

# Joint Causal Inference from Multiple Contexts

**Joris M. Mooij**

*Institute for Informatics, University of Amsterdam  
Amsterdam, The Netherlands*

J.M.MOOIJ@UVA.NL

**Sara Magliacane**

*IBM Research  
Yorktown Heights, United States*

SARA.MAGLIACANE@IBM.COM

*Institute for Informatics, University of Amsterdam  
Amsterdam, The Netherlands*

**Tom Claassen**

*Radboud University Nijmegen  
Nijmegen, The Netherlands*

TOMC@CS.RU.NL

*Institute for Informatics, University of Amsterdam  
Amsterdam, The Netherlands*

**Editor:** ??

## Abstract

The gold standard for discovering causal relations is by means of experimentation. Over the last decades, alternative methods have been proposed that can infer causal relations between variables from certain statistical patterns in purely observational data. We introduce *Joint Causal Inference (JCI)*, a novel approach to causal discovery from multiple data sets that elegantly unifies both approaches. JCI is a causal modeling approach rather than a specific algorithm, and it can be used in combination with any causal discovery algorithm that can take into account certain background knowledge. The main idea is to reduce causal discovery from multiple datasets originating from different contexts (e.g., different experimental conditions) to causal discovery from a single pooled dataset by adding auxiliary context variables and incorporating applicable background knowledge on the causal relationships involving the context variables. We propose different flavours of JCI that differ in the amount of background knowledge that is assumed. JCI can deal with several different types of interventions in a unified fashion, does not require knowledge on intervention targets or types in case of interventional data, and allows one to fully exploit all the information in the joint distribution on system and context variables. We explain how some well-known causal discovery algorithms can be seen as implementations of the JCI framework, but we also propose novel implementations that are simple adaptations of existing causal discovery methods for purely observational data to the JCI setting. We evaluate different implementations of the JCI approach on synthetic data and on flow cytometry protein expression data and conclude that JCI implementations can outperform state-of-the-art causal discovery algorithms.

**Keywords:** Causal Discovery, Structure Learning, Observational and Experimental Data, Interventions, Randomized Controlled Trials

©???? Joris M. Mooij, Sara Magliacane and Tom Claassen.

License: CC-BY 4.0, see <https://creativecommons.org/licenses/by/4.0/>. Attribution requirements are provided at <http://jmlr.org/papers/v?/?????????.html>.

## 1. Introduction

The aim of causal discovery is to learn the causal relations between variables of a system of interest from data. As a simple example, suppose a researcher wants to find out whether playing violent computer games causes aggressive behavior. She gathers observational data by taking a sample from several high schools in different countries and observes a significant correlation between the daily amount of hours spent on playing violent computer games, and aggressive behavior at school. This in itself does not yet imply a causal relation between the two. Indeed, an alternative explanation of the observed correlation could be the presence of a confounder (a latent common cause), for example, a genetic predisposition towards violence that makes the carrier particularly enjoy such games and also make him behave more aggressively. The most reliable way to establish whether playing violent computer games causes aggressive behavior, is to use *experimentation*, for example by means of a randomized controlled trial (Fisher, 1935). This would imply assigning each pupil to one out of two groups randomly, where the pupils in one group are forced to play violent computer games for several hours a day, while the pupils in the other group are forced to abstain from playing those games. After several months, the aggressive behavior in both groups is measured. If a significant correlation between group and outcome is observed (or equivalently, the outcome is significantly different between the two groups) it can then be concluded that playing violent computer games indeed causes aggressive behavior.

Given the ethical and practical problems that such an experiment would involve, one might wonder whether there are alternative ways to answer this question. One such alternative is to combine data from different contexts. For example, in some countries the government may have decided to forbid certain ultra-violent games from being sold, and there is data available both before and after the ban. In addition, some schools may have introduced certain measures to discourage aggressive behavior. By combining the data from these different contexts in an appropriate way, one may be able to identify the presence or absence of a causal effect of playing violent computer games on aggressive behavior. For example, in the setting of Figure 1(c), the causal relationship between the two variables of interest turns out to be generically identifiable from conditional independence relationships in pooled data from all the contexts. In particular, in that case the observed correlation between playing violent computer games and aggressive behavior could be unambiguously attributed to a causal effect of one on the other. In this paper, we propose a simple and general way to combine data sets from different contexts that enables one to draw such strong causal conclusions.

While experimentation is still the gold standard to establish causal relationships, researchers realized in the early nineties that there are other methods that require only *purely observational* data (Spirtes et al., 2000; Pearl, 2009). Many methods for causal discovery from purely observational data have been proposed over the last decades, relying on different assumptions. These can be roughly divided into *constraint-based* causal discovery methods, such as the PC (Spirtes et al., 2000), IC (Pearl, 2009) and FCI algorithms (Spirtes et al., 1995; Zhang, 2008), *score-based* causal discovery methods (e.g., Heckerman et al., 1995; Chickering, 2002; Koivisto and Sood, 2004), and methods exploiting other statistical patterns in the joint distribution (e.g., Mooij et al., 2016; Peters et al., 2017). Originally, these

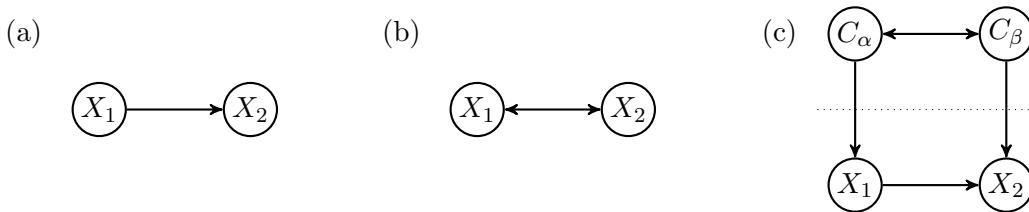


Figure 1: Different causal graphs relating  $X_1$ , the daily amount of hours spent on playing violent computer games, and  $X_2$ , a measure of aggressive behavior. (a) Playing violent computer games causes aggressive behavior; (b) The observed correlation between  $X_1$  and  $X_2$  is explained by a latent confounder, e.g., a genetic predisposition towards violence. (c) Hypothetical causal graph also involving context variables  $C_\alpha$ , which indicates whether ultra-violent games have been banned by the government, and  $C_\beta$ , which indicates school interventions to stimulate social behavior. Without considering contexts, it is not possible to distinguish between (a) and (b) based on conditional independences in the data. In scenario (c), JCI allows one to infer from conditional independences in the pooled data that  $X_1$  causes  $X_2$  and that  $X_1$  and  $X_2$  are not confounded (assuming that context variables  $C_\alpha$  and  $C_\beta$  are not potentially caused by system variables  $X_1$  and  $X_2$ ).

methods were designed to estimate the causal graph of the system from a single dataset corresponding to a single (purely observational) context.

More recently, various causal discovery methods have been proposed that extend these techniques to deal with multiple datasets from different contexts. As an example, the datasets may correspond with a baseline of purely observational data of measurements concerning the “natural” state of the system, and data of measurements under different perturbations of the system caused by external interventions on the system.<sup>1</sup> More generally, they can correspond to measurements of the system in different environments. These methods can be divided into two main approaches:

- (a) methods that obtain statistics or constraints from each context separately and then construct a single context-independent causal graph by combining these statistics (Claassen and Heskes, 2010; Tillman and Spirtes, 2011; Hyttinen et al., 2012, 2014; Triantafillou and Tsamardinos, 2015; Rothenhäusler et al., 2015; Forré and Mooij, 2018),
- (b) methods that pool all data and construct a single context-independent causal graph directly from the pooled data (Cooper, 1997; Cooper and Yoo, 1999; Tian and Pearl, 2001; Sachs et al., 2005; Eaton and Murphy, 2007; Chen et al., 2007; Hauser and Bühlmann, 2012; Mooij and Heskes, 2013; Peters et al., 2016; Oates et al., 2016a; Zhang et al., 2017).

1. In certain parts of the causal discovery literature, the word “intervention” has become synonymous to “surgical/atomic/hard/perfect” intervention (i.e., an intervention that precisely sets a variable or set of variables to a certain value without directly affecting the other variables in the system), but in this work we use it in the more general meaning of any external perturbation of the system.

In this paper, we introduce *Joint Causal Inference (JCI)*, a framework for causal discovery from multiple datasets corresponding to measurements that have been performed in different contexts, which takes the latter approach.

The key idea of JCI is to (i) consider auxiliary context variables that describe the context of each data set, (ii) pool all the data from different contexts, including the values of the context variables, and finally (iii) apply standard causal discovery methods to the pooled data, incorporating appropriate background knowledge on the causal relationships involving the context variables. The framework is simple and very generally applicable as it allows one to deal with latent confounding and cycles (if the causal discovery method supports this), and various types of interventions in a unified way. It does not require background knowledge on the intervention types and targets, making it very suitable to the application on complex systems in which the effects of certain interventions are not known *a priori*, a situation that often occurs in practice. JCI can be implemented using any causal discovery method that can incorporate the appropriate background knowledge on the relationships between context and system variables. This allows one to benefit from the availability of sophisticated and powerful causal discovery methods that have been primarily designed for a single data set from a single context by extending their application domain to the setting of multiple data sets from multiple contexts. At the same time, JCI accommodates various well-known causal discovery methods as special cases, such as the standard randomized controlled trial setting, LCD (Cooper, 1997) and ICP (Peters et al., 2016). By explicitly introducing the context variables and treating them analogously to the system variables (but with additional background knowledge about their causal relations with the system variables), JCI makes it possible to elegantly combine the principles of causal discovery from experimentation with those of causal discovery from purely observational data.

This paper is structured as follows. In Section 2 we describe the relevant causal modeling and discovery concepts and define terminology and notation. In Section 3 we introduce the JCI framework, show how it can be implemented using various causal discovery methods, and compare it with related work. In Section 4 we report experimental results on synthetic and flow cytometry data. We conclude in Section 5.

## 2. Background

In this section, we present the background material on which we will base our exposition. We start in Section 2.1 with a brief subsection stating the basic definitions and results in the field of graphical causal modeling that we will use in this paper. In addition to covering material that is standard in the field, we also discuss more recent extensions to the cyclic setting. Then, in Section 2.2, we discuss the key idea of causal discovery from experimentation (in the setting of a randomized controlled trial, or A/B-testing) in these terms. We finish with Section 2.3 that briefly illustrates the basic idea underlying constraint-based causal discovery from purely observational data in a simple setting.

### 2.1 Graphical Causal Modeling

We briefly summarize some basic definitions and results in the field of graphical causal modeling. For more details, we refer the reader to Pearl (2009) and Bongers et al. (2018).

## 2.1.1 DIRECTED MIXED GRAPHS

A *Directed Mixed Graph* (DMG) is a graph  $\mathcal{G} = \langle \mathcal{V}, \mathcal{E}, \mathcal{F} \rangle$  with nodes  $\mathcal{V}$  and two types of edges: *directed* edges  $\mathcal{E} \subseteq \mathcal{V}^2$ , and *bidirected* edges  $\mathcal{F} \subseteq \{\{i, j\} : i, j \in \mathcal{V}, i \neq j\}$ . We will denote a directed edge  $(i, j) \in \mathcal{E}$  as  $i \rightarrow j$  or  $j \leftarrow i$ , and call  $i$  a *parent* of  $j$  and  $j$  a *child* of  $i$ . We denote all parents of  $j$  as  $\text{PA}_{\mathcal{G}}(j) := \{i \in \mathcal{V} : i \rightarrow j \in \mathcal{E}\}$ , and all children of  $i$  as  $\text{CH}_{\mathcal{G}}(i) := \{j \in \mathcal{V} : i \rightarrow j \in \mathcal{E}\}$ . We allow for self-cycles  $i \rightarrow i$ , so a variable can be its own parent and child. We will denote a bidirected edge  $\{i, j\} \in \mathcal{F}$  as  $i \leftrightarrow j$  or  $j \leftrightarrow i$ , and call  $i$  and  $j$  *spouses*. Two nodes  $i, j \in \mathcal{V}$  are called *adjacent in  $\mathcal{G}$*  if they are connected by an edge (or multiple edges), i.e., if  $i \rightarrow j \in \mathcal{E}$  or  $i \leftarrow j \in \mathcal{E}$  or  $i \leftrightarrow j \in \mathcal{F}$ . For a subset of nodes  $\mathcal{W} \subseteq \mathcal{V}$ , we define the *induced subgraph*  $\mathcal{G}_{\mathcal{W}} := (\mathcal{W}, \mathcal{E} \cap \mathcal{W}^2, \mathcal{F} \cap \{\{i, j\} : i, j \in \mathcal{W}, i \neq j\})$ , i.e., with nodes  $\mathcal{W}$  and exactly those edges of  $\mathcal{G}$  that connect two nodes in  $\mathcal{W}$ .

A *walk between  $i, j \in \mathcal{V}$*  is a tuple  $\langle i_0, e_1, i_1, e_2, i_3, \dots, e_{n-1}, i_n \rangle$  of alternating nodes and edges in  $\mathcal{G}$  ( $n \geq 0$ ), such that all  $i_0, \dots, i_n \in \mathcal{V}$ , all  $e_1, \dots, e_{n-1} \in \mathcal{E} \cup \mathcal{F}$ , starting with node  $i_0 = i$  and ending with node  $i_n = j$ , and such that for all  $k = 1, \dots, n-1$ , the edge  $e_k$  connects the two nodes  $i_{k-1}$  and  $i_k$  in  $\mathcal{G}$ . If the walk contains each node at most once, it is called a *path*. A *trivial walk (path)* consists just of a single node and zero edges. A *directed path from  $i \in \mathcal{V}$  to  $j \in \mathcal{V}$*  is a path between  $i$  and  $j$  such that every edge  $e_k$  on the path is of the form  $i_{k-1} \rightarrow i_k$ , i.e., every edge is directed and points away from  $i$ . By repeatedly taking parents, we obtain the *ancestors* of  $j$ :  $\text{AN}_{\mathcal{G}}(j) := \{i \in \mathcal{V} : i = i_0 \rightarrow i_1 \rightarrow i_2 \rightarrow \dots \rightarrow i_k = j \text{ in } \mathcal{G}\}$ . Similarly, we define the *descendants* of  $i$ :  $\text{DE}_{\mathcal{G}}(i) := \{j \in \mathcal{V} : i = i_0 \rightarrow i_1 \rightarrow i_2 \rightarrow \dots \rightarrow i_k = j \text{ in } \mathcal{G}\}$ . In particular, each node is ancestor and descendant of itself. The *strongly-connected component* of  $i \in \mathcal{V}$  is defined as  $\text{SC}_{\mathcal{G}}(i) := \text{AN}_{\mathcal{G}}(i) \cap \text{DE}_{\mathcal{G}}(i)$ . We extend the definitions to sets  $I \subseteq \mathcal{V}$  by setting  $\text{AN}_{\mathcal{G}}(I) := \cup_{i \in I} \text{AN}_{\mathcal{G}}(i)$ , and similarly for  $\text{DE}_{\mathcal{G}}(I)$  and  $\text{SC}_{\mathcal{G}}(I)$ . A *directed cycle* is a directed path from  $i$  to  $j$  such that in addition,  $j \rightarrow i \in \mathcal{E}$ . A directed mixed graph  $\mathcal{G}$  is *acyclic* if it does not contain any directed cycle, in which case it is known as an *Acyclic Directed Mixed Graph (ADMG)*. A directed mixed graph that does not contain bidirected edges is known as a *Directed Graph (DG)*. If a directed mixed graph does not contain bidirected edges and is acyclic, it is called a *Directed Acyclic Graph (DAG)*.

A subpath (subwalk)  $\langle i_{k-1}, e_k, i_k, e_{k+1}, i_{k+1} \rangle$  of a path (walk)  $\langle i_0, e_1, i_1, e_2, i_3, \dots, e_{n-1}, i_n \rangle$  in  $\mathcal{V}$  is said to form a *collider on  $i_k$*  if the two edges  $e_k, e_{k+1}$  meet head-to-head on their shared node  $i_k$  (i.e., if the two subsequent edges are of the form  $i_{k-1} \rightarrow i_k \leftarrow i_{k+1}$ ,  $i_{k-1} \leftrightarrow i_k \leftarrow i_{k+1}$ ,  $i_{k-1} \rightarrow i_k \leftrightarrow i_{k+1}$ , or  $i_{k-1} \leftrightarrow i_k \leftrightarrow i_{k+1}$ ). Otherwise, the subpath (subwalk) is called a *non-collider on  $i_k$*  (i.e., if the two subsequent edges are of the form  $i_{k-1} \rightarrow i_k \rightarrow i_{k+1}$ ,  $i_{k-1} \leftarrow i_k \leftarrow i_{k+1}$ ,  $i_{k-1} \leftarrow i_k \rightarrow i_{k+1}$ ,  $i_{k-1} \leftrightarrow i_k \rightarrow i_{k+1}$ , or  $i_{k-1} \leftarrow i_k \leftrightarrow i_{k+1}$ ). Also the end points  $i_0, i_n$  of the walk  $\pi$  will be referred to as non-colliders on  $\pi$ . We will denote the colliders on a walk  $\pi$  as  $\text{COL}(\pi)$  and the non-colliders on  $\pi$  (including the end-points of  $\pi$ ) as  $\text{NCOL}(\pi)$ .

## 2.1.2 STRUCTURAL CAUSAL MODELS

Directed Mixed Graphs form a convenient graphical representation for variables (labelled by the nodes) and their functional relations (expressed by the edges) in a *Structural Causal Model (SCM)* (Pearl, 2009), also known as a (non-parametric) *Structural Equation Model (SEM)* (Wright, 1921). Several slightly different definitions of SCMs have been proposed in

the literature, which all have their (dis)advantages. Here we use a variant of the definition in Bongers et al. (2018) that is most convenient for our purposes here.

**Definition 1** A *Structural Causal Model (SCM)* is a tuple  $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \mathcal{H}, \mathcal{X}, \mathcal{E}, \mathbf{f}, \mathbb{P}_{\mathcal{E}} \rangle$  of:

- (i) a finite index set  $\mathcal{I}$  for the endogenous variables in the model;
- (ii) a finite index set  $\mathcal{J}$  for the latent exogenous variables in the model;
- (iii) a directed graph  $\mathcal{H}$  with nodes  $\mathcal{I} \cup \mathcal{J}$ , and directed edges pointing from  $\mathcal{I} \cup \mathcal{J}$  to  $\mathcal{I}$ ;
- (iv) a product of Borel<sup>2</sup> spaces  $\mathcal{X} = \prod_{i \in \mathcal{I}} \mathcal{X}_i$ , which define the domains of the endogenous variables;
- (v) a product of Borel spaces  $\mathcal{E} = \prod_{j \in \mathcal{J}} \mathcal{E}_j$ , which define the domains of the exogenous variables;
- (vi) a measurable function  $\mathbf{f} : \mathcal{X} \times \mathcal{E} \rightarrow \mathcal{X}$ , the causal mechanism, such that each of its components  $f_i$  only depends on a particular subset of the variables, as specified by the directed graph  $\mathcal{H}$ :

$$f_i : \mathcal{X}_{\text{PA}_{\mathcal{H}}(i) \cap \mathcal{I}} \times \mathcal{E}_{\text{PA}_{\mathcal{H}}(i) \cap \mathcal{J}} \rightarrow \mathcal{X}_i, \quad i \in \mathcal{I};$$

- (vii) a product probability measure  $\mathbb{P}_{\mathcal{E}} = \prod_{j \in \mathcal{J}} \mathbb{P}_{\mathcal{E}_j}$  on  $\mathcal{E}$  specifying the exogenous distribution.

