to its parents X_1, X_4. Figure 1 shows the singularity of the function $|g(\Sigma, \delta_2^2, \delta_3^2)|$ in the limit $\delta_2^2 \to 0$. The rate of growth is proportional to the inverse of the standard deviation of the noise variable E_2:

$$|g| \propto \frac{1}{\delta_2} \quad \text{as } \delta_2 \to 0. \tag{22}$$

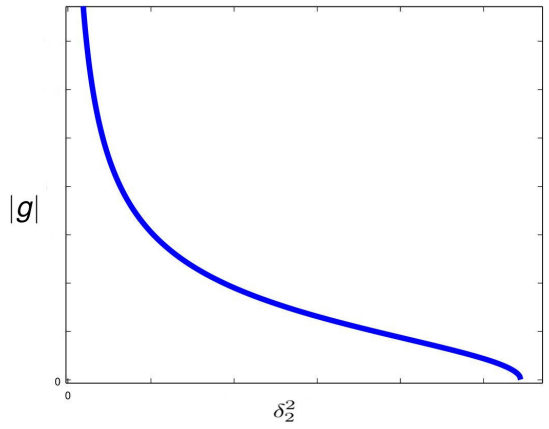


Figure 1: Causal effect estimation error $|g|$ as a function of δ_2^2, for fixed δ_3^2, Σ and $s_1 s_2 = 1$.

Remark 2.3. (Lower bound for δ_2^2) Corollary 2.1 is the main result of our analysis. The right hand term in (20) consists of two terms: the first one, through ϑ, represents the contribution of the partial correlation, and is small if $\rho_{13\cdot 2}$ is small. The second term is a fundamental, intrinsic quantity not controllable from the conditional independence test and the sample size. However, in situations where one is willing to assume a lower bound on δ_2^2:

$$\delta_2^2 \geq \hat{\delta}_2^2,$$

it is possible to give a confidence interval $[b_+, b_-]$ for the function g, depending on the choice of the lower bound $\hat{\delta}_2^2$.

Remark 2.4. (IV estimation error)
In the instrumental variable literature the IV estimator is used, presented in Lemma 1.1. Unfortunately, this estimator and its error function

$$h(\Sigma, A) = \frac{\Sigma_{13}}{\Sigma_{12}} - \alpha_{23} \tag{23}$$

is proportional to α_{13} and from (17) one can deduce a similar growing rate of the function h in terms of the variance of the noise term E_2:

$$|h| \propto \frac{1}{\delta_2} \quad \text{as } \delta_2 \to 0. \tag{24}$$

Remark 2.5. (Singularity analysis)
Figure 2 shows a contour plot of $|g|$ on the rectangle $D(\Sigma) \ni (\delta_2^2, \delta_3^2)$. The singularity in the causal effect

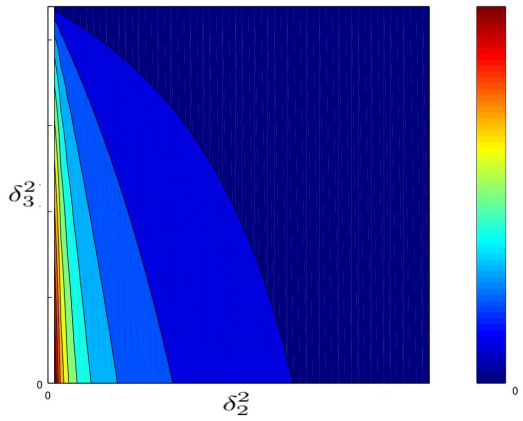


Figure 2: The function $|g|$ has a singularity in the hyperplane $\delta_2^2 = 0$.

function g is reached in the degenerate case, when the conditional distribution of X_2 given X_1 and X_4 approaches a Dirac delta function. This cannot be detected empirically, as we can still have well-defined covariance matrices Σ of the observed system even if the covariance matrix $\bar{\Sigma}$ of the extended one is degenerate.

Let us investigate in detail the limit for $\delta_2^2 \to 0$ from the point of view of the causal model. This proposition will show a simple example of how the causal strengths can be arbitrarily large, keeping the entries of the observed covariance matrix Σ_{ij} finite.

Proposition 2.2. *Assume that the observed covariance matrix Σ is positive-definite. Then, for the limit $\delta_2^2 \to 0$ we have the following scenario for the causal strength parameters:*

$$\begin{cases} \alpha_{23} \approx \pm\, \delta_2^{-1} \\ \alpha_{43} \approx \mp \operatorname{sgn}(\alpha_{42})\, \delta_2^{-1} \\ \alpha_{13} \approx \mp \operatorname{sgn}(\alpha_{12})\, \delta_2^{-1}. \end{cases}$$

This limit, in which our error in the estimated causal effect strength of X_2 on X_3 diverges, is illustrated in Figure 3.

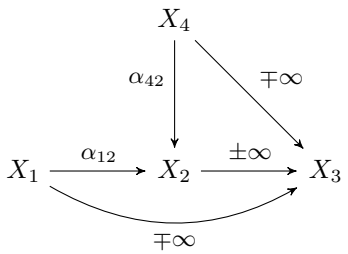


Figure 3: Scenarios in which the error in the causal effect strength of X_2 on X_3 based on the LCD algorithm may become infinitely large.

3 Conclusions and future work

Corollary 2.1 shows how the causal effect estimation error can be extremely sensitive to small perturbations of our model assumptions. Equation (20) holds for any value of ϑ (which is proportional to the partial correlation $\rho_{13.2}$) and the second term vanishes when the confounder is not present. This shows that with a finite sample, a type II error in the conditional independence test may lead to an arbitrarily large error in the estimated causal effect. Even in the infinite sample limit, this error could be arbitrarily large if faithfulness is violated. The result is in agreement with the results in [9], and it shows in a clear algebraic way how type II errors of conditional independence tests can lead to wrong conclusions.