An SCM is often specified informally by specifying only the structural equations and the density<sup>3</sup> of the exogenous distribution with respect to some product measure, for example:

$$\mathcal{M} : \begin{cases} X_i & = f_i(\mathbf{X}_{\text{PA}_{\mathcal{H}}(i) \cap \mathcal{I}}, \mathbf{E}_{\text{PA}_{\mathcal{H}}(i) \cap \mathcal{J}}) \\ p(\mathbf{E}) & = \prod_{j \in \mathcal{J}} p(E_j). \end{cases}$$

In discussing the properties of SCMs, the graphical representation of various objects and their relations in Figure 2 may be helpful. This shows how the SCM is the central object containing all information, and how other representations can be derived from the SCM. In the rest of this section, we will discuss this in more detail.

We refer to the graph  $\mathcal{H}$  as the *augmented functional graph* of  $\mathcal{M}$ . The *functional graph* of  $\mathcal{M}$ , denoted  $\mathcal{G}(\mathcal{M})$ , is the directed mixed graph with nodes  $\mathcal{I}$ , directed edges  $i_1 \rightarrow i_2$  iff  $i_1 \rightarrow i_2 \in \mathcal{H}$ , and bidirected edges  $i_1 \leftrightarrow i_2$  iff there exists  $j \in \text{PA}_{\mathcal{H}}(i_1) \cap \text{PA}_{\mathcal{H}}(i_2) \cap \mathcal{J}$ .<sup>4</sup> If  $\mathcal{G}(\mathcal{M})$  is acyclic, we call the SCM  $\mathcal{M}$  *acyclic*, otherwise we call the SCM *cyclic*.

A pair of random variables  $(\mathbf{X}, \mathbf{E})$  is called a *solution* of the SCM  $\mathcal{M}$  if  $\mathbf{X} = (X_i)_{i \in \mathcal{I}}$  with  $X_i \in \mathcal{X}_i$  for all  $i \in \mathcal{I}$ ,  $\mathbf{E} = (E_j)_{j \in \mathcal{J}}$  with  $E_j \in \mathcal{E}_j$  for all  $j \in \mathcal{J}$ , the distribution  $\mathbb{P}(\mathbf{E})$  is equal to the exogenous distribution  $\mathbb{P}_{\mathcal{E}}$ , and the *structural equations*:

$$X_i = f_i(\mathbf{X}_{\text{PA}_{\mathcal{H}}(i) \cap \mathcal{I}}, \mathbf{E}_{\text{PA}_{\mathcal{H}}(i) \cap \mathcal{J}}) \quad \text{a.s.}$$

- 
- 2. A *Borel space* is both a measurable and a topological space, such that the sigma-algebra is generated by the open sets. Most spaces that one encounters in applications as the domain of a random variable are Borel spaces.
  - 3. We denote a probability measure (or distribution) of a random variable  $\mathbf{X}$  by  $\mathbb{P}(\mathbf{X})$ , and a density of  $\mathbf{X}$  with respect to some product measure by  $p(\mathbf{X})$ .
  - 4. This definition of functional graph makes a slight simplification: a more precise definition would leave out edges that are redundant. For example, if the structural equation for  $X_2$  reads  $X_2 = 0 \cdot X_1 + X_3$  it could be that  $1 \rightarrow 2 \in \mathcal{H}$ , but this edge would not appear in  $\mathcal{G}(\mathcal{M})$ . For the rigorous version of this definition, see Bongers et al. (2018).

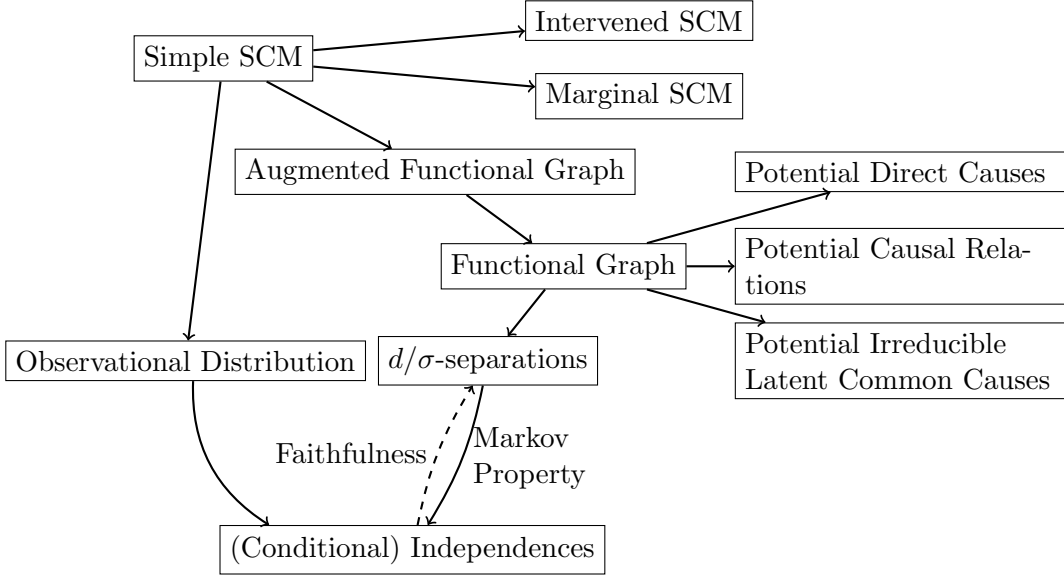


Figure 2: Relationships between various representations for simple SCMs. Directed edges represent mappings. Intervened and marginal SCMs are always defined and are also simple.

hold for all  $i \in \mathcal{I}$ . For acyclic SCMs, solutions exist and have a unique distribution that is determined by the SCM. This is not generally the case in cyclic SCMs, as discussed in detail by Bongers et al. (2018). Indeed, cyclic SCMs could have no solution at all, or could have multiple solutions with different distributions.

**Definition 2** An SCM  $\mathcal{M}$  is said to be uniquely solvable w.r.t.  $\mathcal{O} \subseteq \mathcal{I}$  if there exists a measurable mapping  $\mathbf{g}_{\mathcal{O}} : \mathcal{X}_{(\text{PA}_{\mathcal{H}}(\mathcal{O}) \setminus \mathcal{O}) \cap \mathcal{I}} \times \mathcal{E}_{\text{PA}_{\mathcal{H}}(\mathcal{O}) \cap \mathcal{J}} \rightarrow \mathcal{X}_{\mathcal{O} \cap \mathcal{I}}$  such that for  $\mathbb{P}_{\mathbf{e}}$ -almost every  $\mathbf{e}$  for all  $\mathbf{x} \in \mathcal{X}$ :

$$\mathbf{x}_{\mathcal{O}} = \mathbf{g}_{\mathcal{O}}(\mathbf{x}_{(\text{PA}_{\mathcal{H}}(\mathcal{O}) \setminus \mathcal{O}) \cap \mathcal{I}}, \mathbf{e}_{\text{PA}_{\mathcal{H}}(\mathcal{O}) \cap \mathcal{J}}) \iff \mathbf{x}_{\mathcal{O}} = \mathbf{f}_{\mathcal{O}}(\mathbf{x}, \mathbf{e}).$$

If  $\mathcal{M}$  is uniquely solvable with respect to  $\mathcal{I}$ , then it induces a unique *observational distribution*  $\mathbb{P}_{\mathcal{M}}(\mathbf{X})$ . In general, the SCM might induce several observational distributions.

Given an SCM that models a certain system, we can model the system after an idealized intervention as follows.

**Definition 3** Let  $\mathcal{M}$  be an SCM. The perfect (“surgical”) intervention with target  $I \subseteq \mathcal{I}$  and value  $\xi_I \in \mathcal{X}_I$  induces the intervened SCM  $\mathcal{M}_{\text{do}(I, \xi_I)}$  by copying  $\mathcal{M}$ , but letting  $\tilde{\mathcal{H}}$  be  $\mathcal{H}$  without the edges  $\{j \rightarrow i : j \in \mathcal{I} \cup \mathcal{J}, i \in I\}$ , and modifying the causal mechanism into  $\tilde{\mathbf{f}}$  such that

$$\tilde{f}_i(\mathbf{x}, \mathbf{e}) = \begin{cases} \xi_i & i \in I \\ f_i(\mathbf{x}, \mathbf{e}) & i \notin I. \end{cases}$$

The interpretation is that the causal mechanisms that normally determine the values of the components  $i \in I$  are replaced by mechanisms that assign the values  $\xi_i$ . Other types of

interventions are possible as well (see also Section 3.3). If the intervened SCM  $\mathcal{M}_{\text{do}(I, \xi_I)}$  induces a unique observational distribution, this is denoted as  $\mathbb{P}_{\mathcal{M}}(\mathbf{X} \mid \text{do}(I, \xi_I))$  and referred to as the *interventional distribution of  $\mathcal{M}$  under the perfect intervention  $\text{do}(I, \xi_I)$* . Pearl (2009) derived the *do-calculus* for acyclic SCMs, consisting of three rules that express relationships between interventional distributions of an SCM. Forré and Mooij (2019) recently extended the causal do-calculus to certain cyclic SCMs.

In this work, for simplicity of exposition we will focus on a certain subclass of SCMs that have many convenient properties:

**Definition 4** *We call an SCM  $\mathcal{M}$  simple if it is uniquely solvable with respect to any subset  $\mathcal{O} \subseteq \mathcal{I}$ .*

Simple SCMs provide a special case of *modular* SCMs (Forré and Mooij, 2017). They induce a unique observational distribution, and their marginalizations are always defined (Bongers et al., 2018). The class of simple SCMs is closed under perfect interventions and marginalizations. Hence, also all their perfect interventional distributions are uniquely defined. Later on we will see that simple SCMs have even more convenient properties. All acyclic SCMs are simple. Simple SCMs can be thought of as the “smallest” generalization of acyclic SCMs that allows for cycles yet preserves many of the convenient properties that acyclic SCMs have.

### 2.1.3 STRUCTURAL CAUSAL MODELS: MARKOV PROPERTIES

Under certain conditions, the functional graph  $\mathcal{G}(\mathcal{M})$  of an SCM  $\mathcal{M}$  can be interpreted as a statistical graphical model, i.e., it allows one to read off (conditional) independences that must hold in the observational distribution  $\mathbb{P}_{\mathcal{M}}(\mathbf{X})$ . One of the most common formulations of such *Markov properties* involves the following notion of *d-separation*, first proposed by Pearl (1986) in the context of DAGs, and later shown to be more generally applicable:<sup>5</sup>

**Definition 5 (*d-separation*)** *We say that a walk  $\langle i_0 \dots i_k \rangle$  in DMG  $\mathcal{G} = \langle \mathcal{V}, \mathcal{E}, \mathcal{F} \rangle$  is *d-blocked* by  $C \subseteq \mathcal{V}$  if:*

- (i) *its first node  $i_0 \in C$  or its last node  $i_k \in C$ , or*
- (ii) *it contains a collider on a node  $i \notin \text{AN}_{\mathcal{G}}(C)$ , or*
- (iii) *it contains a non-collider on a node  $i \in C$ .*

*If all paths in  $\mathcal{G}$  between any node in set  $A \subseteq \mathcal{V}$  and any node in set  $B \subseteq \mathcal{V}$  are *d-blocked* by a set  $C \subseteq \mathcal{V}$ , we say that  $A$  is *d-separated* from  $B$  by  $C$ , and we write  $A \perp_{\mathcal{G}}^d B \mid C$ .*

In the general cyclic case, however, the notion of *d-separation* is too strong and needs to be replaced with a non-trivial generalization of *d-separation*, known as  *$\sigma$ -separation* (Forré and Mooij, 2017):

**Definition 6 ( *$\sigma$ -separation*)** *We say that a walk  $\langle i_0 \dots i_k \rangle$  in DMG  $\mathcal{G} = \langle \mathcal{V}, \mathcal{E}, \mathcal{F} \rangle$  is  *$\sigma$ -blocked* by  $C \subseteq \mathcal{V}$  if:*

- (i) *its first node  $i_0 \in C$  or its last node  $i_k \in C$ , or*
- (ii) *it contains a collider on a node  $i \notin \text{AN}_{\mathcal{G}}(C)$ , or*

---

5. It is also sometimes called “*m-separation*” in the ADMG literature.



- (iii) it contains a non-collider on a node  $i \in C$  that points to a neighboring node on the path in another strongly-connected component (i.e.,  $i_{k-1} \rightarrow i_k \rightarrow i_{k+1}$  or  $i_{k-1} \leftrightarrow i_k \rightarrow i_{k+1}$  with  $i_{k+1} \notin \text{SCG}(i_k)$ ,  $i_{k-1} \leftarrow i_k \leftarrow i_{k+1}$  or  $i_{k-1} \leftarrow i_k \leftrightarrow i_{k+1}$  with  $i_{k-1} \notin \text{SCG}(i_k)$ , or  $i_{k-1} \leftarrow i_k \rightarrow i_{k+1}$  with  $i_{k-1} \notin \text{SCG}(i_k)$  or  $i_{k+1} \notin \text{SCG}(i_k)$ ).

If all paths in  $\mathcal{G}$  between any node in set  $A \subseteq \mathcal{V}$  and any node in set  $B \subseteq \mathcal{V}$  are  $\sigma$ -blocked by a set  $C \subseteq \mathcal{V}$ , we say that  $A$  is  $\sigma$ -separated from  $B$  by  $C$ , and we write  $A \perp_{\mathcal{G}}^{\sigma} B \mid C$ .

Forré and Mooij (2017) showed the following fundamental result, where we follow the formulation specifically for the SCM setting from Bongers et al. (2018):

**Theorem 7 (Generalized Directed Global Markov Property)** *If an SCM  $\mathcal{M}$  is uniquely solvable w.r.t. each strongly-connected component in  $\mathcal{G}(\mathcal{M})$ , then its observational distribution  $\mathbb{P}_{\mathcal{M}}(\mathbf{X})$  exists and is unique. Furthermore, for any solution  $(\mathbf{X}, \mathbf{E})$  of  $\mathcal{M}$ , the Generalized Directed Global Markov Property holds with respect to the functional graph  $\mathcal{G}(\mathcal{M})$ :*

$$A \perp_{\mathcal{G}(\mathcal{M})}^{\sigma} B \mid C \implies \mathbf{X}_A \perp_{\mathbb{P}_{\mathcal{M}}(\mathbf{X})} \mathbf{X}_B \mid \mathbf{X}_C \quad \forall A, B, C \subseteq \mathcal{I}.$$

The following stronger statements have been derived for special cases by Forré and Mooij (2017) (where again we follow the terminology of Bongers et al. (2018)):

**Theorem 8 (Directed Global Markov Property)** *If an SCM  $\mathcal{M}$  satisfies at least one of the following three conditions:*

- (i)  $\mathcal{M}$  is acyclic;
- (ii) all endogenous spaces  $\mathcal{X}_i$  are discrete, and  $\mathcal{M}$  is uniquely solvable w.r.t. each ancestral subgraph of  $\mathcal{G}(\mathcal{M})$ ;
- (iii)  $\mathcal{M}$  is linear (i.e.,  $\mathcal{X}_i = \mathbb{R}$  for each  $i \in \mathcal{I}$ ,  $\phi : \mathcal{I} \rightarrow \mathcal{J}$  is a bijection,  $\text{PA}_{\mathcal{H}}(i) = \phi(i)$ ,  $\mathcal{E}_j = \mathbb{R}$  for each  $j \in \mathcal{J}$  and each causal mechanism  $f_i : \mathcal{X}_{\text{PA}_{\mathcal{H}}(i) \cap \mathcal{I}} \times \mathcal{E}_{\phi(i)} \rightarrow \mathcal{X}_i$  is linear, for  $i \in \mathcal{I}$ ), its exogenous variables have a density  $p(\mathbf{E})$  with respect to Lebesgue measure, and  $\mathcal{M}$  has a solution;

then the observational distribution  $\mathbb{P}_{\mathcal{M}}(\mathbf{X})$  exists and is unique. Furthermore, the Directed Global Markov Property holds for any solution  $(\mathbf{X}, \mathbf{E})$  of  $\mathcal{M}$  with respect to the functional graph  $\mathcal{G}(\mathcal{M})$ :

$$A \perp_{\mathcal{G}(\mathcal{M})}^d B \mid C \implies \mathbf{X}_A \perp_{\mathbb{P}_{\mathcal{M}}(\mathbf{X})} \mathbf{X}_B \mid \mathbf{X}_C \quad \forall A, B, C \subseteq \mathcal{I}.$$

The acyclic case is well-known and was first shown by Richardson (2003). The solvability conditions in the discrete and linear cases above are implied by structural solvability w.r.t. each strongly-connected component in  $\mathcal{G}(\mathcal{M})$ .

Simple SCMs (see Definition 4) also have convenient Markov properties: They induce a unique observational distribution that satisfies the Generalized Directed Global Markov Property, and for any perfect intervention, they induce a unique interventional distribution that satisfies the Generalized Directed Global Markov Property with respect to the mutilated graph. Since the do-calculus relies on Markov properties, it can be extended to simple SCMs (Forré and Mooij, 2019).

The starting point for constraint-based approaches to causal discovery from observational data is to assume that the data is modelled by an (unknown) SCM  $\mathcal{M}$ , such that its

observational distribution  $\mathbb{P}_{\mathcal{M}}(\mathbf{X})$  exists, is unique, and satisfies a Markov property with respect to its functional graph  $\mathcal{G}(\mathcal{M})$ . In addition, one usually assumes the *faithfulness assumption* to hold (Spirtes et al., 2000; Pearl, 2009), i.e., that the functional graph explains *all* conditional independences present in the observational distribution. For the cases in which  $d$ -separation applies, this amounts to assuming the following implication:

$$A \underset{\mathcal{G}(\mathcal{M})}{\overset{d}{\perp}} B | C \iff \mathbf{X}_A \underset{\mathbb{P}_{\mathcal{M}}(\mathbf{X})}{\perp} \mathbf{X}_B | \mathbf{X}_C \quad \forall A, B, C \subseteq \mathcal{V}.$$

Meek (1995) has shown completeness properties of  $d$ -separation; more specifically, Meek (1995) showed that  $d$ -separation holds generically for DAGs if (i) all variable domains are finite, or (ii) if all variables are real-valued, linearly related and have a multivariate Gaussian distribution. This provides some justification for assuming faithfulness. On the other hand, no completeness results are known yet for the general cyclic case in which  $\sigma$ -separation applies. Nevertheless, we believe that such results can be shown, and we will assume for simple SCMs a similar faithfulness assumption as for the  $d$ -separation case:

$$A \underset{\mathcal{G}(\mathcal{M})}{\overset{\sigma}{\perp}} B | C \iff \mathbf{X}_A \underset{\mathbb{P}_{\mathcal{M}}(\mathbf{X})}{\perp} \mathbf{X}_B | \mathbf{X}_C \quad \forall A, B, C \subseteq \mathcal{V}.$$

#### 2.1.4 STRUCTURAL CAUSAL MODELS: CAUSAL INTERPRETATION

If the functional graph  $\mathcal{G}(\mathcal{M})$  is a DAG, then it has a straightforward causal interpretation (Peters et al., 2017). However, if  $\mathcal{G}(\mathcal{M})$  contains bidirected edges or cycles, its causal interpretation becomes more subtle (Bongers et al., 2018).