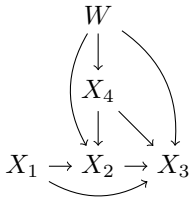
We believe that this conclusion holds more generally: even when we increase the complexity and the number of observed variables, the influence of confounders will still remain hidden, mixing their contribution with the visible parameters, thereby potentially leading to arbitrarily large errors. This means that for individual cases, we cannot give any guarantees on the error in the estimation without making further assumptions. An interesting question for future research is whether this negative worst-case analysis can be supplemented with more positive average-case analysis of the estimation error. Indeed, this is what one would hope if Occam's razor can be of any use for causal inference problems.

Other possible directions for future work are:

- **Study more complex models, in terms of the number of nodes, edges and cycles.**
- **Bayesian model selection:** We hope that the Bayesian approach will automatically prefer a simpler model that excludes a possible weak conditional dependence even though the partial correlation from the data is not exactly zero.
- **Bayesian Information Criterion:** We could directly assign a score based on the likelihood function of the data given the model parameters (A, Δ) and the model complexity, without assuming any prior distribution for the model parameters.
- **Nonlinear structural causal equations:** To deal with nonlinearity it is possible to consider Spearman's correlation instead of the usual one, using the following relationships:

$$
\begin{aligned}
m &= \Sigma_{11}\Sigma_{22}(1-\rho_{12}^2) \\
\eta &= \Sigma_{11}\Sigma_{33}(1-\rho_{13}^2) \\
\omega &= \Sigma_{22}\Sigma_{33}(1-\rho_{23}^2) \\
\gamma &= \Sigma_{11}\sqrt{\Sigma_{22}\Sigma_{33}}(\rho_{23}-\rho_{12}\rho_{13}) \\
\vartheta &= \Sigma_{22}\sqrt{\Sigma_{11}\Sigma_{33}}(\rho_{13}-\rho_{12}\rho_{23})
\end{aligned}
$$

- **"Environment" variable:** In many applications in biology, for example where X_1 is genotype, X_2 gene expression and X_3 phenotype, the observed random variables X_2 and X_3 are strongly dependent on the environmental conditions of the experiment. It might be reasonable to assume that most of the external variability is carried by the covariance between the environment variable W and the other measured ones, including possible confounders. This leads to the following graphical model, which could be useful in deriving some type of guarantees for this scenario:



Acknowledgements

We thank Tom Heskes for posing the problem, and Jonas Peters for inspiring discussions. We thank the reviewers for their comments that helped us improve the manuscript.

References

[1] G. F. Cooper. A simple constraint-based algorithm for efficiently mining observational databases for causal relationships. *Data Mining and Knowledge Discovery*, 1:203–224, 1997.

[2] R.J. Bowden and D.A. Turkington. *Instrumental Variables*. Cambridge University Press, 1984.

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

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

[5] D. A. Lawlor, R. M. Harbord, J. A. C. Sterne, N. Timpson, and G. D. Smith. Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*, 27:1133–1163, 2008.

[6] J.D. Angrista, W. G. Imbens, and D.B. Rubinc. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91:444–455, 1996.

[7] D. A. Jaeger J. Bound and R. M. Baker. Problems with instrumental variables estimation when the correlation between the instruments and the endogeneous explanatory variable is weak. *Journal of the American Statistical Association*, 90:443–450, 1995.

[8] L. S. Chen, F. Emmert-Streib, and J. D. Storey. Harnessing naturally randomized transcription to infer regulatory relationships among genes. *Genome Biology*, 8, 2007.

[9] R. Scheines J. M. Robins, P. Spirtes, and L. Wasserman. Uniform consistency in causal inference. *Biometrika*, 90:491–515, 2003.

[10] J. Zhang and P. Spirtes. Strong faithfulness and uniform consistency in causal inference. In *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence (UAI 2003)*, pages 632–639, 2003.

[11] C. Uhler, G. Raskutti, P. Bühlmann, and B. Yu. Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, 41:436–463, 2013.

[12] P. Spirtes, C. N. Glymour, and R. Scheines. *Causation, prediction, and search*. The MIT Press, 2000.

[13] J. Pearl. *Causality: models, reasoning and inference*. Cambridge University Press, 2000.

[14] M. Drton, R. Foygel, and S. Sullivan. Global identifiability of linear structural equation models. *The Annals of Statistics*, 39:865–886, 2011.

Toward Learning Graphical and Causal Process Models

Christopher Meek
Microsoft Research
One Microsoft Way
Redmond, WA 98052
meek@microsoft.com

Abstract

We describe an approach to learning causal models that leverages temporal information. We posit the existence of a graphical description of a causal process that generates observations through time. We explore assumptions connecting the graphical description with the statistical process and what one can infer about the causal structure of the process under these assumptions.

1 Introduction

Data that measure the temporal dynamics of systems is pervasive. The goal of this paper is to describe an approach to the development of a sound approach to causal inference for dynamic systems. One of the popular extant approaches is Granger causality (Granger 1969) which fails to be sound in the presence of latent variables. Granger causality is typically applied in discrete-time continuous valued time-series. Roughly speaking, in a multivariate time series X a set of variables are the Granger-causes of X_j if the historical values of this set of variables (including X_j) are necessary and sufficient for optimal prediction. Unfortunately a variable deemed a Granger-cause can arise due to either a latent common cause or as a result of a direct causal relationship and thus the approach cannot be used to determine causal relationships if one does not exclude the possibility of latent variables.

In this paper, we explore how one can leverage the assumption that causes must precede effects to inform causal conclusions drawn from observations of a temporal statistical process. The approach taken here is similar to the approach developed by Verma and Pearl (1990) and Spirtes, Glymour, and Scheines (2001) for atemporal causal discovery. One key ingredient in our approach is a new asymmetric graphical separation criterion for directed (possibly cyclic) graphs called δ*separation which plays an analogous role as d-separation in the work of Verma and Pearl (1990) and Spirtes, Glymour, and Scheines (2001). Another key ingredient is the process independence statement that plays an analogous role to the independence statement. Conceptually, we assume that we can test whether a process independence statements about observable quantities holds by observing the process and that these observation provide insight into the causal structure governing the process. In particular, we posit the existence of a graphical description of a causal process and make assumptions that connect δ*separation with observable process independence statements. We explore what can be inferred about the causal structure of the process under various observability assumptions. While the ultimate goal is to create a sound and complete method for causal inference for observations from a stochastic dynamic system, this paper represents some initial steps towards this ultimate goal. In particular, the results in Section 3.2 can be viewed sufficient conditions for Granger causality and, in Section 3.3, we present sufficient conditions under which we can make sound inferences about causal relationships including the existence of causal relationships and the existence and non-existence of latent common causal relationships.

As presented in Section 3, our causal discovery algorithm assumes the existence of an oracle for process independence statements. Our approach of abstracting away the details of how one connects process independence statements with particular statistical processes allows us to simultaneously make progress on the causal discovery problem for multiple distinct statistical processes such as marked point processes, Gaussian processes and dynamic Bayesian networks. In Section 4, we discuss two particular statistical processes and their associated process independence statements. In Section 5, we discuss some related work and open research questions.

2 Graphical Separation

We use $\mathcal{G} = \langle \mathcal{L}, \mathcal{E} \rangle$ to denote a directed graph where $\mathcal{L}$ is a set of vertices and $\mathcal{E} \subseteq \langle \mathcal{L} \times \mathcal{L} \rangle$ is a set of edges represented as ordered pairs. We write $a \rightarrow b$ if $\langle a, b \rangle \in \mathcal{E}$ and say that a is a parent of b and b is the child of a. Note that, in addition to allowing cycles, we also allow that a vertex can be its own parent and child (i.e., a self-edge $a \rightarrow a$). We use the shorthand $a \leftrightarrow b$ to indicate that $a \rightarrow b$ and $b \rightarrow a$.

A *path* in $\mathcal{G}$ is a sequence $\langle l_1, \ldots, l_n \rangle$ where there is an edge between successive pairs of vertices in $\mathcal{G}$. The length of a path $p = \langle l_1, \ldots, l_n \rangle$ is $|p| = n$ and a path p is termed a *trivial path* if $|p| = 1$. A vertex l_i on path $p = \langle l_1, \ldots, l_n \rangle$ is a *collider* on p if $l_{i-1} \rightarrow l_i$ and $l_i \leftarrow l_{i+1}$ and a *non-collider* otherwise. A *directed path* in graph $\mathcal{G} = \langle \mathcal{L}, \mathcal{E} \rangle$ is a sequence of vertices $\langle l_1, \ldots, l_n \rangle$ such that $\langle l_i, l_{i+1} \rangle \in \mathcal{E}$. The *source* of a directed path is the first vertex in the path. We denote the set of ancestors for a set A by $An(A)$. The ancestor relation is reflexive and thus $A \subseteq An(A)$.

We define a graph separation criterion called d*separation for directed graphs which is an extension of d-separation (Pearl 1988). An extension of d-separation is required as a pure vertex separation criterion like d-separation cannot separate a vertex from itself which is required to appropriately handle self-edges in directed graphs. A path p d*connects vertices a and b given the set of vertices C in graph $\mathcal{G}$ if every collider on p is in $An(C)$ and every non-collider on p is not in C. For sets of vertices $A, B, C \subseteq \mathcal{L}$ where $A \cap C = \emptyset$ we say that B is d*separated from A by C in graph $\mathcal{G}$ if and only if there does not exist a non-trivial d*connecting path between some $a \in A$ and some $b \in B$ given C in $\mathcal{G}$.

There are two key differences from Pearl's d-separation that allow us to appropriately handle cyclic directed graphs. First, we restrict d*separation statements to sets in which $A \cap C$ is the empty set but allow the sets A and B to overlap. Second, d*connecting paths must be non-trivial. These modifications enable us to use d*separation statements to distinguish between graphs in which there is a self-edge ($a \rightarrow a$) and one in which there is not.

We use directed graphs to represent temporal statistical processes. We associate the vertices $\mathcal{L}$ with a set of possible observation types (i.e., things that can happen). The edges denote potential dependencies between observations and the absence of a directed edge from observation type a to observation type b indicates that the process that generates observations of type b does not directly depend on the history of observations of type a. Analogous to the use of d-separation for directed acyclic graphs, we would like a graphical separation criterion for directed graphs to answer questions about how past observations influence future observations. Due in part to the fact that a directed graph does not explicitly encode temporal information we cannot simply apply d*separation on the directed graph. Instead, we define δ*separation which extends the graphical δ-separation of Didelez (2008) to handle self-edges. For sets $A, B, C \subset \mathcal{L}$ where $A \cap C = \emptyset$ we say that that B is δ*separated* from A given C (or simply $\delta(A, C, B)$) in $\mathcal{G}$ if an only if B is d-separated from A given C in the B-historical dependency graph $\mathcal{G}^B$ where $\mathcal{G}^B = \langle \mathcal{L}, \mathcal{E}^B \rangle$ and $\mathcal{E}^B = \mathcal{E} \backslash \{\langle b, a \rangle \in \mathcal{E} | b \in B, a \neq b\}$. Note that δ*separation is not symmetric in the first and third arguments due to the use of the graph $\mathcal{G}^B$.

3 Learning the Structure of a Causal Process

Our aim is to connect statistical processes with causal graphs and to learn the causal graph governing a system of observed events. We assume that there is a statistical process governing what and when events happen. We denote a statistical process for a set of observation types $\mathcal{L}$ by $\mathcal{P}_{\mathcal{L}}$. We also assume that we can observe the process to determine the whether process independence statements hold. We will write $PI(A, C, B)$ to indicate that the process associated with observations of type B does not depend on the history of observations of type A given the history of observations of type C in a given process $\mathcal{P}_{\mathcal{L}}$ (where $A \cap C = \emptyset$). We write $\neg PI(A, C, B)$ if this is not the case. We call such statements *process independence statements*. We note that process independence statements need not correspond to statistical independence statements and, as with δ*separation, there is no expectation that such process independence statements ought to be symmetric. In this section, we assume the existence of a process independence oracle for the relevant statistical process. In Section 4, we discuss particular statistical processes and the problem of testing process independence statements for those processes.