**Direct edges** If  $\mathcal{M}$  is acyclic, the directed edges in  $\mathcal{G}(\mathcal{M})$  represent *potential direct causal relations* between endogenous variables. Indeed, if  $i$  is a direct cause of  $j$  with respect to  $\mathcal{I}$  according to  $\mathcal{M}$ , then  $i \rightarrow j \in \mathcal{G}(\mathcal{M})$ . Under faithfulness, the converse holds if  $\mathcal{G}(\mathcal{M})$  is a DAG. If  $\mathcal{G}(\mathcal{M})$  contains bidirected edges, the converse does not necessarily hold under faithfulness, although it still holds generically. The causal interpretation of cyclic SCMs is more subtle. Bongers et al. (2018) show that if  $j \in \mathcal{I}$  has no *self-cycle*, i.e., if  $j \rightarrow j \notin \mathcal{G}(\mathcal{M})$ , then if  $i \in \mathcal{I}$  is a direct cause of  $j$  with respect to  $\mathcal{I}$  according to  $\mathcal{M}$ , there must be a directed edge  $i \rightarrow j \in \mathcal{G}(\mathcal{M})$ . On the other hand, if self-cycles are present, this relationship between direct edges in the functional graph  $\mathcal{G}(\mathcal{M})$  and direct causes may no longer hold. Simple SCMs do not contain self-cycles and hence this relation does hold for simple SCMs. In addition, the converse relation generically holds for simple SCMs.

**Directed paths** If  $\mathcal{M}$  is acyclic, the directed paths in  $\mathcal{G}(\mathcal{M})$  represent *potential causal relations* (indirect ones if the path consists of more than one direct edge). Indeed, if  $i$  is a cause of  $j$  according to  $\mathcal{M}$ , then  $i \in \text{AN}_{\mathcal{G}(\mathcal{M})}(j)$ . The converse only holds generically, even if  $\mathcal{G}(\mathcal{M})$  is a DAG and faithfulness holds. The cyclic case is again more subtle. Bongers et al. (2018) show that even if no self-cycles are present, larger cycles may lead to the situation in which  $i$  causes  $j$  yet there is no directed path from  $i$  to  $j$  in  $\mathcal{G}(\mathcal{M})$ . Additional assumptions need to be made to avoid this counter-intuitive behavior. For example, for simple SCMs it still holds that if  $i$  is a cause of  $j$  according to  $\mathcal{M}$ , then  $i \in \text{AN}_{\mathcal{G}(\mathcal{M})}(j)$ . The converse also holds generically for simple SCMs.

**Bidirected edges** In general, bidirected edges in  $\mathcal{G}(\mathcal{M})$  represent *potential irreducible latent common causes*. “Irreducible” here means that the latent common cause cannot be split into two separate causes that are statistically independent. If  $\mathcal{G}(\mathcal{M})$  is acyclic, bidirected edges in  $\mathcal{G}(\mathcal{M})$  also represent *potential latent confounding*. Indeed, in the acyclic case one can show that if  $i$  and  $j$  have latent confounding with respect to  $\mathcal{I}$  according to  $\mathcal{M}$ , then  $i \leftrightarrow j \in \mathcal{G}(\mathcal{M})$ . The converse holds only generically. For the cyclic case, the situation is again more complicated, and not completely understood at present, in particular because it is not obvious how to define “confounding” in a general cyclic setting. In this paper we will work around this complication.

**Simple SCMs** As we saw, the simple SCMs that we consider in this paper have a simple causal interpretation. They do not have self-cycles and hence the directed edges can be interpreted as potential direct causes. Marginalizations of simple SCMs are always defined, and the simplicity property is inherited by any marginalized SCM. Therefore, also the interpretation of directed paths as potential causes prevails for simple SCMs. Finally, bidirected edges can be interpreted as representing potential irreducible latent common causes. Taken together, we conclude that the functional graph  $\mathcal{G}(\mathcal{M})$  of a simple SCM can be interpreted as its *potential causal graph*.

## 2.2 Causal Discovery by Experimentation

The gold standard for causal discovery is by means of experimentation. For example, randomized controlled trials (Fisher, 1935) form the foundation of modern evidence-based medicine. In engineering, A/B-testing is a common protocol to optimize certain causal effects of an engineered system. Toddlers learn causal representations of the world through playful experimentation.

We will discuss here the simplest randomized controlled trial setting by formulating it in terms of the graphical causal terminology introduced in the last section. The experimental procedure is as follows. Consider two variables, “treatment”  $C$  and “outcome”  $X$ . In the simplest setting, one considers a binary treatment variable, where  $C = 1$  corresponds to “treat with drug” and  $C = 0$  corresponds to “treat with placebo”. For example, the drug could be aspirin, and outcome could be the severity of headache perceived two hours later. Patients are split into two groups, the treatment and the control group, by means of a coin flip that assigns a value of  $C$  to every patient.<sup>6</sup> Patients are treated depending on the assigned value of  $C$ . Some time after treatment, the outcome  $X$  is measured for each patient. This yields a dataset  $(C_n, X_n)_{n=1}^N$  with two measurements  $(C_n, X_n)$  for the  $n^{\text{th}}$  patient. If the distribution of outcome  $X$  significantly differs between the two groups, one concludes that treatment is a cause of outcome.

The important underlying causal assumptions that ensure the validity of the conclusion are:

- (i) outcome  $X$  is not a potential cause of treatment  $C$  (which is commonly deemed justified if the outcome is an event that occurs later in time than the treatment event);

---

6. Usually this is done in a double-blind way, so that neither the patient nor the doctor knows which group a patient has been assigned to.

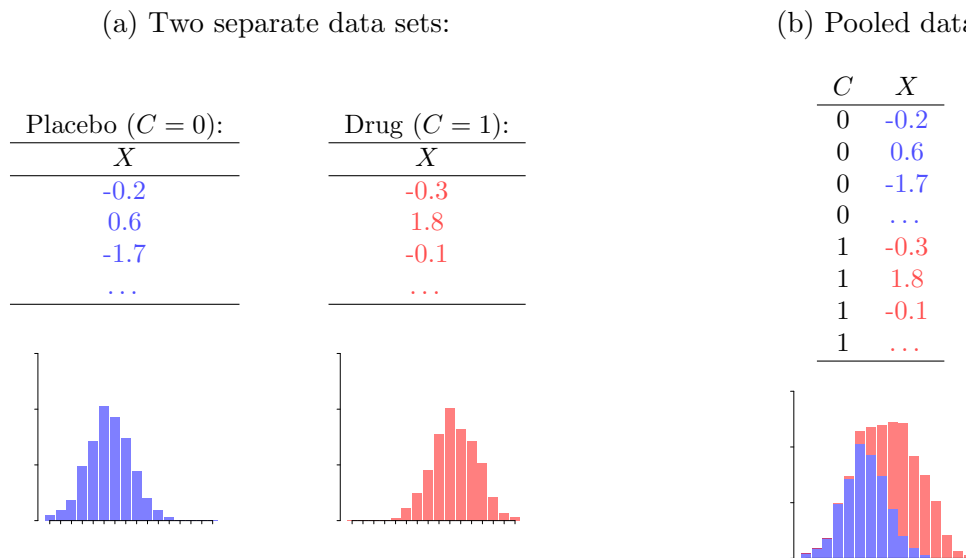


Figure 3: Illustration of the data from an example randomized controlled trial. The data can either be interpreted as (a) two separate data sets, one for the treatment and one for the control group, or (b) as a single data set including a context variable indicating treatment/control. Note that in this particular example,  $C$  is dependent on  $X$  in the pooled data (or equivalently, the distribution of  $X$  differs between contexts  $C = 0$  and  $C = 1$ ), which implies that  $C$  is a cause of  $X$ .

- (ii) there is no potential irreducible latent common cause of treatment and outcome (this is where the randomization comes in: if treatment is decided solely by a proper coin flip, then it seems reasonable to assume that there cannot be any latent common cause of the coin flip  $C$  and the outcome  $X$  that is not just a combination of two statistically independent separate causes of  $C$  and  $X$ ),
- (iii) no selection bias is present in the data (in other words, data samples have not been selected afterwards based on the measured values of  $C$  or  $X$ , which could happen for example if somebody would have selected only a subset of patients that suffered from certain treatment side effects).

Under these assumptions, one can show that if the distribution of the outcome  $X$  differs between the two groups of patients (“treatment group” with  $C = 1$  vs. “control group” with  $C = 0$ ), then treatment must be a cause of outcome (in this population of patients). There are two conceptually slightly different ways of testing this in the data, depending on whether we treat the data as a single pooled dataset, or rather as two separate datasets (each one corresponding to a particular patient group), see also Figure 3. If we consider the data about outcome  $X$  in the two groups as two *separate* data sets (corresponding to the same variable  $X$ , but measured in different contexts  $C$ ), then the question is whether the distribution of  $X$  is statistically different in the two data sets. This can be tested with a two-

sample test, for example, a  $t$ -test or a Wilcoxon test. The other alternative is to consider the data as a single *pooled* data set (by pooling the data for the two groups), and let the value of  $C$  indicate the context of each sample (treatment or control). The question now becomes whether the conditional distribution of  $X$  given  $C = 0$  differs from the conditional distribution of  $X$  given  $C = 1$ , i.e., whether  $\mathbb{P}(X | C = 0) \neq \mathbb{P}(X | C = 1)$ . In other words, we have to test whether there is a statistically significant *dependence*  $C \not\perp X$  in the pooled data between treatment  $C$  and outcome  $X$ ; if there is, it must be due to the treatment  $C$  causing the outcome  $X$ , as the following proposition shows:

**Proposition 9** *Suppose that the data-generating process on context variable  $C$  and outcome variable  $X$  can be modeled by a simple SCM and no selection bias is present. Under the randomized controlled trial assumptions:*

- (i)  $C \leftarrow X \notin \mathcal{G}$  (“outcome  $X$  is not a potential cause of treatment  $C$ ”)
- (ii)  $C \leftrightarrow X \notin \mathcal{G}$  (“there is no potential irreducible latent common cause of treatment  $C$  and outcome  $X$ ”),

*a dependence  $C \not\perp X$  in the joint distribution  $\mathbb{P}(C, X)$  implies that  $C$  causes  $X$ , and furthermore,  $C$  and  $X$  are unconfounded, i.e., the causal effect of  $C$  on  $X$  is given by:*

$$\mathbb{P}(X | \text{do}(C = c)) = \mathbb{P}(X | C = c). \quad (1)$$

**Proof** By the Markov property, if the edge  $C \rightarrow X$  were absent in  $\mathcal{G}(\mathcal{M})$ , then  $C$  would be independent of  $X$ . Therefore, if  $C \not\perp X$ , the edge  $C \rightarrow X$  must be in  $\mathcal{G}(\mathcal{M})$ . In both cases, the causal do-calculus applied to  $\mathcal{G}(\mathcal{M})$  yields the identity (1). If  $C \not\perp X$ , then there must be two values  $c, c'$  such that  $\mathbb{P}(X | C = c) \neq \mathbb{P}(X | C = c')$ , and hence, because of (1),  $\mathbb{P}(X | \text{do}(C = c)) \neq \mathbb{P}(X | \text{do}(C = c'))$ , i.e.,  $C$  is a cause of  $X$ . ■

Of course, in this straightforward example the equivalence between the two approaches (differences between two separate data sets vs. properties of a single pooled data set) is trivial, and the reader may wonder why we emphasize it. The reason is that the key idea of our approach is precisely this: reducing an apparently complicated causal discovery problem with multiple datasets to a more standard causal discovery problem involving a single pooled data set. The Joint Causal Inference framework that we propose in this paper can be considered as a straightforward extension of this randomized controlled trial setting to multiple treatment and outcome variables.

It is important to realize that the simple causal reasoning for the RCT *cannot* be made when looking at the two data sets in isolation (i.e., by considering only properties of  $\mathbb{P}(X | C = 0)$  and  $\mathbb{P}(X | C = 1)$  separately, and not using in addition any other properties of the joint distribution  $\mathbb{P}(X, C)$ ). The latter approach is commonly used by constraint-based methods for causal discovery from multiple data sets (e.g., Tillman, 2009; Claassen and Heskens, 2010; Tillman and Spirtes, 2011; Hyttinen et al., 2014; Triantafillou and Tsamardinos, 2015; Rothenhäusler et al., 2015; Forré and Mooij, 2018). Under the assumptions made, the crucial (and possibly very strong) signal in the data that allows one to draw the conclusion that  $C$  causes  $X$  is the dependence  $C \not\perp X$  in the pooled data. Methods that only test for conditional independences *within* each context and subsequently combine these into a single context-independent causal model will not yield any conclusion in this setting. The

approach taken by JCI, on the other hand, is to analyze the pooled data jointly, so that informative signals like these can be taken into account.

### 2.3 Causal Discovery from Purely Observational Data

In the previous section, we discussed the current ground truth for discovering causal relations. Over the last two decades, alternative methods have been proposed to perform causal discovery from *purely observational* data. This is intriguing and potentially of high relevance, since experiments may be impossible, unfeasible, impractical, unethical or too expensive to perform. These causal discovery methods can be divided into *constraint-based* causal discovery methods, such as the PC (Spirtes et al., 2000), IC (Pearl, 2009) and FCI algorithms (Spirtes et al., 1995; Zhang, 2008), and *score-based* causal discovery methods (e.g., Heckerman et al., 1995; Chickering, 2002; Koivisto and Sood, 2004). The PC and IC algorithms and most score-based methods assume causal sufficiency (i.e., the absence of potential irreducible latent common causes), while the FCI algorithm and other modern constraint-based algorithms allow for latent confounders and selection bias. Originally, these methods have been designed to estimate the causal graph of the system from a single dataset corresponding to a single (purely observational) context.

All these methods try to infer causal relationships on the basis of subtle statistical patterns in the data. The most important of these patterns are (conditional) independences between variables. These are exploited by most constraint-based methods, and implicitly, by score-based methods. Other patterns, such as “Verma constraints” (Shpitser et al., 2014), algebraic constraints in the linear-Gaussian case (van Ommen and Mooij, 2017), non-Gaussianity in linear models (Kano and Shimizu, 2003), and non-additivity of noise in nonlinear models (Peters et al., 2014) can also be exploited. Another class of methods that has become popular more recently are methods that try to infer the causal direction ( $A \rightarrow B$  vs.  $B \rightarrow A$ ) from purely observational data of variable pairs (see e.g., Mooij et al., 2016). Since our main goal is to enable constraint-based causal discovery from multiple contexts, we will focus on this approach here, while noting that the JCI framework that we propose in the next section is compatible with all approaches to causal discovery from purely observational data that allow for multiple variables and can handle certain background knowledge.

As discussed in detail by Spirtes et al. (2000), causal discovery from conditional independence patterns in purely observational data becomes possible under strong assumptions. The simplest example of how certain patterns of conditional independences in the observational distribution can lead to conclusions about the causal relations of the variables is given by the “Y-structure” pattern (Mani, 2006), which is illustrated in Figure 4. We show here that the Y-structure pattern also generalizes to the cyclic case.

**Proposition 10** *Suppose that the data-generating process on four variables  $X_1, X_2, X_3, X_4$  can be modeled by a simple SCM. Assume that the sampling procedure is not subject to selection bias, and that faithfulness holds. If the following conditional (in)dependences hold in the observational distribution:*

$$\begin{aligned} X_1 \not\perp\!\!\!\perp X_4, & \quad X_2 \not\perp\!\!\!\perp X_4, & \quad X_1 \perp\!\!\!\perp X_2, \\ X_1 \perp\!\!\!\perp X_4 | X_3, & \quad X_2 \perp\!\!\!\perp X_4 | X_3, & \quad X_1 \not\perp\!\!\!\perp X_2 | X_3, \end{aligned}$$

*then  $X_3$  is a potential direct cause of  $X_4$  with respect to  $\{X_1, X_2, X_3, X_4\}$ . Furthermore,  $X_3$  and  $X_4$  are unconfounded with respect to  $\{X_3, X_4\}$ , i.e., the causal effect of  $X_3$  on  $X_4$*

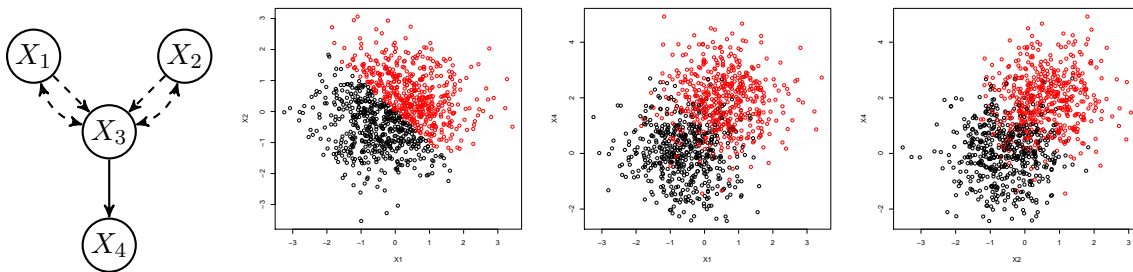


Figure 4: Left: Causal graphs satisfying the “Y-structure” pattern on four variables  $(X_1, X_2, X_3, X_4)$ . Right: Scatter plots illustrating the Y-structure pattern in purely observational data, where  $X_3$  is discrete-valued and its value is indicated by color (black, red); left:  $X_1$  vs.  $X_2$ , center:  $X_1$  vs.  $X_4$ , right:  $X_2$  vs.  $X_4$ .

is given by:

$$\mathbb{P}(X_4 | \text{do}(X_3 = x_3)) = \mathbb{P}(X_4 | X_3 = x_3). \quad (2)$$

**Proof** By the assumed Markovness and faithfulness properties, one can easily see that the only possible functional graphs that are compatible with the observed conditional independences are the ones in Figure 4, where  $X_1$  must be adjacent to  $X_3$  via at least one of the two dashed edges, and similarly,  $X_2$  must be adjacent to  $X_3$  via at least one of the two dashed edges. Hence,  $X_3$  is a potential direct cause of  $X_4$  with respect to  $\{X_1, X_2, X_3, X_4\}$ , but  $X_4$  is not a potential direct cause of  $X_3$  with respect to  $\{X_1, X_2, X_3, X_4\}$ . Also, there cannot be a bidirected edge between  $X_3$  and  $X_4$ . By applying the causal do-calculus, we arrive at (2). ■

This example illustrates how conditional independence patterns in the observational distribution allow one to infer certain features of the underlying causal model. This principle is exploited more generally by constraint-based methods, and implicitly, by score-based methods that optimize a penalized likelihood over causal graphs.

Typically, the functional graph cannot be completely identified from purely observational data. For example, in the Y-structure case, the conditional independences in the observational data do not allow to conclude whether the dependence between  $X_1$  and  $X_3$  is explained by  $X_1$  being a potential cause of  $X_3$ , or by  $X_1$  and  $X_3$  having an irreducible latent common cause, or both. Another disadvantage of causal discovery methods from purely observational data is that they typically need very large sample sizes and strong assumptions in order to work reliably. These are some of the motivations to combine these ideas with those of causal discovery by experimentation, as we will do in the next section.

### 3. Joint Causal Inference

In this section we introduce Joint Causal Inference (JCI), a framework for causal discovery from multiple data sets corresponding to measurements that have been performed in different contexts, that combines the main ideas presented in Sections 2.2 and 2.3.