A process $\mathcal{P}_{\mathcal{L}}$ satisfies the *Causal Factorization Assumption* with respect to a causal process graph $\mathcal{G} = \langle \mathcal{L}, \mathcal{E} \rangle$ if and only if for all $A, B, C \subset \mathcal{L}$ where $A \cap B = \emptyset$ it is the case that $\delta(A, B, C) \Rightarrow PI(A, B, C)$

A process $\mathcal{P}_{\mathcal{L}}$ satisfies the *Causal Dependence Assumption* with respect to a causal process graph $\mathcal{G} = \langle \mathcal{L}, \mathcal{E} \rangle$ if and only if for all $A, B, C \subset \mathcal{L}$ where $A \cap B = \emptyset$ it is the case that $PI(A, B, C) \Rightarrow \delta(A, B, C)$

The Causal Analysis (CA) Algorithm (Algorithm 1) uses a process independence oracle to construct a di-

rected graph. We use $\pi_l^{\mathcal{G}}$ to denote the parents of l in graph $\mathcal{G}$ and $|B|$ to denote the cardinality of the set B. The basic idea is to use process independence statements to remove edges from an initially complete graph. This algorithm is analogous to the PC Algorithm of Spirtes, Glymour and Scheines (2001) but does not have an orientation phase.

Note that the output of the CA algorithm is a directed graph and that any edges presented do not necessarily indicate a causal relationship. In the remainder of this section we explore the interpretation of the output of the CA algorithm under various assumptions. Recall that $a \leftrightarrow b$ simply indicates that $a \rightarrow b$ and $b \rightarrow a$ and not the existence of a latent common cause.

Input: A set of events $\mathcal{L}$ and a process $\mathcal{P}_\mathcal{L}$
Output: A directed graph $\mathcal{G}$
Let $\mathcal{G} = \langle \mathcal{L}, \mathcal{E} \rangle$ be a complete directed graph.;
foreach $l \in \mathcal{L}$ **do**
  Let $n = 0$;
  foreach $l' \in \pi_l^{\mathcal{G}}$ **do**
    foreach $B \subseteq \pi_l^{\mathcal{G}} \setminus \{l'\}$ *where* $|B| = n$ **do**
      if $PI(l', B, l)$ *holds in* $\mathcal{P}_\mathcal{L}$ **then**
        $\mathcal{E} = \mathcal{E} \setminus \langle l', l \rangle$
      end
    end
    Let $n = n + 1$;
  end
end
Return $\mathcal{G} = \langle \mathcal{L}, \mathcal{E} \rangle$;
Algorithm 1: The Causal Analysis (CA) Algorithm

Theorem 1 (Complete Observations)**.** *If $\mathcal{P}_\mathcal{L}$ satisfies both the causal dependence and factorization assumptions with respect to $\mathcal{G}$ then algorithm $CA(\mathcal{L}, \mathcal{P}_\mathcal{L})$ returns $\mathcal{G}' = \mathcal{G}$.*

Lemma 1. *If $\mathcal{P}_\mathcal{L}$ satisfies the causal dependence assumption for $\mathcal{G} = \langle \mathcal{L}, \mathcal{E} \rangle$ and algorithm $CA(\mathcal{L}, \mathcal{P}_\mathcal{L})$ returns $\mathcal{G}' = \langle \mathcal{L}, \mathcal{E}' \rangle$ then if $l' \rightarrow l \in \mathcal{E}$ then $l' \rightarrow l \in \mathcal{E}'$.*

Lemma 2. *If $\mathcal{P}_\mathcal{L}$ satisfies both the causal dependence and factorization assumptions for $\mathcal{G} = \langle \mathcal{L}, \mathcal{E} \rangle$ and algorithm $CA(\mathcal{L}, \mathcal{P}_\mathcal{L})$ returns $\mathcal{G}' = \langle \mathcal{L}, \mathcal{E}' \rangle$ then if $l' \rightarrow l \notin \mathcal{E}$ then $l' \rightarrow l \notin \mathcal{E}'$.*

Proof of Theorem 1: The theorem follows from Lemmas 1 and 2.

3.1 Absence of a direct causal relationship

Next we consider the case in which some of the event types in the system are not observed. We let $\mathcal{O} \subseteq \mathcal{L}$ be the set of observed event types. In this case we will assume that the causal factorization and dependence assumptions hold for a process $\mathcal{P}_\mathcal{L}$ and some causal process graph $\mathcal{G}$. Our causal factorization and dependence assumptions allow us to focus on δ*separation in $\mathcal{G}$ by assuming that the observed process independence statements accurately reflect the δ*separation statements about $\mathcal{G}$ for the observed observation types. In order to understand and interpret the output of the CA algorithm we need to understand the conditions that lead to edges in the final output. We begin by defining the concept of vertex blockability relative to a set of observed event types.

We say that a vertex a is b-unblockable relative to $\mathcal{O}$ in $\mathcal{G}$ if and only if for all $C \subseteq \mathcal{O} \setminus \{a, b\}$ $\neg\delta(a, C, b)$ is true of $\mathcal{G}$. Otherwise the vertex is said to be b-blockable relative to $\mathcal{O}$. Note that if $b \rightarrow b$ then if $b \in \mathcal{O}$ b is b-unblockable relative to $\mathcal{O}$.

We say that l is a *direct cause* of l' relative to $\mathcal{O}$ for causal process graph $\mathcal{G}$ if and only if there exists a directed path $\langle l_1, \ldots, l_n \rangle$ where $l_1 = l$ and $l_n = l'$ and $l_i \notin \mathcal{O}$ for $(1 < i < n)$. We call the path in the definition of direct cause a *witnessing path* that l is a direct cause of l'. We let D_b denote the set of observed direct causes of the variable b relative to $\mathcal{O}$, that is, members of $\mathcal{O}$ that are direct causes of b relative to $\mathcal{O}$.

Example 1. Let $\mathcal{E} = \{a \rightarrow c, c \rightarrow b\}$, $\mathcal{L} = \{a, b, c\}$ and $\mathcal{O} = \{a, b\}$. The vertex a is b-unblockable relative to $\mathcal{O}$ for $\mathcal{G} = \langle \mathcal{L}, \mathcal{E} \rangle$ but the vertex b is a-blockable relative to $\mathcal{O}$. In this example, a is a direct cause of b relative to $\mathcal{O}$ in graph $\mathcal{G}$ and $a \rightarrow c \rightarrow b$ is a witnessing path for this fact.

Lemma 3. *If l' is a direct cause of l relative to $\mathcal{O}$ in $\mathcal{G}$ then l' is l-unblockable relative to $\mathcal{O}$ in $\mathcal{G}$.*

The following lemma allows us to make causal inferences using the causal analysis algorithm about the absence of a direct causal relationship.

Lemma 4. *If $\mathcal{P}_\mathcal{L}$ satisfies the causal dependence assumption with respect to $\mathcal{G}$ then, in the graph $\mathcal{G}'$ output by CA($\mathcal{O}$,$\mathcal{P}_\mathcal{L}$), the set of parents for each event type include all of its direct causes relative to $\mathcal{O}$.*

In particular, if the algorithm finds that an event type a is not a parent of event type b then a is not a direct cause of b.

3.2 Causal sufficiency

In the section, we restrict the type of unobserved event types which enables us to make strong inferences about the causal structure of a process. In particular we assume causal sufficiency which is essentially an assumption that there are no latent confounding processes.

A set of event types $\mathcal{O} \subset \mathcal{L}$ is *causally sufficient* with respect to a graph $\mathcal{G} = \langle \mathcal{L}, \mathcal{E} \rangle$ if and only if every common cause of $l, l' \in \mathcal{O}$ is in the set of event types

$\mathcal{O}$.

A directed graph $\mathcal{G}' = \langle \mathcal{O}, \mathcal{E}' \rangle$ is *causally correct* with respect to a graph $\mathcal{G} = \langle \mathcal{L}, \mathcal{E} \rangle$ if for every edge $\langle a, b \rangle \in \mathcal{E}'$ a is a direct cause of b with respect to $\mathcal{O}$ in $\mathcal{G}$.

Theorem 2 (Causal Sufficiency)**.** *If $\mathcal{P}_{\mathcal{L}}$ satisfies both the causal dependence and factorization assumptions for $\mathcal{G} = \langle \mathcal{L}, \mathcal{E} \rangle$ and $\mathcal{O} \subseteq \mathcal{L}$ is causally sufficient with respect to $\mathcal{G}$ then the graph $\mathcal{G}'$ returned by algorithm $CA(\mathcal{O}, \mathcal{P}_{\mathcal{L}})$ is causally correct with respect to $\mathcal{G}$ and $\mathcal{O}$.*

Lemma 5. *If $\mathcal{P}_{\mathcal{L}}$ satisfies the causal dependence and factorization assumptions with respect to $\mathcal{G}$ and $\mathcal{O}$ is causally sufficient for $\mathcal{G}$ then the output of the CA algorithm removes the edge $a \to b$ if a is not a direct cause of b relative to $\mathcal{O}$.*

3.3 Causal insufficiency

We have shown that the CA algorithm can provide causally accurate information under the assumptions of causal sufficiency, causal factorization and causal dependence. In this section we consider removing the assumption of causal sufficiency.

Example 2. Let $\mathcal{E} = \{a \leftarrow c, c \to b\}$, $\mathcal{L} = \{a, b, c\}$ and $\mathcal{O} = \{a, b\}$. The observed event types $\mathcal{O}$ are not causally sufficient for the graph $\mathcal{G} = \langle \mathcal{L}, \mathcal{E} \rangle$. In addition, the CA algorithm fails to provide output that is causally correct. In particular, the CA algorithm yields the graph in which $a \to b$ and $b \to a$ despite the fact that neither is a a cause of b in $\mathcal{G}$ nor is b a cause of a.

Our aim is to graphically characterize vertex separability. We do so using the idea of an inducing path in a directed graph that was introduced for directed acyclic graphs by Verma and Pearl (1990). For a pair of vertices a, b, we define $A_{ab} = An(\{a\}) \cup An(\{b\}) \setminus \{a, b\}$. A path p between $\langle a, b \rangle$ is an *inducing path* relative to $\mathcal{O}$ if and only if (1) every vertex on $p \in \mathcal{O}$ is a collider on p and (2) Every collider on p is in A_{ab}. An inducing path $p = \langle l_1 = a, \ldots, l_n = b \rangle$ from a to b is into b if $l_{n-1} \to l_n$. An inducing path $p = \langle l_1 = a, \ldots, l_n = b \rangle$ from a to b is out of a if $l_1 \to l_2$.

Lemma 6. *For a directed graph $\mathcal{G}$ the following three statements are equivalent:*

(a) *A vertex a is b-unblockable relative to $\mathcal{O}$ in graph $\mathcal{G}$*

(b) *There is an inducing path between a and b relative to $\mathcal{O}$ in graph $\mathcal{G}^b$. Note this inducing path must be into b.*

(c) *$\neg\delta(a, \mathcal{O} \cap A_{ab}, b)$ in $\mathcal{G}$.*

We say that a is a *cause* of b in $\mathcal{G}$ and if there is a directed path from a to b in $\mathcal{G}$.

We aim to find common features of all graphs that are consistent with the observed pattern of process independence statements. Latent processes, however, can mask the causal nature of the observed pattern of dependencies.

For a pair of vertices a, b and graph $\mathcal{G}$ we say that there is a *potential indirect inducing path* into b relative to $\mathcal{O}$ if and only if (1) there is a vertex $c_1 \in \mathcal{O} \setminus \{a, b\}$ such that $a \to b$ in $\mathcal{G}$ and (2) there is a sequence of vertices $c_1, \ldots, c_n \subseteq \mathcal{O} \setminus \{a, c\}$ such that $c_i \leftrightarrow c_{i+1}$ and $c_n \leftrightarrow b$ in $\mathcal{G}$.