### 3.1 Context Variables

Henceforth, we will distinguish *system variables*  $(X_i)_{i \in \mathcal{I}}$  describing the system of interest, and *context variables*  $(C_k)_{k \in \mathcal{K}}$  describing the context in which the system has been observed. The decision of what to consider part of the “system” and what to consider part of its “context” does not reflect an objective property of nature, but is a choice of the modeler. While the system variables are treated as *endogenous* variables of the system of interest, we usually (but not necessarily) think of the context variables as observed *exogenous* variables for the system of interest. In particular, context variables could describe which interventions have been performed on the system (or more specifically, how these interventions have been performed), in which case we will also refer to them as *intervention variables*. The possible interventions are not limited to the perfect (“surgical”) interventions modeled by the do-operator of Pearl (2009), but can also be more general types of interventions that appear in practice, like mechanism changes (Tian and Pearl, 2001), soft interventions (Markowitz et al., 2005), fat-hand interventions (Eaton and Murphy, 2007), activity interventions (Mooij and Heskes, 2013), and stochastic versions of all these. This will be discussed in more detail in Section 3.3. Even more generally, a context variable could describe any property of the environment of the system, including those properties that one would not normally think about as an intervention. Examples are the lab in which measurements have been done, the time of the day, the patient population, variables like “gender” or “age”, etc.

The idea of explicitly considering context variables is not novel: they have been discussed in the literature under various names, such as the “force variables” introduced by Pearl (1993), “decision variables” in influence diagrams (Dawid, 2002), “regime indicators” in Didelez et al. (2006), “selection variables” in selection diagrams (Bareinboim and Pearl, 2013), and the “environment variable” in Peters et al. (2016). Some formal aspects in how these variables are treated vary across accounts, however. For example, Dawid (2002) treats system variables as random variables and is careful to not treat context (“decision”) variables as random variables. In this work we simply consider context variables as random variables with added background knowledge on their causal relations.

Context variables may appear to be a more general concept than intervention variables, since every intervention can be seen as a *change of* context, but not every change of context is naturally thought of as an intervention. For example, the causal effect of some drug on a certain health outcome may differ for males and females. Taking “gender” as a context variable that just encodes the specific subpopulation of patients we are considering would be more natural than considering it to be an intervention variable that encodes the result of a potentially gender-changing operation on the patient. Furthermore, interventions usually come with an “observational baseline” of “doing nothing”, but this is not always naturally available. On the other hand, for context variables, no baseline is necessary. In the following, we do not need to commit to a particular interpretation, as the two interpretations can be treated equally from a mathematical modeling perspective.

That being said, the basic idea of JCI is simple: rather than considering a causal model of the system alone (i.e., modeling only the endogenous system variables), we broaden its scope to include relevant parts of the environment of the system (i.e., we include the context variables as additional endogenous variables). Thereby, we “internalize” parts of the environment of the system, which makes the meta-system (consisting of both system and



its environment) amenable to formal causal modeling. The meta-system can now formally be considered as occurring in just a single (meta)-context, and thereby we have reduced the problem of how to deal with multiple contexts to one of dealing with a single context only. We will formalise this idea in the next subsection.

### 3.2 Joint Causal Modeling of Multiple Contexts

Different approaches to modeling multiple contexts can be taken, e.g., using influence diagrams (Dawid, 2002), using selection diagrams (Bareinboim and Pearl, 2013), considering only conditional models (i.e., for the conditional likelihood of the data given the context) (Eaton and Murphy, 2007; Mooij and Heskes, 2013), or using ioSCMs (Forré and Mooij, 2019). Here, we will take what is perhaps the simplest approach: we treat both context and system variables as endogenous variables in an SCM.

We will use a simple SCM to model the meta-system (i.e., the system and its contexts) causally. The endogenous variables of the SCM consist of the system variables  $\mathbf{X} = (X_i)_{i \in \mathcal{I}}$  with values  $\mathbf{x} \in \mathcal{X} = \prod_{i \in \mathcal{I}} \mathcal{X}_i$  and the context variables  $\mathbf{C} = (C_k)_{k \in \mathcal{K}}$  with values  $\mathbf{c} \in \mathcal{C} = \prod_{k \in \mathcal{K}} \mathcal{C}_k$ . The latent exogenous variables of the SCM are denoted  $\mathbf{E} = (E_j)_{j \in \mathcal{J}}$  with values  $\mathbf{e} \in \mathcal{E} = \prod_{j \in \mathcal{J}} \mathcal{E}_j$ . The SCM modeling the meta-system is then assumed to be of the following form:

$$\mathcal{M} : \begin{cases} C_k & = f_k(\mathbf{X}_{\text{PA}_{\mathcal{H}}(k) \cap \mathcal{I}}, \mathbf{C}_{\text{PA}_{\mathcal{H}}(k) \cap \mathcal{K}}, \mathbf{E}_{\text{PA}_{\mathcal{H}}(k) \cap \mathcal{J}}), & k \in \mathcal{K}, \\ X_i & = f_i(\mathbf{X}_{\text{PA}_{\mathcal{H}}(i) \cap \mathcal{I}}, \mathbf{C}_{\text{PA}_{\mathcal{H}}(i) \cap \mathcal{K}}, \mathbf{E}_{\text{PA}_{\mathcal{H}}(i) \cap \mathcal{J}}), & i \in \mathcal{I}, \\ \mathbb{P}(\mathbf{E}) & = \prod_{j \in \mathcal{J}} \mathbb{P}(E_j). \end{cases} \quad (3)$$

The system variables  $\mathbf{X}$  and context variables  $\mathbf{C}$  are all treated as endogenous variables of the system, and the exogenous variables  $\mathbf{E}$  are independent latent variables that are assumed not to be potentially caused by the system variables  $\mathbf{X}$  or the context variables  $\mathbf{C}$ . The augmented functional graph  $\mathcal{H}$  has nodes  $\mathcal{I} \cup \mathcal{J} \cup \mathcal{K}$  and directed edges corresponding to the functional dependencies of the causal mechanisms on the variables. The functional graph  $\mathcal{G}(\mathcal{M})$  has only nodes  $\mathcal{I} \cup \mathcal{K}$ , and may contain both directed and bidirected edges between the nodes, expressing potential direct causal relations and potential irreducible latent common causes.

Note that the most general way to use SCMs to model multiple contexts would be to use separate SCMs, one for each context. In that approach, we could have a different functional graph for each context. Representing the contexts jointly, as in (3), we simply obtain the union of those functional graphs. In particular, even if within each context, the system is acyclic, it could be that the mixture of systems in different contexts has a cyclic functional graph. As a simple example, consider a system with two system variables  $X_1$  and  $X_2$ , and consider two different contexts, where in the first context  $X_1$  causes  $X_2$  (but not vice versa), and in the second context,  $X_2$  causes  $X_1$  (but not vice versa); see also Figure 5. As a more concrete example, the engine drives the wheels of a car when going uphill, but when going downhill, the rotation of the wheels drives the engine. Modeling this in a joint SCM as in (3) requires a cyclic functional graph.

The model (3) imposes a probability distribution  $\mathbb{P}(\mathbf{C})$  on the context variables, the *context distribution*. The context distribution will reflect the empirical distribution of the context variables in the *pooled* data, by using as the probability of a context the fraction of

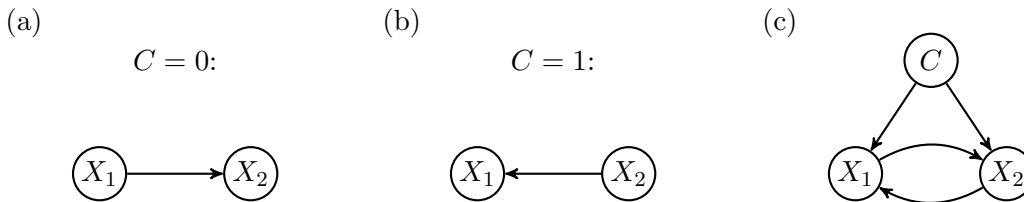


Figure 5: The functional graph of a mixture of two acyclic SCMs can be cyclic. (a)  $X_1$  causes  $X_2$  in context  $C = 0$ ; (b)  $X_2$  causes  $X_1$  in context  $C = 1$ ; (c)  $X_1$  and  $X_2$  cause each other in the joint model.

the total number of samples that have been measured in that context. One might object that this makes the model very specific to the particular setting, since it also specifies the relative numbers of samples in each dataset, but as it turns out, the conclusions of the causal discovery procedure do not depend on these details under reasonable assumptions, and therefore generalize to other context distributions.

Because the context variables are treated as endogenous variables (similarly to the system variables), we have “internalized” them. The main advantage of our modeling approach over alternative approaches is that in (3), context variables are formally treated in exactly the same way as the system variables. This implies in particular that all standard definitions and terminology of Section 2.1, and all causal discovery methods that are applicable in that setting, can be directly applied.

### 3.3 Modeling Interventions as Context Changes

The causal model in (3) allows one to model a perfect (“surgical”) intervention in the usual way (Pearl, 2009). Indeed, the perfect intervention that forces  $\mathbf{X}_I$  to take on the value  $\boldsymbol{\xi}_I$  (“do( $\mathbf{X}_I = \boldsymbol{\xi}_I$ )”) for some subset  $I \subseteq \mathcal{I}$  and some value  $\boldsymbol{\xi}_I \in \prod_{i \in I} \mathcal{X}_i$  can be modeled by replacing the structural equations for the system variables in (3) by:

$$X_i = \begin{cases} \xi_i & i \in I \\ f_i(\mathbf{X}_{\text{PA}(i) \cap \mathcal{I}}, \mathbf{C}_{\text{PA}(i) \cap \mathcal{K}}, \mathbf{E}_{\text{PA}(i) \cap \mathcal{J}}) & i \in \mathcal{I} \setminus I, \end{cases}$$

while leaving the rest of the model invariant.<sup>7</sup>

Alternatively, the context variables can be used to model interventions. For example, the same perfect intervention could be modeled by introducing a context variable  $C_k$  that has  $\text{CH}(k) = I$  and domain  $\mathcal{C}_k = \{\emptyset\} \cup \prod_{i \in I} \mathcal{X}_i$ , by taking  $\mathbf{f}_I$  to be of the following form:

$$f_i(\mathbf{X}_{\text{PA}(i) \cap \mathcal{I}}, \mathbf{C}_{\text{PA}(i) \cap \mathcal{K}}, \mathbf{E}_{\text{PA}(i) \cap \mathcal{J}}) = \begin{cases} \tilde{f}_i(\mathbf{X}_{\text{PA}(i) \cap \mathcal{I}}, \mathbf{C}_{\text{PA}(i) \cap \mathcal{K} \setminus \{k\}}, \mathbf{E}_{\text{PA}(i) \cap \mathcal{J}}) & C_k = \emptyset \\ (C_k)_i & C_k \in \prod_{i \in I} \mathcal{X}_i \end{cases} \quad (4)$$

for  $i \in I$ . Modeling a perfect intervention in this way is similar to the concept of “force variables” introduced by Pearl (1993). If  $k$  has no parents and no spouses, the observational

7. For brevity, we dropped the subscript  $\mathcal{H}$  of  $\text{PA}_{\mathcal{H}}(\cdot)$ .

distribution (without the intervention) is given by the conditional distribution  $\mathbb{P}(\mathbf{X} | C_k = \emptyset)$ , the interventional distribution corresponding to the perfect intervention  $\text{do}(\mathbf{X}_I = \boldsymbol{\xi}_I)$  is given by the conditional distribution  $\mathbb{P}(\mathbf{X} | C_k = \boldsymbol{\xi}_I)$ , and the marginal distribution  $\mathbb{P}(\mathbf{X})$  represents a mixture of those. This is illustrated in Figure 6.

More general types of interventions such as mechanism changes (Tian and Pearl, 2001) can be modeled in a similar way, simply by not enforcing the dependence on  $C_k$  to be of the form (4), but allowing more general forms of functional dependence. For example, switching the causal mechanism of system variable  $X_i$  from mechanism  $A$  to mechanism  $B$  can be modeled as follows by introducing a context variable  $C_k$  with  $\text{CH}(k) = \{i\}$  and domain  $\mathcal{C}_k = \{A, B\}$ :

$$f_i(\mathbf{X}_{\text{PA}(i) \cap \mathcal{I}}, \mathbf{C}_{\text{PA}(i) \cap \mathcal{K}}, \mathbf{E}_{\text{PA}(i) \cap \mathcal{J}}) = \begin{cases} \tilde{f}_i^A(\mathbf{X}_{\text{PA}(i) \cap \mathcal{I}}, \mathbf{C}_{\text{PA}(i) \cap \mathcal{K} \setminus \{k\}}, \mathbf{E}_{\text{PA}(i) \cap \mathcal{J}}) & C_k = A \\ \tilde{f}_i^B(\mathbf{X}_{\text{PA}(i) \cap \mathcal{I}}, \mathbf{C}_{\text{PA}(i) \cap \mathcal{K} \setminus \{k\}}, \mathbf{E}_{\text{PA}(i) \cap \mathcal{J}}) & C_k = B. \end{cases} \quad (5)$$

As another example, a stochastic surgical intervention on  $X_i$  that is only successful with a certain probability can be modeled by having one of the latent exogenous variables  $E_j$  with  $j \in \text{PA}(i)$  determine whether the intervention was successful:

$$f_i(\mathbf{X}_{\text{PA}(i) \cap \mathcal{I}}, \mathbf{C}_{\text{PA}(i) \cap \mathcal{K}}, \mathbf{E}_{\text{PA}(i) \cap \mathcal{J}}) = \begin{cases} \tilde{f}_i(\mathbf{X}_{\text{PA}(i) \cap \mathcal{I}}, \mathbf{C}_{\text{PA}(i) \cap \mathcal{K} \setminus \{k\}}, \mathbf{E}_{\text{PA}(i) \cap \mathcal{J} \setminus \{j\}}) & C_k = \emptyset \text{ or } E_j = 0 \\ (C_k)_i & C_k \in \prod_{i \in \mathcal{I}} \mathcal{X}_i \text{ and } E_j = 1. \end{cases} \quad (6)$$

This approach of modeling interventions by means of context variables is very general, as it allows to treat various types of interventions in a unified way: it can deal with perfect interventions (Pearl, 2009), mechanism changes (Tian and Pearl, 2001), soft interventions (Markowitz et al., 2005), fat-hand interventions (Eaton and Murphy, 2007), activity interventions (Mooij and Heskes, 2013), and stochastic versions of all these. In case the context variables are used to model interventions in this way, we also refer to those as *intervention variables*, and to the context distribution  $\mathbb{P}(\mathbf{C})$ , the probability for each context to occur, as the *experimental design*.

### 3.4 JCI Assumptions

In this subsection, we discuss additional background knowledge on the causal relationships of context variables that one may often have in practice, and that can be very helpful for causal discovery.

#### 3.4.1 JCI ASSUMPTION 0

First, we remind the reader of our basic modeling assumption:

**Assumption 0** (“Joint SCM”) *The data-generating mechanism is described by a simple SCM  $\mathcal{M}$  of the form:*

$$\mathcal{M} : \begin{cases} C_k & = f_k(\mathbf{X}_{\text{PA}_{\mathcal{H}}(k) \cap \mathcal{I}}, \mathbf{C}_{\text{PA}_{\mathcal{H}}(k) \cap \mathcal{K}}, \mathbf{E}_{\text{PA}_{\mathcal{H}}(k) \cap \mathcal{J}}), & k \in \mathcal{K}, \\ X_i & = f_i(\mathbf{X}_{\text{PA}_{\mathcal{H}}(i) \cap \mathcal{I}}, \mathbf{C}_{\text{PA}_{\mathcal{H}}(i) \cap \mathcal{K}}, \mathbf{E}_{\text{PA}_{\mathcal{H}}(i) \cap \mathcal{J}}), & i \in \mathcal{I}, \\ \mathbb{P}(\mathbf{E}) & = \prod_{j \in \mathcal{J}} \mathbb{P}(E_j), \end{cases} \quad (7)$$

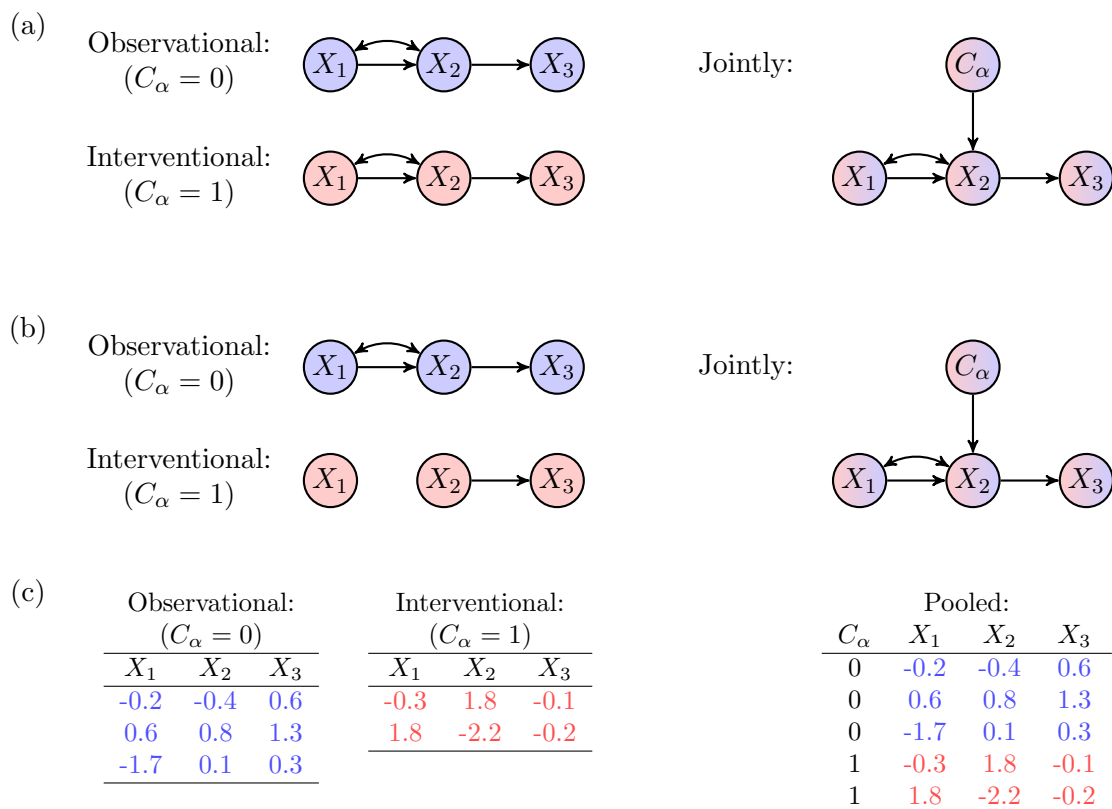


Figure 6: (a–b) Two ways of representing interventions, either through separate context-dependent causal graphs of the system (left), or through the joint causal graph of system and context (right). Two different intervention types are considered: (a) a mechanism change on  $X_2$ ; (b) a perfect (surgical) intervention on  $X_2$ . (c) shows different ways of grouping the data: as separate datasets for each context (left), or as a single joint data set after pooling (right).

that jointly models the context distribution  $\mathbb{P}(\mathbf{C})$  and the system in all contexts. Its functional graph  $\mathcal{G}(\mathcal{M})$  has nodes  $\mathcal{I} \cup \mathcal{K}$  (corresponding to system variables  $\{X_i\}_{i \in \mathcal{I}}$  and context variables  $\{C_k\}_{k \in \mathcal{K}}$ ).

Whereas we will always make this assumption in order to facilitate the formulation of JCI, the following three assumptions that we discuss are optional, and their applicability has to be decided based on a case-by-case basis.