Lemma 7. *For any set of observed variable $\mathcal{O}$, if a graph has an inducing path between observed variables a, b into b containing another observed variable then the output of the CA algorithm will contain a potential indirect inducing path into b.*

Theorem 3 (Sufficient Cause)**.** *If $\mathcal{P}_{\mathcal{L}}$ satisfies both the causal dependence and factorization assumptions for $\mathcal{G} = \langle \mathcal{L}, \mathcal{E} \rangle$ then if CA produces $\mathcal{G}'$ with vertices $\mathcal{O} \subseteq \mathcal{L}$ for which the subgraph over $\{a, b\}$ is $a \to b$ and $\mathcal{G}'$ contains no potential inducing path between a, b into b then a is a cause of b in $\mathcal{G}$.*

Lemma 8. *If $\mathcal{P}_{\mathcal{L}}$ satisfies both the causal dependence and factorization assumptions for $\mathcal{G} = \langle \mathcal{L}, \mathcal{E} \rangle$ and CA produces $\mathcal{G}'$ with vertices $\mathcal{O} \subseteq \mathcal{L}$ for which the subgraph over $\{a, b, c\}$ is $a \leftrightarrow b \leftrightarrow c$ then*

- *if $PI(a, \emptyset, c)$ and $PI(c, \emptyset, a)$ then there is a latent common causes of a, b and a (possibly distinct) latent cause of b, c and b is not a direct cause of c and b is not a direct cause of a.*
- *if $PI(a, b, c)$ then there is no latent common causes of b, c, b is a cause of c in $\mathcal{G}$.*

4 Statistical Processes and Process Independence

Our approach to causal discovery through the observation of a dynamic process is applicable to different temporal statistical processes. The key connection required is a connection between process independence statements and the observations from a particular statistical process. In this section we consider two distinct statistical processes and discuss process independence for these processes.

4.1 Dynamic Bayesian Networks

Dynamic Bayesian networks (DBNs) are a popular discrete-time model that can capture temporal dynamics of a statistical process. A DBN is a statis-

tical model of an infinite set of variables indexed by time. A variable X_i^t denotes the i^{th} variable at time t. We use $X = X_1, \ldots, X_n$ to denote the set of *variable types* in the DBN, that is, a variable with an unspecified time component and X^t to denote the set of variables at time t. The DBN specifies the evolution of X^t as a stochastic function of the value of previous variables X^{t-i} $(i > 0)$. In particular, the variable X_i^t is a stochastic function of the value of its parents in a graph. The causal process graph associated with a causal DBN is a graph over the variable types of the DBN X where there is an edge $X_i \rightarrow X_j$ if there exists a t, i such that there is an edge $X_i^{t-i} \rightarrow X_j$ in the DBN. Thus, the parent relationship of the causal process graph captures the dependence of a variable type on the history of other variable types. Furthermore, process independence statements $PI(X_i, C, X_j)$ correspond to a set of independence statements of the form $I(X_i^1, \ldots, X_i^{t-1}, X_C^1, \ldots, X_C^{t-1}, X_j^t)$. Without further assumptions, testing process independence would be unfeasible but if we focus on stationary processes with finite temporal dependency we can potentially test process independence statements.

4.2 Graphical Event Models

In this section, we define Conditional Intensity Models and Graphical Event Models (GEMs) and connect these models with previous work on the class of Piecewise-Constant Conditional Intensity Models and Poisson Networks. We assume that events of different types are distinguished by labels l drawn from a finite alphabet $\mathcal{L}$. An event is then composed of a non-negative time-stamp t and a label l. A *history* is an event sequence $h = \{(t_i, l_i)\}_{i=1}^n$ where $0 < t_1 < \cdots < t_n$, and our data is a specific history denoted by $\mathcal{D}$. Given data $\mathcal{D}$, we define the *history at time t* as $h(t, \mathcal{D}) = \{(t_i, l_i) \mid (t_i, l_i) \in \mathcal{D}, t_i \leq t\}$. We suppress $\mathcal{D}$ from $h(t, \mathcal{D})$ when clear from context and write $h_i = h(t_{i-1})$. By convention $t_0 = 0$. We define the *ending time* $t(h)$ of a history h as the time of the last event in h: $t(h) = \max_{(t,l) \in h} t$ so that $t(h_i) = t_{i-1}$.

A *Conditional Intensity Model* (CIM) is a set of non-negative *conditional intensity functions* indexed by label $\{\lambda_l(t|h;\theta)\}_{l \in \mathcal{L}}$. The data likelihood for this model is

$$p(\mathcal{D}|\theta) = \prod_{l \in \mathcal{L}} \prod_{i=1}^{n} \lambda_l(t_i|h_i, \theta)^{\mathbf{1}_l(l_i)} e^{-\Lambda_l(t_i|h_i;\theta)} \qquad (1)$$

where $\Lambda_l(t|h;\theta) = \int_{-\infty}^{t} \lambda_l(\tau|h;\theta)d\tau$ and the function $\mathbf{1}_l(l')$ is one if $l' = l$ and zero otherwise. The conditional intensities are assumed to satisfy $\lambda_l(t|h;\theta) = 0$ for $t \leq t(h)$ to ensure that $t_i > t_{i-1} = t(h_i)$. These modeling assumptions are quite weak. In fact, any distribution for $\mathcal{D}$ in which the timestamps are continuous random variables can be written in this form. For more details see [1, 2]. Despite the fact that the modeling assumptions are weak, these models offer a powerful approach for decomposing the dependencies of different event types on the past. In particular, this per label conditional specification allows one to model detailed label-specific dependence on past events.

Next we define a graphical conditional intensity model that we call a graphical event model (GEM). A filtered history for $A \subseteq \mathcal{L}$ as $[h]_A = \{(t_i, l_i) | (t_i, l_i) \in h \wedge l_i \in A\}$. A GEM is a pair $< \mathcal{G}, \theta >$, where $\mathcal{G} =< \mathcal{L}, \mathcal{E} >$ is a directed graph over a set of event types and edges in $\mathcal{E}$ represent potential dependencies among event types. The parameters $\theta = \{\theta_l\}_{l \in \mathcal{L}}$ parameterize the intensity functions for each event type. In particular, $\lambda_l(t|h_t, \theta_l) = \lambda_l(t|[h_t]_{\pi_l}, \theta_l)$ where π_l is the set of parents for l in $\mathcal{G}$. As in the case of the DBN, a process independence statement correspond to testing a dependence of an event type on set of event histories. One potential approach to testing a process independence $PI(a, C, b)$ is to estimate/learn an intensity function for b using the event histories for $\{a\} \cup C$ and see if the intensity model depends on the event history for a. The work by Gunawardana et al (2011) on learning piecewise continuous intensity models is a good starting point for this approach.

5 Discussion

One of the goals for the research direction described in this paper is the development a sound approach to causal inference for dynamic systems. One of the popular extant approaches is that of Granger causality which fails on this account. This approach is typically applied in a discrete-time continuous valued time-series and, thus, can be viewed as a dynamic Bayesian network. Roughly speaking, in a multivariate time series X a set of variables are the Granger-causes of X_j if the historical values of this set of variables (including X_j) are necessary and sufficient for optimal prediction. Unfortunately this approach does not appropriately handle latent common causes. In particular, for both of the scenarios described in Lemma 8 it is the case that each of the variables is a Granger cause of its neighbors while this relationships need not be causal as the lemma demonstrates. In fact, it is easy to construct stochastic processes with latent factors which demonstrate that the inferential approach to Granger causality is not sound with respect to causal relations.

There has been much work related to causal discovery and the estimation of causal effects in time-series. As discussed above, the work on Granger causality (Granger 1969) is the most well known. The short-

comings of this approach are also well known (e.g., Eichler 2007) and there has been some work in trying to address these known short comings. For instance, Eichler (2007) proposes a similar approach to the approach described here but differs in that it allows for the possibility of "simultaneous correlation" which requires the use of an alternative definition of separation. In addition, while providing definitions of cause and spurious cause, sufficient conditions for the identification of causal relationships are not presented. The work of Entner and Hoyer (2010) considers the problem of causal discovery from time series data using limited dependence vector autoregressive models and the FCI algorithm that uses conditional independence tests to identify the structure. Our approach of using δ*separation is inspired by the work of Didelez (2008) who defined δ-separation and shows the connection between that graphical separation criterion and local independence of marked point processes. Our extension to δ*separation allows for the appropriate treatment of self-edges which are essential in any self-excitatory or self-inhibitory dynamic process. Another more loosely connected work is that of Eichler and Didelez (2007) that considers the estimation of causal effects based on an intervention in a time-series.

While the results described in this paper offer hope for developing a methodologically sound approach to causal inference for dynamic systems, there is much work that needs to be done. Here are some of the open research questions.

- Non-parametric tests for process independence for various type of temporal statistical processes
- Soundness and completeness results for δ*separation analogous to those provided by Pearl (1988), Meek (1995) and Spirtes et al (2001) for d-separation. Note that Didelez (2008) has shown the soundness of δ-separation for a family of marked point processes related to GEMs.
- A representation for equivalence classes of causal graphs with respect to δ*separation in the case of causal insufficiency ($\mathcal{O} \subset \mathcal{L}$) analogous to those developed by Verma and Pearl (1990) and Spirtes et al (2001) that captures the common casual aspects of the set of graphs in the equivalence class.

Acknowledgments

Thanks to Asela Gunawardana and two anonymous reviewers for their comments on an earlier draft of this paper.

References

[1] D. J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes: Elementary Theory and Methods*, volume I. Springer, second edition, 2002.

[2] Vanessa Didelez. Graphical models for marked point processes based on local independence. *JRSS-B*, 70(1):245–264, 2008.

[3] Michael Eichler. Granger causality and path diagrams for multivariate time series. *Journal of Econometrics*, 137:334–353, 2007.

[4] Michael Eichler and Vanessa Didelez. Causal reasoning in graphical time series models. In *Uncertainty in Artificial Intelligence*, pages 109–116, 2007.

[5] Doris Entner and Patrik O. Hoyer. On causal discovery from time series data using FCI. In *Probabilistic Graphical Models*, pages 121–128, 2010.

[6] C.W.J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37:424–438, 1969.

[7] Asela Gunawardana, Christopher Meek, and Puyang Xu. A model for temporal dependencies in event streams. In *Advances in Neural Information Processing Systems*, 2011.

[8] C. Meek. Strong completeness and faithfulness in Bayesian networks. In *Proceedings of Eleventh Conference on Uncertainty in Artificial Intelligence,* Montreal, QU, pages 411–418. Morgan Kaufmann, August 1995.

[9] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search, Second Edition.* MIT Press, Cambridge, MA, second edition, 2001.

[10] T. Verma and J. Pearl. Equivalence and synthesis of causal models. In *Proceedings of Sixth Conference on Uncertainty in Artificial Intelligence,* Boston, MA, pages 220–227. Morgan Kaufmann, July 1990.

On Causal Explanations of Quantum Correlations

Robert W. Spekkens
Perimeter Institute for Theoretical Physics
31 Caroline St. N, Waterloo, Ontario, Canada, N2L 2Y5

Abstract

The framework of causal models is ideally suited to formalizing certain conceptual problems in quantum theory, and conversely, a variety of tools developed by physicists studying the foundations of quantum theory have applications for causal inference. This talk reviews some of the connections between the two fields. In particular, it is shown that certain correlations predicted by quantum theory and observed experimentally cannot be explained by any causal model while respecting the core principles of causal discovery algorithms. Nonetheless, it is argued that by understanding quantum theory as an innovation to the theory of Bayesian inference, one can introduce a quantum generalization of the notion of a causal model and salvage a causal explanation of these correlations without fine-tuning. Furthermore, experiments exhibiting certain quantum features, namely, coherence and entanglement, enable solutions to causal inference problems that are intractable classically. In particular, while passive observation of a pair of variables cannot determine the causal relation that holds between them according to classical physics, this is not the case in quantum physics. In other words, according to quantum theory, certain kinds of correlation *do* imply causation. The results of a quantum-optical experiment confirming these predictions will be presented.
This talk is based on the work described in Refs. [1] and [2].