### 3.4.2 JCI ASSUMPTION 1

Typically, when a modeler decides to distinguish a *system* from its *context*, the modeler possesses background knowledge that expresses that the context is *exogenous* to the system:

**Assumption 1** (“Exogeneity”, optional) *No system variable is a potential cause of any context variable (directly with respect to  $\mathcal{I} \cup \mathcal{K}$ ), i.e.,*

$$\forall k \in \mathcal{K}, \forall i \in \mathcal{I}: \quad i \rightarrow k \notin \mathcal{G}(\mathcal{M}).$$

This exogeneity assumption is often easy to justify, for example if context is gender or age. Another common case is that the context encodes interventions on the system that have been decided and performed on the system *before* measurements on the system are performed: this already rules out any potential causal influence of system variables on the intervention (context) variables if time travel is not deemed possible. Of course, one can imagine settings in which a system variable describes an event that precedes an intervention event that is performed on the system described by an intervention variable. For example, a doctor typically first diagnoses a patient *before* deciding on treatment. Thus, if some of the system variables are measurements that are performed as part of the medical examination used for the diagnosis, and the intervention is the treatment that was decided *after*—and based upon—these measurements, this assumption may not be applicable.

### 3.4.3 JCI ASSUMPTION 2

The second JCI assumption generalizes the randomization assumption for Randomized Controlled Trials:

**Assumption 2** (“Complete randomized context”, optional) *There is no potential irreducible latent common cause of the system variables and the context variables, i.e.,*

$$\forall k \in \mathcal{K}, \forall i \in \mathcal{I}: \quad i \leftrightarrow k \notin \mathcal{G}(\mathcal{M}).$$

This assumption is often harder to justify in practice. It is justifiable in experimental protocols in which the decision of which intervention to perform on the system does not depend on anything else that might also affect the system of interest, and in which the observed context variables provide a complete description of the context. This is ensured for example in case of proper randomization in a double-blind randomized trial setting, i.e., in which neither the patient nor the physician know whether the patient was assigned a drug or a placebo.

Many experimental protocols that do not involve explicit coin flips or random number generators are implicitly performing randomization. For example, in the experimental

procedure described by Sachs et al. (2005), one starts with a collection of human immune system cells. These are divided into batches randomly, without taking into account any property of the cells. When done carefully, the experimenter tries to ensure that for example the weight of a cell cannot influence the batch it ends up in, by stirring the liquid that contains the cells before pipetting. Then, after randomly assigning cells to batches, interventions are performed on each batch separately, by adding some chemical compound to the batch of cells. Finally, properties of each individual cell within each batch are measured. If we take the system variables to reflect the measured properties of the individual cells, and the context variables to encode the batch ID, this experimental procedure justifies JCI Assumption 2.

However, one should be careful not to jump to the conclusion that the chemical compound administered to the batch is what actually causes the observed system behavior, as there may be other factors that vary across batches due to unintentional side effects of the experimental procedure. For example, which person carries out the experiment for a particular batch of cells might influence the outcome, because slightly different experimental procedures are used by different lab assistants. Another example is that the time of the day may affect the measurements, and also correlate with batch ID. In situations like those, *identifying* the batch ID with the chemical compound administered to that batch could be misleading, and could lead one to incorrectly attribute the inferred causal relation between batch ID and a certain system variable to the causal effect of the intended intervention corresponding to that batch on the system variable. This is a subtle type of error that the causal modeler should beware of. Even though we have good reasons to assume that proper randomization was performed for batch ID in the Sachs et al. (2005) experiment, it is questionable whether the interpretation of the context variables as concerning solely the addition of certain chemical compounds (and not any other factors that actually varied across batches) is appropriate.

The issue can also be understood by noting that JCI Assumption 2 may not be preserved when marginalizing out context variables, as illustrated in Figure 7. As an example in which the situation in Figure 7 may occur, consider a randomized trial setup for establishing whether sugar causes plants to grow. Context variable  $C_\alpha$  denotes the coin flip result,  $C_\beta$  indicates whether sugar is administered to the plant, and  $C_\gamma$  indicates whether water is administered to the plant. The experimenter decided to use an experimental design with two groups, and assigning plants to groups with a coin flip. One group of plants was administered a solution consisting of sugar dissolved in water on a daily basis, the other (control) group was not treated in any way. The growth rate  $X_1$  of the plants was measured for both groups. Suppose the following experimental design was used:

$\mathbb{P}(\mathbf{C} = \mathbf{c})$	$C_\alpha$ (coin flip)	$C_\beta$ (sugar)	$C_\gamma$ (water)
$\frac{1}{2}$	0	0	0
$\frac{1}{2}$	1	1	1

If one would only take context variable  $C_\beta$  (did the plant get sugar?) into account and would treat  $C_\alpha$  and  $C_\gamma$  as latent, as in Figure 7(b), and would make JCI Assumptions 1 and 2, one would arrive at the (wrong) conclusion that sugar causes plants to grow. However, if one would take all three context variables into account, and make JCI Assumptions 1 and 2, one would obtain the right conclusion that at least one of the three context variables

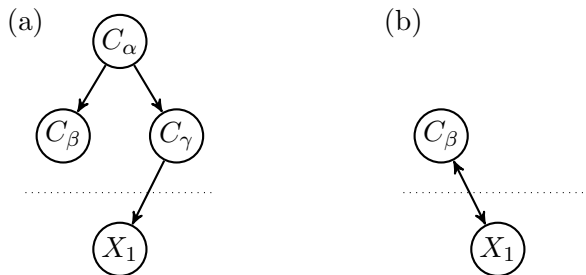


Figure 7: Confounding between system and context variables due to unobserved context variables. (a) If all three context variables  $C_\alpha, C_\beta, C_\gamma$  are observed, JCI Assumption 2 would be valid. (b) After marginalizing out  $C_\alpha$  and  $C_\gamma$ , leaving only context variable  $C_\beta$  as observed, JCI Assumption 2 is no longer valid.

must cause plants to grow. A simple remedy to avoid the wrong conclusion if only  $C_\beta$  is observed would be to drop JCI Assumption 2: then it is no longer identifiable whether  $C_\beta$  causes  $X_1$ , or whether  $C_\beta$  and  $X_1$  are just confounded.

As a more realistic example, consider the *placebo effect* in a randomized controlled trial that is not performed double-blindly. This is the same setting as in Figure 7, but with different meanings of the variables. Context variable  $C_\alpha$  would still encode the result of the coin flip that indicates whether a patient is assigned to the treatment or control group,  $C_\beta$  indicates taking the drug, and  $C_\gamma$  indicates being under treatment by a medical professional. In this scenario, the drug itself has no effect on health outcome  $X_1$ . If the RCT is not performed double-blindly, it could be that there is a placebo effect, i.e.,  $C_\gamma$  could cause  $X_1$ . JCI Assumption 2 would be valid if all three context variables are observed and taken into account (Figure 7(a)), but if only context variable  $C_\beta$  (taking the drug) is considered, then this results in confounding between context node  $C_\beta$  and system node  $X_1$ , even though  $C_\beta$  was decided by a random coin flip and therefore it was properly randomized (Figure 7(b)), in which case JCI Assumption 2 would be invalid.

### 3.4.4 JCI ASSUMPTION 3

In order to discuss the last JCI assumption, we need more notation and theory.

**Definition 11** *Given an SCM  $\mathcal{M}$  satisfying JCI assumption 0, define the influence diagram  $\mathcal{G}(\mathcal{M})_{\mathcal{I}|\text{do}(\mathcal{K})}$  as the (colored) DMG with nodes  $\mathcal{I} \cup \mathcal{K}$ , and as directed and bidirected edges a subset of those in  $\mathcal{G}(\mathcal{M})$ , namely precisely those that involve one or more system nodes in  $\mathcal{I}$ . We will graphically represent the system nodes  $\mathcal{I}$  of  $\mathcal{G}(\mathcal{M})_{\mathcal{I}|\text{do}(\mathcal{K})}$  by circles and the context nodes  $\mathcal{K}$  of  $\mathcal{G}(\mathcal{M})_{\mathcal{I}|\text{do}(\mathcal{K})}$  by squares.*

We have seen that JCI Assumption 1 is often easily justifiable, but the applicability of JCI Assumption 2 may be less clear in practice. If both assumptions are made, then one can apply the following result, which basically says that when one is only interested in modeling the causal relations involving the system variables, one does not need to care about the causal relations between the context variables, as long as the right context distribution is induced.

**Theorem 12** *Assume that JCI Assumptions 0, 1 and 2 hold for SCM  $\mathcal{M}$ :*

$$\mathcal{M} : \begin{cases} C_k &= f_k(\mathbf{C}_{\text{PA}_{\mathcal{H}}(k) \cap \mathcal{K}}, \mathbf{E}_{\text{PA}_{\mathcal{H}}(k) \cap \mathcal{J}}), & k \in \mathcal{K}, \\ X_i &= f_i(\mathbf{X}_{\text{PA}_{\mathcal{H}}(i) \cap \mathcal{I}}, \mathbf{C}_{\text{PA}_{\mathcal{H}}(i) \cap \mathcal{K}}, \mathbf{E}_{\text{PA}_{\mathcal{H}}(i) \cap \mathcal{J}}), & i \in \mathcal{I}, \\ \mathbb{P}(\mathbf{E}) &= \prod_{j \in \mathcal{J}} \mathbb{P}(E_j), \end{cases}$$

*For any other SCM  $\tilde{\mathcal{M}}$  satisfying JCI Assumptions 0, 1 and 2 that is the same as  $\mathcal{M}$  except that it models the context differently, i.e., of the form*

$$\tilde{\mathcal{M}} : \begin{cases} C_k &= \tilde{f}_k(\mathbf{C}_{\text{PA}_{\tilde{\mathcal{H}}}(k) \cap \mathcal{K}}, \mathbf{E}_{\text{PA}_{\tilde{\mathcal{H}}}(k) \cap \tilde{\mathcal{J}}}), & k \in \mathcal{K}, \\ X_i &= f_i(\mathbf{X}_{\text{PA}_{\mathcal{H}}(i) \cap \mathcal{I}}, \mathbf{C}_{\text{PA}_{\mathcal{H}}(i) \cap \mathcal{K}}, \mathbf{E}_{\text{PA}_{\mathcal{H}}(i) \cap \mathcal{J}}), & i \in \mathcal{I}, \\ \mathbb{P}(\mathbf{E}) &= \prod_{j \in \tilde{\mathcal{J}}} \mathbb{P}(E_j), \end{cases} \quad (8)$$

*with  $\mathcal{J} \subseteq \tilde{\mathcal{J}}$  and  $\text{PA}_{\mathcal{H}}(i) = \text{PA}_{\tilde{\mathcal{H}}}(i)$  for all  $i \in \mathcal{I}$ , we have that*

- (i) the influence diagrams coincide:  $\mathcal{G}(\mathcal{M})_{\mathcal{I} | \text{do}(\mathcal{K})} = \mathcal{G}(\tilde{\mathcal{M}})_{\mathcal{I} | \text{do}(\mathcal{K})}$ ;*
- (ii) if  $\tilde{\mathcal{M}}$  and  $\mathcal{M}$  induce the same context distribution, i.e.,  $\mathbb{P}_{\mathcal{M}}(\mathbf{C}) = \mathbb{P}_{\tilde{\mathcal{M}}}(\mathbf{C})$ , then for any perfect intervention on the system variables  $\text{do}(I, \xi_I)$  with  $I \subseteq \mathcal{I}$  (including the non-intervention  $I = \emptyset$ ),  $\tilde{\mathcal{M}}_{\text{do}(I, \xi_I)}$  is observationally equivalent to  $\mathcal{M}_{\text{do}(I, \xi_I)}$ .*
- (iii) if  $\mathcal{G}(\tilde{\mathcal{M}})_{\mathcal{K}}$  and  $\mathcal{G}(\mathcal{M})_{\mathcal{K}}$  induce the same separations, then also  $\mathcal{G}(\tilde{\mathcal{M}})$  and  $\mathcal{G}(\mathcal{M})$  induce the same separations (where “separations” can refer to either  $d$ -separations or  $\sigma$ -separations).*

**Proof** Let  $\mathcal{M}$  be an SCM of the form (7). Under JCI Assumption 1, the structural equations for the context variables do not depend on the system variables:

$$C_k = f_k(\mathbf{C}_{\text{PA}_{\mathcal{H}}(k) \cap \mathcal{K}}, \mathbf{E}_{\text{PA}_{\mathcal{H}}(k) \cap \mathcal{J}}), \quad k \in \mathcal{K}.$$

Because of JCI Assumption 2,  $\text{PA}_{\mathcal{H}}(\mathcal{K}) \cap \text{PA}_{\mathcal{H}}(\mathcal{I}) \cap \mathcal{J} = \emptyset$ , i.e., the context variables do not share any exogenous variable with the system variables. This means that in  $\mathcal{G}(\mathcal{M})$ , any edge between a context variable and a system variable must be a directed edge pointing from context to system variable, i.e., of the form  $k \rightarrow i$  with  $k \in \mathcal{K}$ ,  $i \in \mathcal{I}$ .

Since the structural equations for the system variables of  $\tilde{\mathcal{M}}$  coincide with those of  $\mathcal{M}$ , their solutions (in terms of the context and exogenous variables) also coincide, even after any perfect intervention on a subset of the system variables. Since  $\mathbf{C}$  is independent of  $\mathbf{E}_{\text{PA}_{\mathcal{H}}(\mathcal{I})}$  (both for  $\mathcal{M}$  as well as for  $\tilde{\mathcal{M}}$ ), and since  $\mathbb{P}_{\mathcal{E}} = \mathbb{P}_{\tilde{\mathcal{E}}_{\mathcal{J}}}$  by assumption, this implies that the interventional distributions of  $\mathcal{M}$  and  $\tilde{\mathcal{M}}$  coincide for any perfect intervention on a subset of system variables if  $\mathbb{P}_{\mathcal{M}}(\mathbf{C}) = \mathbb{P}_{\tilde{\mathcal{M}}}(\mathbf{C})$ .

Assume now that  $\mathcal{G}(\tilde{\mathcal{M}})_{\mathcal{K}}$  and  $\mathcal{G}(\mathcal{M})_{\mathcal{K}}$  induce the same separations. In the remainder of this proof, “open” can be read either consistently as “ $\sigma$ -open” or as “ $d$ -open”. Note that by assumption,  $\mathcal{G}(\mathcal{M})_{\mathcal{I} | \text{do}(\mathcal{K})} = \mathcal{G}(\tilde{\mathcal{M}})_{\mathcal{I} | \text{do}(\mathcal{K})}$ , and that the edges in  $\mathcal{G}(\mathcal{M})_{\mathcal{I} | \text{do}(\mathcal{K})}$  are a subset of those in  $\mathcal{G}(\tilde{\mathcal{M}})$ , and of those in  $\mathcal{G}(\mathcal{M})$ . We will prove that  $\mathcal{G}(\tilde{\mathcal{M}})$  and  $\mathcal{G}(\mathcal{M})$  induce the same separations by first showing that for any two context nodes connected by a path  $\pi$  in  $\mathcal{G}(\mathcal{M})_{\mathcal{K}}$  such that  $\pi$  is  $A$ -open in  $\mathcal{G}(\mathcal{M})$  for some  $A \subseteq \mathcal{I} \cup \mathcal{K}$ , we can find a



path  $\pi'$  in  $\mathcal{G}(\mathcal{M})_{\mathcal{K}}$  between the two nodes that is  $A'$ -open in  $\mathcal{G}(\mathcal{M})$  where  $A' = A \cap \mathcal{K} \cup B$  with  $B \subseteq \mathcal{K} \cap \text{AN}_{\mathcal{G}(\mathcal{M})_{\mathcal{I}|\text{do}(\mathcal{K})}}(A \setminus \mathcal{K})$ . For  $\pi$  to be  $A$ -open in  $\mathcal{G}(\mathcal{M})$ , any collider on  $\pi$  that is not a  $\mathcal{G}(\mathcal{M})$ -ancestor of  $A \cap \mathcal{K}$  must be a  $\mathcal{G}(\mathcal{M})$ -ancestor of  $A \setminus \mathcal{K}$ . Since the latter does not necessarily imply that the collider must also be  $\mathcal{G}(\tilde{\mathcal{M}})$ -ancestor of  $A \setminus \mathcal{K}$ , the idea will be to replace the variables from  $A \setminus \mathcal{K}$  in the conditioning set by variables in  $\mathcal{K} \cap \text{AN}_{\mathcal{G}(\mathcal{M})_{\mathcal{I}|\text{do}(\mathcal{K})}}(A \setminus \mathcal{K})$  (i.e., context nodes that are guaranteed to be both  $\mathcal{G}(\mathcal{M})$ -ancestors and  $\mathcal{G}(\tilde{\mathcal{M}})$ -ancestors of  $A \setminus \mathcal{K}$ ) that are  $\mathcal{G}(\mathcal{M})$ -descendants of those colliders that are not already  $\mathcal{G}(\mathcal{M})$ -ancestors of  $A \cap \mathcal{K}$ . It will turn out that this is not always possible to achieve for  $\pi$ , but that we can construct another path  $\pi'$  for which this can be done.

Consider a path  $\pi$  in  $\mathcal{G}(\mathcal{M})_{\mathcal{K}}$  between  $k_0 \in \mathcal{K}$  and  $k_n \in \mathcal{K}$  that is  $A$ -open in  $\mathcal{G}(\mathcal{M})$  for some  $A \subseteq \mathcal{I} \cup \mathcal{K}$ . We will iteratively construct a walk in  $\mathcal{G}(\mathcal{M})_{\mathcal{K}}$  between the same two nodes  $k_0$  and  $k_n$  that is both  $A$ -open in  $\mathcal{G}(\mathcal{M})$  and  $(A \cap \mathcal{K}) \cup B$ -open in  $\mathcal{G}(\mathcal{M})$ , where  $B \subseteq \mathcal{K} \cap \text{AN}_{\mathcal{G}(\mathcal{M})_{\mathcal{I}|\text{do}(\mathcal{K})}}(A \setminus \mathcal{K})$ . We will proceed by induction. Suppose a walk  $\pi_m$  in  $\mathcal{G}(\mathcal{M})_{\mathcal{K}}$  between  $k_0$  and  $k_n$  is  $A$ -open in  $\mathcal{G}(\mathcal{M})$ . Then it is  $A \cup B_m$ -open in  $\mathcal{G}(\mathcal{M})$  where  $B_m = (\mathcal{K} \cap \text{AN}_{\mathcal{G}(\mathcal{M})_{\mathcal{I}|\text{do}(\mathcal{K})}}(A \setminus \mathcal{K})) \setminus \text{NCOL}(\pi_m)$ . Consider the “problematic” colliders  $\text{COL}(\pi_m) \setminus \text{AN}_{\mathcal{G}(\mathcal{M})}(A \cap \mathcal{K} \cup B_m)$  on  $\pi_m$ , i.e., the ones that are not ancestors of  $A \cap \mathcal{K} \cup B_m$ . If there are any, choose one such problematic collider  $c \in \mathcal{K}$  on  $\pi_m$ . Since  $c$  is not  $\mathcal{G}(\mathcal{M})$ -ancestor of  $A \cap \mathcal{K} \cup B_m$ , but  $\pi_m$  is  $A$ -open, it has to be  $\mathcal{G}(\mathcal{M})$ -ancestor of  $A \setminus \mathcal{K}$ . This means that there is a directed path in  $\mathcal{G}(\mathcal{M})$  that starts at  $c$ , passes through zero or more context nodes, none of which lie in  $A \cap \mathcal{K} \cup B_m$  by assumption, and then through zero or more system nodes, until it ends at a system node in  $A \setminus \mathcal{K}$ . Let  $k_c \in \mathcal{K}$  be the last context node on this directed path before the path crosses the context-system boundary. By assumption,  $k_c$  must exist as a non-collider on  $\pi_m$  (otherwise it would be in  $B_m$  and  $c$  would be  $\mathcal{G}(\mathcal{M})$ -ancestor of  $B_m$ ), hence we can make a shortcut by replacing the subwalk of  $\pi_m$  between  $c$  and  $k_c$  by a directed path  $c \rightarrow \dots \rightarrow k_c$  in  $\mathcal{G}(\mathcal{M})_{\mathcal{K}}$ , which necessarily entirely consists of context nodes that are not in  $A$ . If  $k_c$  occurs more than once on this new walk, remove the entire subwalk between the two outermost occurrences of  $k_c$ , such that  $k_c$  only occurs once. This new walk  $\pi_{m+1}$  must be  $A$ -open:  $c$  (if still present) is now a non-collider that is not in  $A$ , none of the (non-collider) nodes on the directed path (if still present) between  $c$  and  $k_c$  are in  $A$ , and  $k_c$  itself is not in  $A$  and is a  $\mathcal{G}(\mathcal{M})$ -ancestor of  $A$ , so it does not matter whether it is a collider or non-collider. The number of problematic colliders on  $\pi_{m+1}$  is at least one less than on  $\pi_m$ :  $c$  is no longer a collider, and if  $k_c$  became a collider on  $\pi_{m+1}$ , it won't be problematic (as it is itself in  $(\mathcal{K} \cap \text{AN}_{\mathcal{G}(\mathcal{M})_{\mathcal{I}|\text{do}(\mathcal{K})}}(A \setminus \mathcal{K}))$  and cannot also occur as non-collider on  $\pi_{m+1}$ ). We repeat this procedure until no problematic colliders are present anymore. This yields a walk  $\pi_M$  that is both  $A$ -open and  $A'$ -open, with  $A' = (A \cap \mathcal{K}) \cup B$  where  $B = B_M = (\mathcal{K} \cap \text{AN}_{\mathcal{G}(\mathcal{M})_{\mathcal{I}|\text{do}(\mathcal{K})}}(A \setminus \mathcal{K})) \setminus \text{NCOL}(\pi_M)$ . We now shorten this  $A'$ -open walk  $\pi_M$  in  $\mathcal{G}(\mathcal{M})_{\mathcal{K}}$  into an  $A'$ -open path  $\pi'$  in  $\mathcal{G}(\mathcal{M})_{\mathcal{K}}$ .

This implies that there must be an  $A'$ -open path  $\tilde{\pi}'$  in  $\mathcal{G}(\tilde{\mathcal{M}})_{\mathcal{K}}$  connecting  $k_0$  and  $k_n$ , by assumption. Every collider on  $\tilde{\pi}'$  is a  $\mathcal{G}(\tilde{\mathcal{M}})_{\mathcal{K}}$ -ancestor of  $A \cap \mathcal{K} \cup B$ , and hence  $\mathcal{G}(\tilde{\mathcal{M}})$ -ancestor of  $A \cap \mathcal{K} \cup B$ , and hence  $\mathcal{G}(\tilde{\mathcal{M}})$ -ancestor of  $A$ . Therefore,  $\tilde{\pi}'$  is also  $A'$ -open in  $\mathcal{G}(\tilde{\mathcal{M}})$ . But then it must also be  $A$ -open, as we can add  $A \setminus \mathcal{K}$  to the conditioning set without blocking any non-collider on  $\tilde{\pi}'$ , and then remove  $B \setminus (A \setminus \mathcal{K})$  from the conditioning set as all colliders are still kept open due to either being  $\mathcal{G}(\tilde{\mathcal{M}})$ -ancestor of  $A \cap \mathcal{K}$  or of  $A \setminus \mathcal{K}$ .

Consider now any path in  $\mathcal{G}(\mathcal{M})$  that is  $A$ -open, for  $A \subseteq \mathcal{I} \cup \mathcal{K}$ . Any edge on the path between a system node and a context node must be of the form  $i \leftarrow k$  (with  $i \in \mathcal{I}$ ,  $k \in \mathcal{K}$ ) or  $k \rightarrow i$ , where  $i$  is in another strongly-connected component than  $k$  and  $k$  cannot be in  $A$  (because the path was assumed to be  $A$ -open). Replacing each longest subpath consisting entirely of context nodes  $k_0 \dots k_n$  (with all  $k_0, \dots, k_n \in \mathcal{K}$ ) by a corresponding  $A$ -open path in  $\mathcal{G}(\tilde{\mathcal{M}})_{\mathcal{K}}$  between  $k_0$  and  $k_n$  gives a walk in  $\mathcal{G}(\tilde{\mathcal{M}})$  that by construction is also  $A$ -open in  $\mathcal{G}(\tilde{\mathcal{M}})$ . Any system collider on this walk must be a collider on the original path, and therefore  $\mathcal{G}(\mathcal{M})$ -ancestor of  $A$ , and therefore also  $\mathcal{G}(\tilde{\mathcal{M}})$ -ancestor of  $A$ . Any system non-collider on this walk is also a system non-collider on the original path and therefore not in  $A$  or, in case of  $\sigma$ -separation, pointing only to nodes in the same strongly-connected component of  $\mathcal{G}(\mathcal{M})$ , and hence of  $\mathcal{G}(\tilde{\mathcal{M}})$ . Any context non-collider on this walk cannot be in  $A$ , or, in case of  $\sigma$ -separation, points to the same strongly-connected component in  $\mathcal{G}(\tilde{\mathcal{M}})$ , since the replacing path in  $\mathcal{G}(\tilde{\mathcal{M}})_{\mathcal{K}}$  was  $A$ -open by construction. Any context collider on this walk that is a  $\mathcal{G}(\mathcal{M})_{\mathcal{K}}$ -ancestor of  $(A \cap \mathcal{K}) \cup B$ , and therefore must be  $\mathcal{G}(\tilde{\mathcal{M}})$ -ancestor of  $A$ . The walk can be shortened into an  $A$ -open path in  $\mathcal{G}(\tilde{\mathcal{M}})$ .

Similarly, one can show that any path in  $\mathcal{G}(\tilde{\mathcal{M}})$  that is  $A$ -open, there must be a corresponding path in  $\mathcal{G}(\mathcal{M})$  that is  $A$ -open.  $\blacksquare$

We can now state the last JCI assumption. This one can be useful whenever JCI Assumptions 1 and 2 are made:

**Assumption 3** (“Generic context model”, optional) *The context graph  $\mathcal{G}(\mathcal{M})_{\mathcal{K}}$  is of the following special form:*

$$\forall k, k' \in \mathcal{K} : \quad k \neq k' \implies k \leftrightarrow k' \in \mathcal{G} \wedge k \rightarrow k' \notin \mathcal{G}(\mathcal{M}).$$

This assumption can be useful to avoid having to infer the causal relations between the context variables, when one is only interested in the causal relations involving the system variables. It is illustrated in Figure 8. The following Corollary states that this assumption can be made without loss of generality if the context distribution contains no conditional independences:

**Corollary 13** *Assume that JCI Assumptions 0, 1 and 2 hold for SCM  $\mathcal{M}$ . Then there exists an SCM  $\tilde{\mathcal{M}}$  that satisfies JCI Assumptions 0, 1 and 2 and 3, such that*

- (i) *the influence diagrams coincide:  $\mathcal{G}(\mathcal{M})_{\mathcal{I}|\text{do}(\mathcal{K})} = \mathcal{G}(\tilde{\mathcal{M}})_{\mathcal{I}|\text{do}(\mathcal{K})}$ ;*
- (ii) *for any perfect intervention on the system variables  $\text{do}(I, \xi_I)$  with  $I \subseteq \mathcal{I}$  (including the non-intervention  $I = \emptyset$ ),  $\tilde{\mathcal{M}}_{\text{do}(I, \xi_I)}$  is observationally equivalent to  $\mathcal{M}_{\text{do}(I, \xi_I)}$ ;*
- (iii) *if the context distribution  $\mathbb{P}_{\mathcal{M}}(\mathbf{C})$  contains no conditional independences, the same  $\sigma$ -separations hold in  $\mathcal{G}(\tilde{\mathcal{M}})$  as in  $\mathcal{G}(\mathcal{M})$ ; if in addition, the Directed Global Markov Property holds for  $\mathcal{M}$ , then also the same  $d$ -separations hold in  $\mathcal{G}(\tilde{\mathcal{M}})$  as in  $\mathcal{G}(\mathcal{M})$ .*

**Proof** Let  $\mathcal{M}$  be an SCM of the form (7). Under JCI Assumption 1, the structural equations for the context variables do not depend on the system variables:

$$C_k = f_k(\mathbf{C}_{\text{PA}_{\mathcal{H}}(k) \cap \mathcal{K}}, \mathbf{E}_{\text{PA}_{\mathcal{H}}(k) \cap \mathcal{J}}), \quad k \in \mathcal{K}.$$

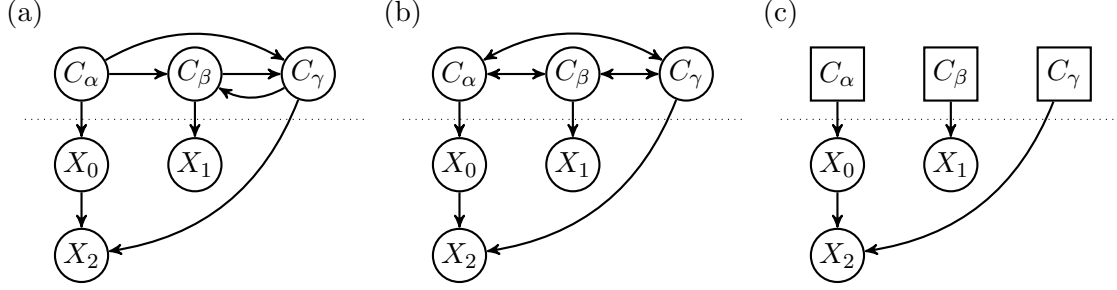


Figure 8: Example functional graphs of (a) a true SCM  $\mathcal{M}$  and (b) the modified SCM  $\tilde{\mathcal{M}}$  constructed in the proof of Corollary 13 that satisfies JCI Assumption 3, and (c) their corresponding influence diagram  $\mathcal{G}(\mathcal{M})_{\mathcal{I}|\text{do}\mathcal{K}}$ . Corollary 13 gives sufficient conditions for  $\tilde{\mathcal{M}}$  and  $\mathcal{M}$  to be equivalent for our purposes.

Because of JCI Assumption 2,  $\text{PA}_{\mathcal{H}}(\mathcal{K}) \cap \text{PA}_{\mathcal{H}}(\mathcal{I}) \cap \mathcal{J} = \emptyset$ , i.e., the context variables do not share any exogenous variable with the system variables.

Consider now the modified SCM  $\tilde{\mathcal{M}}$  of the form:

$$\tilde{\mathcal{M}} : \begin{cases} C_k & = g_k(\mathbf{E}_C), & k \in \mathcal{K} \\ X_i & = f_i(\mathbf{X}_{\text{PA}_{\mathcal{H}}(i) \cap \mathcal{I}}, \mathbf{C}_{\text{PA}_{\mathcal{H}}(i) \cap \mathcal{K}}, \mathbf{E}_{\text{PA}_{\mathcal{H}}(i) \cap \mathcal{J}}), & i \in \mathcal{I}, \\ \mathbb{P}(\mathbf{E}) & = \prod_{j \in \tilde{\mathcal{J}}} \mathbb{P}(E_j), \end{cases}$$

where  $\tilde{\mathcal{J}} = \mathcal{J} \cup \{C\}$  contains an additional exogenous variable  $\mathbf{E}_C \in \prod_{k \in \mathcal{K}} \mathcal{E}_k$  with components  $(\mathbf{E}_C)_k \in \mathcal{C}_k$  with distribution  $\mathbb{P}(\mathbf{E}_C) = \mathbb{P}_{\mathcal{M}}(\mathbf{C})$  and  $g_k$  the projection on the  $k$ 'th component  $g_k : \mathbf{E}_C \mapsto (\mathbf{E}_C)_k$ . By construction, this SCM  $\tilde{\mathcal{M}}$  satisfies JCI Assumptions 1 and 2. The only aspect that requires a little bit of thought is the simplicity of  $\tilde{\mathcal{M}}$ . Take  $\mathcal{O} \subseteq \mathcal{I}$  and consider the solution function for  $\mathcal{O}$  according to  $\mathcal{M}$ :  $\mathbf{g}_{\mathcal{O}} : \mathcal{X}_{(\text{PA}_{\mathcal{H}}(\mathcal{O}) \setminus \mathcal{O}) \cap \mathcal{I}} \times \mathcal{C}_{(\text{PA}_{\mathcal{H}}(\mathcal{O}) \setminus \mathcal{O}) \cap \mathcal{K}} \times \mathcal{E}_{\text{PA}_{\mathcal{H}}(\mathcal{O}) \cap \mathcal{J}} \rightarrow \mathcal{X}_{\mathcal{I} \setminus \mathcal{O}}$ . This solves the structural equations for  $\mathcal{O} \setminus \mathcal{I}$ , and since these are the same for  $\tilde{\mathcal{M}}$  as for  $\mathcal{M}$ , the same solution function works also for  $\tilde{\mathcal{M}}$ . Now take  $\mathcal{Q} \subseteq \mathcal{K}$  and consider the solution function  $\mathbf{g}_{\mathcal{Q}} : \mathcal{E}_C \rightarrow \mathcal{C}_{\mathcal{Q}}$  with components  $\mathcal{E}_C \rightarrow \mathcal{C}_k : \mathbf{e}_C \mapsto g_k(\mathbf{e}_C)$ ,  $k \in \mathcal{Q}$ . Any other solution function can be obtained by composition. We conclude that  $\tilde{\mathcal{M}}$  is simple.  $\tilde{\mathcal{M}}$  also induces the same context distribution  $\mathbb{P}_{\tilde{\mathcal{M}}}(\mathbf{E}) = \mathbb{P}_{\mathcal{M}}(\mathbf{E})$  and satisfies JCI Assumption 3 by construction. The other statements now follow by applying Theorem 12, where the only thing left to show is that  $\mathcal{G}(\tilde{\mathcal{M}})_{\mathcal{K}}$  and  $\mathcal{G}(\mathcal{M})_{\mathcal{K}}$  induce the same  $\sigma$ -separations if the context distribution  $\mathbb{P}_{\mathcal{M}}(\mathbf{C})$  contains no conditional independences, and the same  $d$ -separations if in addition the Directed Global Markov Property holds for  $\mathcal{M}$ .

Marginalizing out the system variables (both in  $\mathcal{M}$  as well as in  $\tilde{\mathcal{M}}$ ) yields  $\mathcal{M}_{\setminus \mathcal{I}}$  and  $\tilde{\mathcal{M}}_{\setminus \mathcal{I}}$ , with functional graphs  $\mathcal{G}(\mathcal{M}_{\setminus \mathcal{I}}) = \mathcal{G}(\mathcal{M})_{\mathcal{K}}$  and  $\mathcal{G}(\tilde{\mathcal{M}}_{\setminus \mathcal{I}}) = \mathcal{G}(\tilde{\mathcal{M}})_{\mathcal{K}}$ , respectively. By the Generalized Directed Global Markov property, since  $\mathbb{P}_{\mathcal{M}}(\mathbf{C})$  has no conditional independences, there must be a  $K$ - $\sigma$ -open path in  $\mathcal{G}(\mathcal{M})_{\mathcal{K}}$  between any two context nodes  $k \neq k' \in \mathcal{K}$ , for any  $K \subseteq \mathcal{K}$  with  $\{k, k'\} \cap K = \emptyset$ . If the Directed Global Markov property holds for  $\mathcal{M}$ , then it holds for  $\mathcal{G}(\mathcal{M}_{\setminus \mathcal{I}})$ , and hence there must even be a  $K$ - $d$ -open path in

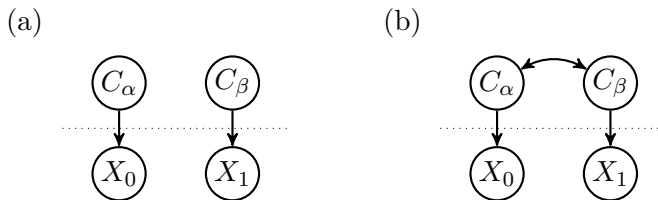


Figure 9: Functional graphs of (a) the true SCM  $\mathcal{M}$  and (b) the modified SCM  $\tilde{\mathcal{M}}$  constructed in the proof of Corollary 13 that are not Markov equivalent. The graph in (a) is identifiable under JCI Assumptions 0–2. The joint distribution  $\mathbb{P}_{\mathcal{M}}(\mathbf{X}, \mathbf{C})$  is not faithful with respect to the graph in (b), which is the minimal one that also satisfies JCI Assumption 3 such that  $\mathbb{P}_{\mathcal{M}}(\mathbf{X}, \mathbf{C})$  is Markov with respect to it.

$\mathcal{G}(\mathcal{M})_{\mathcal{K}}$  between any two context nodes  $k \neq k' \in \mathcal{K}$ , for any  $K \subseteq \mathcal{K}$  with  $\{k, k'\} \cap K = \emptyset$ . Since by construction  $\mathcal{G}(\tilde{\mathcal{M}})_{\mathcal{K}}$  contains all bidirected edges  $k \leftrightarrow k'$ , there is a  $K$ - $d$ -open path in  $\tilde{\mathcal{M}}_{\setminus \mathcal{I}}$  between any two context nodes  $k \neq k' \in \mathcal{K}$ , for any  $K \subseteq \mathcal{K}$  with  $\{k, k'\} \cap K = \emptyset$ . ■

JCI Assumption 3 is typically made for convenience. When our aim is not to model the causal relations *between* the context variables, but just to use the context variables as an aid to model the causal relations between system variables and between context and system variables, Corollary 13 shows that without loss of generality we can assume JCI Assumption 3 (if JCI Assumptions 1 and 2 are made and the context distribution contains no conditional independences). The causal discovery algorithm then need not waste time on learning the causal relations between context variables but can focus directly on learning the causal relations involving the system variables.

Note that the genericity assumption in statement (iii) of Corollary 13 (i.e.,  $\mathbb{P}_{\mathcal{M}}(\mathbf{C})$  containing no conditional independences) is necessary, as the simple counterexample in Figure 9 shows. Depending on how well the causal discovery algorithm can handle faithfulness violations, model misspecification due to incorrectly assuming JCI Assumption 3 even though  $\mathbb{P}(\mathbf{C})$  contains conditional independences might prevent successful identification of the causal relationships between system variables. Therefore, it is prudent to check that the empirical context distribution  $\mathbb{P}(\mathbf{C})$  indeed contains no conditional independences before making JCI Assumption 3.

An example of a common situation in which the context distribution contains no conditional independences is what we refer to as a *diagonal design* (see also Table 1). This is a simple experimental design that is often used to discover the effects of single interventions when one is not interested in understanding the interactions that multiple interventions might have. Note that two non-constant binary variables  $X, Y$  can only be independent if  $\mathbb{P}(X = 1, Y = 1) > 0$ . Even more, they can only be conditionally independent given a third discrete variable  $\mathbf{Z}$  if  $\mathbb{P}(X = 1, Y = 1 | \mathbf{Z} = \mathbf{z}) > 0$  for all  $\mathbf{z}$  with  $\mathbb{P}(\mathbf{Z} = \mathbf{z}) > 0$ . Therefore, each pair of context variables is dependent in a diagonal design (as there is no context in which a pair of context variables simultaneously obtains the value 1), even conditionally on any subset of the other context variables. In other words, the context distribution  $\mathbb{P}(\mathbf{C})$  corresponding to any diagonal design contains no (conditional) independences.