References

[1] Christopher J. Wood and Robert W. Spekkens, *The lesson of causal discovery algorithms for quantum correlations: Causal explanations of Bell-inequality violations require fine-tuning*, preprint arXiv:1208.4119, (2012).

[2] Katja Ried, Megan Agnew, Lydia Vermeyden, Dominik Janzing, Robert W. Spekkens and Kevin J. Resch, *Inferring causal structure: a quantum advantage*, preprint arXiv:1406.5036, (2014).

Generalizability of Causal and Statistical Relations

Elias Bareinboim
Cognitive Systems Laboratory
Computer Science Department
University of California, Los Angeles
eb@cs.ucla.edu

Abstract

The problem of generalizability of empirical findings (experimental and observational) to new environments, settings, and populations is one of the central problems in causal inference. Experiments in the sciences are invariably conducted with the intent of being used elsewhere (e.g., outside the laboratory), where conditions are likely to be different. This practice is based on the premise that, due to certain commonalities between the source and target environments, causal claims would be valid even where experiments have never been performed. Despite the extensive amount of empirical work relying on this premise, practically no formal treatments have been able to determine the conditions under which generalizations are valid, in some formal sense.

Our work develops a theoretical framework for understanding, representing, and algorithmizing the generalization problem as encountered in many practical settings in data-intensive fields. Our framework puts many apparently disparate generalization problems under the same theoretical umbrella. In this talk, I will start with a brief review of the basic concepts, principles, and mathematical tools necessary for reasoning about causal and counterfactual relations [1, 2, 3]. I will then introduce two special problems under the generalization umbrella.

First, I will discuss "transportability" [4, 5, 6], that is, how information acquired by experiments in one setting can be reused to answer queries in another, possibly different setting where only limited information can be collected. This question embraces several sub-problems treated informally in the literature under rubrics such as "external validity" [7, 8], "meta-analysis" [9], "heterogeneity" [10], "quasi-experiments" [11, Ch. 3]. Further, I will discuss selection bias [12, 13, 14], that is, how knowledge from a sampled subpopulation can be generalized to the entire population when sampling selection is not random, but determined by variables in the analysis, which means units are preferentially excluded from the sample.

In both problems, we provide complete conditions and algorithms to support the inductive step required in the corresponding task. This characterization distinguishes between estimable and non-estimable queries, and identifies which pieces of scientific knowledge need to be collected in each study to construct a bias-free estimate of the target query. The problems discussed in this work have applications in several empirical sciences such as Bioinformatics, Medicine, Economics, Social Sciences as well as in data-driven fields such as Machine Learning, Artificial Intelligence and Statistics.

References

[1] J. Pearl. The deductive approach to causal inference. *Journal of Causal Inference*, 2(2):115–130, 2014.

[2] P. Spirtes, C.N. Glymour, and R. Scheines. *Causation, Prediction, and Search*. Springer-Verlag, New York, 1993.

[3] E. Bareinboim, C. Brito, and J. Pearl. Local characterizations of causal bayesian networks. *Lecture Notes in Artificial Intelligence*, 7205:1–17, 2012.

[4] J. Pearl and E. Bareinboim. External validity: From *do*-calculus to transportability across populations. *Statistical Science*, forthcoming, 2014.

[5] E. Bareinboim and J. Pearl. Causal transportability with limited experiments. In *Proceedings of the Twenty-Seventh National Conference on Artificial Intelligence*, pages 95–101, Menlo Park, CA, 2013. AAAI Press.

[6] E. Bareinboim and J. Pearl. A general algorithm for deciding transportability of experimental results. *Journal of Causal Inference*, 1(1):107–134, 2013.

[7] D. Campbell and J. Stanley. *Experimental and Quasi-Experimental Designs for Research*. Wadsworth Publishing, Chicago, 1963.

[8] C. Manski. *Identification for Prediction and Decision*. Harvard University Press, Cambridge, Massachusetts, 2007.

[9] Gene V. Glass. Primary, secondary, and meta-analysis of research. *Educational Researcher*, 5(10):pp. 3–8, 1976.

[10] M. Höfler, A.T. Gloster, and J. Hoyer. Causal effects in psychotherapy: Counterfactuals counteract overgeneralization. *Psychotherapy Research*, 2010.

[11] W.R. Shadish, T.D. Cook, and D.T. Campbell. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton-Mifflin, Boston, second edition, 2002.

[12] V. Didelez, S. Kreiner, and N. Keiding. Graphical models for inference under outcome-dependent sampling. *Statistical Science*, 25(3):368–387, 2010.

[13] E. Bareinboim and J. Pearl. Controlling selection bias in causal inference. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 100–108. JMLR, April 21-23 2012.

[14] E. Bareinboim, J. Tian, and J. Pearl. Recovering from selection bias in causal and statistical inference. In *Proceedings of the Twenty-Eight National Conference on Artificial Intelligence (AAAI 2014)*, Menlo Park, CA, 2014. AAAI Press.

Estimating Causal Effects by Bounding Confounding

Philipp Geiger, Dominik Janzing, Bernhard Schölkopf
Max Planck Institute for Intelligent Systems
Spemannstraße 38, 72076 Tübingen, Germany
{pgeiger, janzing, bs}@tuebingen.mpg.de

Abstract

Assessing the causal effect of a treatment variable X on an outcome variable Y is usually difficult due to the existence of unobserved common causes. Without further assumptions, observed dependences do not even prove the existence of a causal effect from X to Y. It is intuitively clear that strong statistical dependences between X and Y do provide evidence for X influencing Y if the influence of common causes is known to be weak. We propose a framework that formalizes effect versus confounding in various ways and derive upper/lower bounds on the effect in terms of a priori given bounds on confounding. The formalization includes information theoretic quantities like information flow and causal strength, as well as other common notions like effect of treatment on the treated (ETT). We discuss several scenarios where upper bounds on the strength of confounding can be derived. This justifies to some extent human intuition which assumes the presence of causal effect when strong (e.g., close to deterministic) statistical relations are observed.