Table 1: Example of a diagonal design with 5 context variables. If the context variables are indicators of interventions, the context with  $C_k = 0$  for all  $k \in \mathcal{K}$  corresponds with the purely observational setting, and the other contexts in which one  $C_k = 1$  and the other  $C_l = 0$  for  $l \neq k$  correspond with a particular intervention each.

$C_\alpha$	$C_\beta$	$C_\gamma$	$C_\delta$	$C_\epsilon$	possible interpretation
0	0	0	0	0	observational
1	0	0	0	0	intervention $\alpha$
0	1	0	0	0	intervention $\beta$
0	0	1	0	0	intervention $\gamma$
0	0	0	1	0	intervention $\delta$
0	0	0	0	1	intervention $\epsilon$

JCI Assumption 3 could be modified for situations in which the context distribution does contain conditional independences. For example, in the extreme case in which all context variables are jointly independent, one would simply assume that  $\mathcal{G}(\mathcal{M})$  contains no directed and no bidirected edges between context variables. Such situations may occur for symmetric experimental designs in which all context variables are jointly independent by design (for example, factorial designs with equal sample sizes in each experimental context). However, we believe that this occurs less often in practice than the generic case in which all context variables are (conditionally) dependent, because resource constraints often lead experimenters to deviate from completely symmetric experimental designs. Therefore, rather than assuming the context variables to be jointly independent as a default, we have opted here for the more generic default of assuming that no conditional independences hold between context variables in the context distribution. More generally, one could replace JCI Assumption 3 by assuming that  $\mathcal{G}(\mathcal{M})_{\mathcal{K}}$  equals a certain graph describing the known conditional independences in the experimental design. Theorem 12 can be applied to these more general situations as well and shows that for the purpose of constraint-based causal discovery, any functional graph for the context variables that implies the observed conditional independences (i.e., any functional graph that is Markov equivalent to the true causal graph) works. An even easier option in general is to just ignore JCI Assumption 3 and instead try to infer the context subgraph  $\mathcal{G}_{\mathcal{K}}$  from the data. This is computationally more expensive, but doesn't seem to make much of a difference in terms of accuracy, as our experimental results will show.

JCI Assumption 3 only makes sense when both JCI Assumptions 1 and 2 are made. If we would not make JCI Assumption 1 or 2, the causal relations between the observed context variables will have testable consequences in the joint distribution in general. For an example, see Figure 10. Here,  $C_\alpha$  could be “lab”, and  $C_2$  could be the “temperature” at which an experiment is performed. In this case, we get different conditional independences in the joint distribution  $\mathbb{P}(X_0, X_1, C_\alpha, C_\beta)$  if lab causes temperature rather than that they are confounded (for example, by geographical location). Something similar can happen if context variables can be caused by system variables.

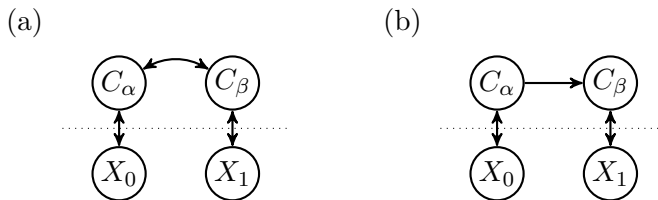


Figure 10: If JCI Assumption 2 does not apply, the causal relations between context variables have testable consequences for the conditional independences in the joint distribution. (a)  $X_0 \not\perp\!\!\!\perp X_1 \mid \{C_\alpha, C_\beta\}$ ; (b)  $X_0 \perp\!\!\!\perp X_1 \mid \{C_\alpha, C_\beta\}$ .

### 3.4.5 SUMMARY OF JCI ASSUMPTIONS

Summarizing, the JCI framework rests on different assumptions, one of which is required, whereas the others are all optional. The basic assumption that is required is JCI Assumption 0, which states that the meta-system consisting of context and system can be described by a simple SCM. This is really just the standard assumption made throughout the causal discovery literature, but now applied to the meta-system rather than to the system. In addition, assumptions about the causal relationships of the context variables can be made, which are all optional and can be decided on a case-by-case basis. In most cases, we would expect JCI Assumption 1 (no system variable is a potential cause of any context variable) to apply. In some cases, also JCI Assumption 2 (there is no potential irreducible latent common cause of context and system variables) applies. If both apply, one can assume JCI Assumption 3 for convenience if the context distribution contains no conditional independences. In other words, we only need to model the observed conditional independences in the context distribution, not necessarily their causal relations, when our interest is in modeling the causal relations involving system variables only.

For causal discovery in the JCI framework, knowledge of the intervention *targets* (or more generally, which system variables are affected directly by which context variables) is not necessary, but it is certainly helpful and can be exploited similarly to other available background knowledge, depending on the algorithm used to implement JCI. When using JCI for a combination of different interventional data sets, intervention targets can be learnt from data when they are not known, similarly to how the effects of system variables can be learnt. One main advantage of the JCI framework is that it offers a unified way to deal with different types of interventions, as discussed in Section 3.3. Therefore, knowledge of intervention *types* (e.g., is it a perfect surgical intervention, or a mechanism change?) is also not necessary, but can still be helpful as it provides additional background knowledge that may be exploited for causal discovery.

In concluding this subsection, we observe that the JCI framework generalizes and combines the ideas of causal discovery from purely observational data and of causal discovery by means of randomized controlled trials. Indeed, note that if JCI is applied to a single context (i.e., 0 context variables), it reduces to the standard setting of causal discovery from purely observational data described in Section 2.3. If JCI is applied to a setting with a single context variable and a single system variable, JCI (with Assumptions 1 and 2)

reduces to the Randomized Controlled Trial setting described in Section 2.2. Therefore, the Joint Causal Inference framework truly generalizes both these special cases.

### 3.5 Causal Discovery from Multiple Datasets

In this section, we discuss how causal discovery from multiple contexts is performed in the Joint Causal Inference framework. Our starting point is the assumption that some model of the form (7) is an appropriate causal model for the system and its context, and we have obtained samples of the system variables in multiple contexts. Suppose that the exact model  $\mathcal{M}$  (and in particular, its causal graph  $\mathcal{G}(\mathcal{M})$ ) is unknown to us. The goal of *causal discovery* is to infer as much as possible about the causal graph  $\mathcal{G}(\mathcal{M})$  from the available data and from available background knowledge about context and system.

Let us denote the dataset for context  $\mathbf{c} \in \mathcal{C}$  as  $\mathcal{D}^{(\mathbf{c})} = ((x_{in}^{(\mathbf{c})})_{i \in \mathcal{I}})_{n=1}^{N_{\mathbf{c}}}$ , and for simplicity, assume that no values are missing. The number of samples in each context, given by  $N_{\mathbf{c}}$ , is allowed to depend on the context. As a first step, which is conceptually different from typical constraint-based approaches to causal discovery from multiple contexts, we *pool* the data, thereby representing it as a single dataset  $\mathcal{D} = (\mathbf{x}_n, \mathbf{c}_n)_{n=1}^N$  where  $N = \sum_{\mathbf{c} \in \mathcal{C}} N_{\mathbf{c}}$ . We then assume that  $\mathcal{D}$  is an i.i.d. sample of  $\mathbb{P}_{\mathcal{M}}(\mathbf{X}, \mathbf{C})$ , where  $(\mathbf{X}, \mathbf{C}, \mathbf{E})$  is a solution of the SCM  $\mathcal{M}$  of the form (7).

In setting up the problem, we have made the simplifying assumptions that the measurement procedure is not subject to selection bias, nor to (independent) measurement error (Blom et al., 2018). We will assume that the data has been generated by an SCM in accordance with JCI Assumption 0, and optionally, a subset of JCI Assumptions 1, 2 and 3. To enable constraint-based causal discovery, we will assume that the joint distribution  $\mathbb{P}_{\mathcal{M}}(\mathbf{X}, \mathbf{C})$  is faithful with respect to the functional graph  $\mathcal{G}(\mathcal{M})$ , using the appropriate separation criterion ( $\sigma$ -separation in general, or  $d$ -separation for specific cases, as discussed in Section 3.6). We will discuss the ramifications of the faithfulness assumption in more detail in Section 3.6.

**Definition 14** *We say that a particular feature of  $\mathcal{G}(\mathcal{M})$  is identifiable from  $\mathbb{P}_{\mathcal{M}}(\mathbf{X}, \mathbf{E})$  and background knowledge if the feature is present in the graph  $\mathcal{G}(\tilde{\mathcal{M}})$  of any SCM  $\tilde{\mathcal{M}}$  with  $\mathbb{P}_{\tilde{\mathcal{M}}}(\mathbf{X}, \mathbf{E}) = \mathbb{P}_{\mathcal{M}}(\mathbf{X}, \mathbf{E})$  that incorporates the background knowledge.*

“Feature” could refer to the presence or absence of a direct edge, a directed path, a bidirected edge, arbitrary subgraphs, or even the complete graph. The task of causal discovery is then to identify as many features of  $\mathcal{G}(\mathcal{M})$  as possible from the data, the i.i.d. sample  $\mathcal{D}$  of  $\mathbb{P}_{\mathcal{M}}(\mathbf{X}, \mathbf{E})$ , and the available background knowledge.

The trick of the Joint Causal Inference framework that allows one to deal with data from multiple contexts  $(\mathcal{D}^{(\mathbf{c})})_{\mathbf{c} \in \mathcal{C}}$  is that by incorporating the context variables explicitly, and pooling the data, we have now reduced the causal discovery problem to one that is mathematically equivalent to causal discovery from purely observational data  $\mathcal{D}$  and applicable background knowledge on the causal relations between context and system variables (a subset of JCI Assumptions 1 and 2). If applicable, JCI Assumption 3 can be made to reduce the computational effort.

After discussing the faithfulness assumption in more detail, we will give a few suggestions of how JCI can be implemented in Section 3.7.

### 3.6 Structural Causal Models: Faithfulness Assumption

In this subsection we will discuss the subtleties of the faithfulness assumption in the JCI setting. We will see how it allows us to deal also with perfect interventions. On the one hand, our faithfulness assumption is weaker than the usual assumption made in constraint-based causal discovery methods that combine different datasets by analyzing each dataset separately before combining the results into one coherent model. On the other hand, it is stronger, since it implies restrictions on the context distribution: it must be faithful to some DMG.

Given a simple SCM  $\mathcal{M}$  of the form (7), the joint distribution  $\mathbb{P}_{\mathcal{M}}(\mathbf{X}, \mathbf{C})$  induced by the SCM satisfies the Generalized Directed Global Markov Property (Theorem 7) with respect to the functional graph  $\mathcal{G} := \mathcal{G}(\mathcal{M})$  of the SCM, i.e., any  $\sigma$ -separation  $U \perp_{\mathcal{G}}^{\sigma} V | W$  between sets of nodes  $U, V, W \subseteq \mathcal{I} \cup \mathcal{K}$  in the graph  $\mathcal{G}$  implies a conditional independence  $\tilde{\mathbf{X}}_U \perp_{\mathbb{P}(\mathbf{X}, \mathbf{C})} \tilde{\mathbf{X}}_V | \tilde{\mathbf{X}}_W$ , where we write  $\tilde{\mathbf{X}} := (\mathbf{X}, \mathbf{C})$ . Under the additional assumptions of Theorem 8, the stronger Directed Global Markov Property holds (i.e., the  $d$ -separation criterion).

For constraint-based causal discovery, some type of faithfulness assumption is usually made. For simplicity, the faithfulness assumption that we make in this work is the standard one, but we apply it to the combination of system and its environment: we assume that the joint distribution  $\mathbb{P}_{\mathcal{M}}(\mathbf{X}, \mathbf{C})$  is faithful with respect to the functional graph  $\mathcal{G}(\mathcal{M})$  of  $\mathcal{M}$ . In other words, any conditional independence  $\tilde{\mathbf{X}}_U \perp_{\mathbb{P}(\mathbf{X}, \mathbf{C})} \tilde{\mathbf{X}}_V | \tilde{\mathbf{X}}_W$  for sets of nodes  $U, V, W \subseteq \mathcal{I} \cup \mathcal{K}$  is due to the  $\sigma$ -separation  $U \perp_{\mathcal{G}}^{\sigma} V | W$  in  $\mathcal{G}$  (or  $d$ -separation, if applicable), and no other conditional independences in  $\mathbb{P}_{\mathcal{M}}(\mathbf{X}, \mathbf{C})$  exist. In particular, this assumption rules out any (conditional) independence between context variables in case JCI Assumption 3 is made.

This faithfulness assumption allows us to deal with different types of interventions, including perfect (surgical) interventions. For example, for the surgical intervention on  $X_2$  in the example illustrated in Figure 6(b), the causal graphs  $\mathcal{G}_{\mathcal{I}}^{(c)}$  (restricted to the system variables  $\{X_i\}_{i \in \mathcal{I}}$ ) depend on the context  $\mathbf{c} \in \mathcal{C}$ : in the observational context ( $C_{\alpha} = 0$ ),  $X_1 \rightarrow X_2$ , whereas in the interventional context ( $C_{\alpha} = 1$ ), this direct causal relation is no longer present (as it has been overruled by the perfect intervention). This does not invalidate the faithfulness of the joint distribution  $\mathbb{P}(C_{\alpha}, X_1, X_2, X_3)$  with respect to the joint causal graph. Indeed, even though  $X_1 \perp X_2 | C_{\alpha} = 0$ , we still have  $X_1 \not\perp X_2 | C_{\alpha}$  because  $X_1 \not\perp X_2 | C_{\alpha} = 1$ .<sup>8</sup> In other words, the fact that  $\mathbb{P}(\mathbf{X} | C_{\alpha} = 1)$  is *not* faithful to the system subgraph  $\mathcal{G}_{\mathcal{I}}$  (i.e., the induced subgraph of the causal graph  $\mathcal{G}$  on the system nodes  $\mathcal{I}$ ) does not lead to any problem as long as we are not going to test for independences in the subset of data corresponding to context  $C_{\alpha} = 1$  separately, but restrict ourselves to testing independences only in the *pooled* data set that combines all contexts.

Causal discovery methods that analyze data from each context separately (e.g., Hauser and Bühlmann, 2012; Triantafillou and Tsamardinos, 2015; Hyttinen et al., 2014) typically make a stronger faithfulness assumption. In our notation, such approaches assume that  $\mathbb{P}(\mathbf{X} | \mathbf{C} = \mathbf{c})$  is faithful w.r.t. a causal subgraph  $\mathcal{G}_{\mathcal{I}}^{(c)}$  that may be context-dependent, and must then reason about how these context-dependent subgraphs are related, explicitly

8. Note that for a discrete context domain  $\mathcal{C}$ , we have that  $A \perp B | \mathbf{C}$  if and only if  $A \perp B | \mathbf{C} = \mathbf{c}$  for all  $\mathbf{c}$  with  $p(\mathbf{c}) > 0$ . More generally,  $A \perp B | \mathbf{C}$  if and only if  $A \perp B | \mathbf{C} = \mathbf{c}$  for almost all  $\mathbf{c}$ .



relying on knowledge about the type of interventions (typically assuming that the interventions are surgical interventions with known targets). One advantage of our approach is its broader applicability, since it does not rely on knowledge of the intervention types, nor of the intervention targets.

Under JCI Assumption 1, the faithfulness assumption for the joint distribution  $\mathbb{P}_{\mathcal{M}}(\mathbf{C}, \mathbf{X})$  implies in particular that the context distribution  $\mathbb{P}_{\mathcal{M}}(\mathbf{C})$  is faithful with respect to some DMG (indeed, more specifically, to  $\mathcal{G}(\mathcal{M})_{\mathcal{K}}$ ). When JCI Assumption 3 is made, the faithfulness assumption even means that the context distribution  $\mathbb{P}(\mathbf{C})$  does not contain any (conditional) independences. In case the empirical context distribution  $\mathbb{P}(\mathbf{C})$  *does* contain (conditional) independences, one has several options. The first is to modify the assumed functional graph of the context variables in JCI Assumption 3 such that the context distribution is faithful to it. For example, if all context variables are jointly independent, one could simply assume that the context graph  $\mathcal{G}(\mathcal{M})_{\mathcal{K}}$  has no (directed or bidirected) edges at all. The second option is to omit JCI Assumption 3 or some analogue of it completely. In that case, the faithfulness assumption still imposes the restriction that the context distribution can be faithfully modeled by *some* DMG. The third option is to use only data corresponding to a certain subset of the contexts and to ignore data from other contexts. In addition, one can sometimes work around conditional independences by partitioning the set of context variables into groups of context variables, and using combined context variables instead of the original context variables. This will be illustrated in the next paragraph. Finally, one could ignore the faithfulness violations in the context distribution and hope that the causal discovery algorithm will handle them well. This last approach usually means that it will be harder to guarantee consistency of the approach. Note that the faithfulness assumption for the context variables is actually testable, since the empirical context distribution is available, and can be directly tested for conditional independences.

The faithfulness assumption also rules out deterministic relations between the variables that lead to faithfulness violations. In particular, there could be deterministic relations between context variables. For example, in the experimental design of the experiments in Sachs et al. (2005) described in Tables 2 and 3,  $C_{\alpha}$  is a (deterministic) function of  $C_{\theta}$  and  $C_l$ :  $C_{\alpha} = \neg(C_{\theta} \vee C_l)$ . One might naïvely believe that this could be dealt with by simply removing context variable  $C_{\alpha}$  from consideration, leaving only context variables  $C_{\beta}, \dots, C_l$  as observed context variables, none of which is a (deterministic) function of the others. However, marginalizing out context variables may give rise to violations of JCI Assumption 2, as we have seen in Section 3.4.3. An operation that is generally allowed is *grouping* context variables together. In the case of the Sachs et al. (2005) experimental design, we can combine  $C_{\alpha}$ ,  $C_{\theta}$  and  $C_l$  together into a single context variable given by the 3-tuple  $(C_{\alpha}, C_{\theta}, C_l)$ . Accidentally, in this case this would be mathematically equivalent to the tuple  $(C_{\theta}, C_l)$ , but the interpretation of  $(C_{\alpha}, C_{\theta}, C_l)$  is different from that of  $(C_{\theta}, C_l)$ . Another option would be to ignore a subset of the contexts. In this case, one could exclude the two contexts with  $C_{\theta} = 1$  or  $C_l = 1$ . Then,  $C_{\alpha}$  becomes a constant, and constants can be safely ignored (or, trivially combined with any other context variable).<sup>9</sup>

9. In an earlier draft of this work (Magliacane et al., 2016a), we proposed to handle deterministic relations between context variables by using the notion of  $D$ -separation, first presented in (Geiger et al., 1990) and later extended in (Spirtes et al., 2000). However, this notion does not provide a complete characterization of conditional independences due to a combination of graph structure and deterministic relations.

Table 2: Left: Experimental design used by Sachs et al. (2005).  $N_{\mathbf{C}}$  is the number of data samples in context  $\mathbf{C}$ . Interpretation of context variables is provided in Table 3. Right: Different choice of context variables:  $C_\alpha$ ,  $C_\theta$  and  $C_\iota$  have been grouped together into a single combined context variable  $(C_\alpha, C_\theta, C_\iota)$  in order to deal with the deterministic relation  $C_\alpha = \neg(C_\theta \vee C_\iota)$ .

$C_\alpha$	$C_\beta$	$C_\gamma$	$C_\delta$	$C_\epsilon$	$C_\zeta$	$C_\eta$	$C_\theta$	$C_\iota$	$N_{\mathbf{C}}$	$C_\beta$	$C_\gamma$	$C_\delta$	$C_\epsilon$	$C_\zeta$	$C_\eta$	$(C_\alpha, C_\theta, C_\iota)$	$N_{\mathbf{C}}$
1	0	0	0	0	0	0	0	0	853	0	0	0	0	0	0	(1,0,0)	853
1	1	0	0	0	0	0	0	0	902	1	0	0	0	0	0	(1,0,0)	902
1	0	1	0	0	0	0	0	0	911	0	1	0	0	0	0	(1,0,0)	911
1	0	0	1	0	0	0	0	0	723	0	0	1	0	0	0	(1,0,0)	723
1	0	0	0	1	0	0	0	0	810	0	0	0	1	0	0	(1,0,0)	810
1	0	0	0	0	1	0	0	0	799	0	0	0	0	1	0	(1,0,0)	799
1	0	0	0	0	0	1	0	0	848	0	0	0	0	0	1	(1,0,0)	848
0	0	0	0	0	0	0	0	1	913	0	0	0	0	0	0	(0,1,0)	913
0	0	0	0	0	0	0	0	1	707	0	0	0	0	0	0	(0,0,1)	707
1	1	1	0	0	0	0	0	0	899	1	1	0	0	0	0	(1,0,0)	899
1	1	0	1	0	0	0	0	0	753	1	0	1	0	0	0	(1,0,0)	753
1	1	0	0	1	0	0	0	0	868	1	0	0	1	0	0	(1,0,0)	868
1	1	0	0	0	1	0	0	0	759	1	0	0	0	1	0	(1,0,0)	759
1	1	0	0	0	0	1	0	0	927	1	0	0	0	0	1	(1,0,0)	927

Table 3: For each context variable in Table 2: reagents used in this experimental setting, and expected intervention type and targets as based on (our interpretation of) biological background knowledge described in Sachs et al. (2005).

	Reagent	Intervention
$C_\alpha$	$\alpha$ -CD3, $\alpha$ -CD28	global activator
$C_\beta$	ICAM-2	global activator
$C_\gamma$	AKT inhibitor	activity of AKT
$C_\delta$	G0076	activity of PKC
$C_\epsilon$	Psitectorigenin	abundance of PIP2
$C_\zeta$	U0126	MEK activity
$C_\eta$	LY294002	PIP2, PIP3 mechanism change
$C_\theta$	PMA	PKC activity
$C_\iota$	$\beta$ 2CAMP	PKA activity

Under JCI Assumption 3, it is advisable to check that there are indeed no conditional independences between context variables in the empirical context distribution. More generally, one should check whether the conditional independences in the context distribution can be faithfully described by a DMG. If not, one can try to work around by applying the tricks discussed in the last paragraph (grouping context variables, and leaving out certain contexts). When grouping context variables, one should note that the inferred causal relations from context variables to system variables may no longer be easily interpretable. A simple example that illustrates this is to consider two interventions that are always performed together: when drug A is prescribed, also drug B is prescribed, and vice versa. In that case we cannot be sure whether the effect on outcome is due to drug A or to drug B. Nonetheless, we can still use the inferred causal relations between system variables.

---

Therefore, in this work we use the simpler techniques of grouping context variables and ignoring certain contexts to deal with faithfulness violations due to deterministic relations between context variables.

### 3.7 Implementing JCI

Any causal discovery method that is applicable under the assumptions described in Section 3.5 can be used for Joint Causal Inference. Identifiability greatly benefits from taking into account the available background knowledge on the causal graph stemming from the applicable JCI assumptions as discussed in Section 3.4. In addition, taking into account background knowledge on targets of intervention variables may help considerably. Some logic-based causal discovery methods (e.g., Hyttinen et al., 2014; Triantafillou and Tsamardinos, 2015; Forré and Mooij, 2018), are ideally suited to exploit such background knowledge. For other methods (e.g., FCI (Spirtes et al., 1995; Zhang, 2008) or methods that focus on ancestral relations, e.g., ACI (Magliacane et al., 2016b)), incorporating all background knowledge is less straightforward and as far as we know cannot be done with off-the-shelf implementations.

Given a causal discovery algorithm for purely observational data that can exploit the JCI background knowledge, we can implement JCI in a straightforward fashion: (i) introduce context variables, if not already provided; (ii) pool all data sets, including the values of the context variables; (iii) handle faithfulness violations between context variables by grouping context variables and/or leaving out certain contexts, if necessary; (iv) apply the causal discovery algorithm on the pooled data, while taking into account the appropriate JCI background knowledge. Any soundness and consistency results for the causal discovery algorithm in the purely observational setting (plus background knowledge) directly apply to the JCI setting, as long as there is no model misspecification (i.e., if the assumed JCI assumptions do hold for the true model). If we use JCI Assumption 3 (or something similar), we can use Theorem 12 to show that what the algorithm concludes about the causal relations concerning system variables is still correct. In the remainder of this subsection, we discuss four specific JCI implementations that we will study in our experiments in Section 4.

#### 3.7.1 LOCAL CAUSAL DISCOVERY (LCD)

Perhaps the first implementation of JCI is provided by the LCD algorithm by Cooper (1997). LCD is a very simple constraint-based causal discovery algorithm that can be used for the purely observational causal discovery setting where certain background knowledge is available, and in particular, in the JCI setting. The basic idea behind the LCD algorithm is the following result (which we generalized to allow for cycles):

**Proposition 15** *Suppose that the data-generating process on three variables  $X_1, X_2, X_3$  can be represented by a faithful, simple SCM  $\mathcal{M}$  and that the sampling procedure is not subject to selection bias. If  $X_2$  is not a potential cause of  $X_1$ , the following conditional (in)dependences in the observational distribution  $\mathbb{P}(X_1, X_2, X_3)$*

$$X_1 \not\perp\!\!\!\perp X_2, \quad X_2 \not\perp\!\!\!\perp X_3, \quad X_1 \perp\!\!\!\perp X_3 \mid X_2$$

*imply that the functional graph  $\mathcal{G}(\mathcal{M})$  must be one of the three graphs in Figure 11. In particular,*

- (i)  $X_3$  is not a potential cause of  $X_2$ ;
- (ii)  $X_2$  is a direct cause of  $X_3$  w.r.t.  $\{X_1, X_2, X_3\}$ ;
- (iii)  $X_2$  and  $X_3$  are unconfounded, i.e., the causal effect of  $X_2$  on  $X_3$  is given by:

$$\mathbb{P}(X_3 \mid \text{do}(X_2 = x_2)) = \mathbb{P}(X_3 \mid X_2 = x_2). \quad (9)$$



Figure 11: All possible DAGs detected by LCD.

**Proof** The proof proceeds by enumerating all DAGs on three variables and ruling out the ones that do not satisfy the assumptions. Since  $\mathcal{M}$  is assumed to be simple, the assumption that  $X_2$  is not a potential cause of  $X_1$  implies that there is no directed edge  $X_2 \rightarrow X_1$  in the functional graph  $\mathcal{G}(\mathcal{M})$ . If there were an edge between  $X_1$  and  $X_3$ ,  $X_1 \perp\!\!\!\perp X_3 \mid X_2$  would not hold (faithfulness). Also, since  $X_1 \not\perp\!\!\!\perp X_2$ ,  $X_1$  and  $X_2$  must be adjacent (Markovness). Similarly,  $X_2$  and  $X_3$  must be adjacent.  $X_2$  cannot be a collider on any path between  $X_1$  and  $X_3$  (faithfulness). Since the only possible edges between  $X_1$  and  $X_2$  are  $X_1 \rightarrow X_2$  and  $X_1 \leftrightarrow X_2$  (both of which have an arrowhead at  $X_2$ ), this means that there must be a directed edge  $X_2 \rightarrow X_3$ , but there cannot be a bidirected edge  $X_2 \leftrightarrow X_3$  or directed edge  $X_2 \leftarrow X_3$ . In other words, the only three possible functional graphs are the ones in Figure 11. The causal do-calculus applied to  $\mathcal{G}(\mathcal{M})$  yields (9). Because  $X_2 \not\perp\!\!\!\perp X_3$ ,  $X_2$  is a direct cause of  $X_3$  with respect to  $\{X_1, X_2, X_3\}$ . ■

In a JCI setting where JCI Assumption 1 is made, we can directly apply LCD for causal discovery on tuples  $\langle C_k, X_i, X_{i'} \rangle$  with  $k \in \mathcal{K}$ ,  $i \neq i' \in \mathcal{I}$  on the pooled data. In our experiments, we use  $-\log p_{13}$ , where  $p_{13}$  denotes the  $p$ -value of a (marginal) independence test on  $X_1$  and  $X_3$ , as a heuristic confidence measure of the LCD prediction. In other words, the more evidence we have against the null hypothesis that  $X_1$  and  $X_3$  are independent, the higher our confidence that the conclusions are correct.

A conservative version of LCD has been applied by Triantafillou et al. (2017) to the task of inferring signaling networks from mass cytometry data. An algorithm closely related to LCD, named “Trigger”, has been applied on genomics data (Chen et al., 2007). Chen et al. (2007) motivate the JCI assumptions in the setting of learning the causal relations between gene expression levels using SNPs as context variables. Since the DNA content cannot be caused by gene expression levels, the JCI Assumption 1 is satisfied. Chen et al. (2007) then argue that Mendelian randomization justifies the JCI Assumption 2. Finally, a single conditional independence in the pooled data (as in LCD) provides the desired evidence for an unconfounded causal relation between two gene expression levels. In this sense, it is a direct combination of RCTs with constraint-based causal discovery, using a single context variable and two system variables at a time.

### 3.7.2 INVARIANT CAUSAL PREDICTION (ICP)

Invariant Causal Prediction (ICP) exploits invariance in the conditional distribution of a target variable given its direct causes across multiple contexts, assuming that none of the contexts corresponds with an intervention that targets the target variable (Peters et al., 2016). The implementation described in Peters et al. (2016) handles linear relationships, arbitrary interventions (as long as they do not change the conditional distribution of the effect variable given its direct causes), assumes the absence of latent confounders between target variable and its direct causes, and the absence of cycles involving the target variable.

One of the main advantages of this method over others is that it provides (conservative) confidence intervals on direct causal relationships that do not require the faithfulness assumption to be made. The authors discuss several possible extensions to broaden the scope of the method, but do not address this in all generality. A nonlinear extension of the method has been proposed recently (Heinze-Deml et al., 2017).

We will show here that one can also interpret ICP as a particular implementation of the JCI framework, even in a very general setting with nonlinear relations between variables and with irreducible latent common causes and cycles present (although faithfulness is then required).

We start by introducing a notation and two useful Lemmas that generalize their acyclic  $d$ -separation analogues from (Claassen and Heskes, 2011).

**Definition 16** *Let  $X, Y, Z, S \subseteq \mathcal{V}$  be sets of nodes in a DMG  $\mathcal{G} = \langle \mathcal{V}, \mathcal{E}, \mathcal{F} \rangle$ . Let  $\perp$  denote a DMG-separation property, e.g.,  $d$ -separation ( $\perp^d$ ) or  $\sigma$ -separation ( $\perp^\sigma$ ). We say that the minimal separation*

$$X \perp_{\mathcal{G}} Y \mid S \cup [Z]$$

*holds if and only if*

$$X \perp_{\mathcal{G}} Y \mid S \cup Z \quad \wedge \quad \forall Q \subsetneq Z : X \not\perp_{\mathcal{G}} Y \mid S \cup Q.$$

*Similarly: we say that the minimal connection*

$$X \not\perp_{\mathcal{G}} Y \mid S \cup [Z]$$

*holds if and only if*

$$X \not\perp_{\mathcal{G}} Y \mid S \cup Z \quad \wedge \quad \forall Q \subsetneq Z : X \perp_{\mathcal{G}} Y \mid S \cup Q.$$

Note that despite the notation, a minimal connection is not the logical negation of a minimal separation. Minimal connections imply the absence of certain ancestral relations:

**Lemma 17** *Let  $\{X\}, \{Y\}, S, \{Z\} \subseteq \mathcal{V}$  be mutually disjoint sets of nodes in a DMG  $\mathcal{G} = \langle \mathcal{V}, \mathcal{E}, \mathcal{F} \rangle$ . For both  $d$ -separation ( $\perp^d$ ) and  $\sigma$ -separation ( $\perp^\sigma$ ), we have that:*

$$X \not\perp_{\mathcal{G}} Y \mid S \cup [\{Z\}] \implies Z \notin \text{AN}_{\mathcal{G}}(\{X, Y\} \cup S).$$

**Proof** The minimal connection means that all paths between  $X$  and  $Y$  are closed when conditioning on  $S$  and there exists at least one path between  $X$  and  $Y$  that is open when conditioning on  $S \cup \{Z\}$ . For  $d$ -separation, this means that such a path (i) contains a collider not in  $\text{AN}_{\mathcal{G}}(S)$ , (ii) every collider is in  $\text{AN}_{\mathcal{G}}(S \cup \{Z\})$ , (iii) every non-collider is not in  $S \cup \{Z\}$ . For  $\sigma$ -separation, this means that such a path (i) contains a collider not in  $\text{AN}_{\mathcal{G}}(S)$ , (ii) every collider is in  $\text{AN}_{\mathcal{G}}(S \cup \{Z\})$ , (iii) every non-collider is either not in  $S \cup \{Z\}$ , or if it is, it points to neighboring nodes in the same strongly-connected component only.

Thus there exists a path between  $X$  and  $Y$  that contains a collider in  $\text{AN}_{\mathcal{G}}(\{Z\})$  that is not in  $\text{AN}_{\mathcal{G}}(S)$ . If  $Z \in \text{AN}_{\mathcal{G}}(S)$  this would be a contradiction. If  $Z \in \text{AN}_{\mathcal{G}}(X)$ , then

we can consider the walk between  $X$  and  $Y$  obtained from composing the subpath of the original path between  $Y$  and the first collider (starting from  $Y$ ) in  $\text{AN}_{\mathcal{G}}(\{Z\}) \setminus \text{AN}_{\mathcal{G}}(S)$  with a directed path to  $Z$  and then on to  $X$ , without passing through nodes in  $S$ . This walk between  $X$  and  $Y$  must be open when conditioning on  $S$ , and hence there exists a path between  $X$  and  $Y$  that is open when conditioning on  $S$ , a contradiction. Similarly we obtain a contradiction if  $Z \in \text{AN}_{\mathcal{G}}(Y)$ . ■

On the other hand, minimal separations imply the presence of certain ancestral relations:

**Lemma 18** *Let  $\{X\}, \{Y\}, S, Z \subseteq \mathcal{V}$  be mutually disjoint sets of nodes in a DMG  $\mathcal{G} = \langle \mathcal{V}, \mathcal{E}, \mathcal{F} \rangle$ . For both  $d$ -separation ( $\perp^d$ ) and  $\sigma$ -separation ( $\perp^\sigma$ ), we have that:*

$$X \perp_{\mathcal{G}} Y \mid S \cup [Z] \implies Z \subseteq \text{AN}_{\mathcal{G}}(\{X, Y\} \cup S).$$

**Proof** Let  $Q \subsetneq Z$ . Consider a path between  $X$  and  $Y$  that is open when conditioning on  $S \cup Q$ , but becomes blocked when conditioning on  $S \cup Z$ . For  $d$ -separation, this means that (i) every collider on the path is in  $\text{AN}_{\mathcal{G}}(S \cup Q)$ , (ii) every non-collider is not in  $S \cup Q$ , and (iii) it contains a non-collider in  $S \cup Z$ . For  $\sigma$ -separation, this means that (i) every collider on the path is in  $\text{AN}_{\mathcal{G}}(S \cup Q)$ , (ii) every non-collider is either not in  $S \cup Q$  or if it is, it points to a neighboring node on the path in another strongly-connected component, and (iii) it contains a non-collider in  $S \cup Z$  that points to a neighboring node on the path in another strongly-connected component. In both cases, we have that (i) every collider on the path is in  $\text{AN}_{\mathcal{G}}(S \cup Q)$  and (ii) it contains a non-collider in  $Z \setminus Q$ . Consider a maximal directed subpath of the path starting at a non-collider  $U$  in  $Z \setminus Q$  and stopping at a collider or at an endpoint. Then  $U \in \text{AN}_{\mathcal{G}}(\{X, Y\} \cup S \cup Q)$ .

So, for each  $Q \subsetneq Z$ , there exists a  $U \in Z \setminus Q$  with  $U \in \text{AN}_{\mathcal{G}}(\{X, Y\} \cup S \cup Q)$ . Thus for every  $Z_i \in Z$ , we either obtain (taking  $Q = Z \setminus \{Z_i\}$ ) an ancestral relation of the form  $Z_i \in \text{AN}_{\mathcal{G}}(\{X, Y\} \cup S)$ , or, otherwise, at least  $Z_i \in \text{AN}_{\mathcal{G}}(Z_j)$  for some  $Z_j \in Z \setminus \{Z_i\}$ . Define a directed graph  $\mathcal{A}$  with nodes  $Z \cup \{\omega\}$  (where  $\omega$  represents  $\{X, Y\} \cup S$ ) and add an edge  $Z_i \rightarrow \omega$  whenever our construction yields an ancestral relation of the form  $Z_i \in \text{AN}_{\mathcal{G}}(\{X, Y\} \cup S)$ , or otherwise, an edge  $Z_i \rightarrow Z_j$  if our construction yields  $Z_i \in \text{AN}_{\mathcal{G}}(Z_j)$ .

Then, taking the transitive closure of the constructed directed graph  $\mathcal{A}$  and using transitivity of ancestral relations, for any  $Z_i \in Z$  we either obtain  $Z_i \in \text{AN}_{\mathcal{G}}(\{X, Y\} \cup S)$ , or  $Z_i$  is in some strongly-connected component  $C \subseteq Z$  in  $\mathcal{A}$ . In the latter case, we can apply the reasoning above (taking now  $Q = Z \setminus C$ ) to conclude that there exists a  $Z_j \in C$  with  $Z_j \in \text{AN}_{\mathcal{G}}(\{X, Y\} \cup S)$  or  $Z_j \in \text{AN}_{\mathcal{G}}(C')$  where  $C' \subseteq Z$  is another strongly-connected component. Since the strongly-connected components of  $Z$  form an acyclic structure, repeating this reasoning a finite number of times, we ultimately conclude that  $Z_i \in \text{AN}_{\mathcal{G}}(\{X, Y\} \cup S)$ . ■

An immediate Corollary is:

**Corollary 19** *Let  $X, Y \in \mathcal{V}$  be different nodes in a DMG  $\mathcal{G} = \langle \mathcal{V}, \mathcal{E}, \mathcal{F} \rangle$ . For  $d$ -separation ( $\perp^d$ ) or  $\sigma$ -separation ( $\perp^\sigma$ ), consider  $Z^* := \bigcap \{Z : X \notin Z, Y \notin Z, X \perp_{\mathcal{G}} Y \mid Z\}$ . Then  $Z^* \subseteq \text{AN}_{\mathcal{G}}(\{X, Y\})$ .*

**Proof** First, note that  $Z^* = \bigcap \{Z : X \notin Z, Y \notin Z, X \perp_{\mathcal{G}} Y \mid [Z]\}$ . From Lemma 18,  $X \perp_{\mathcal{G}} Y \mid [Z]$  implies  $Z \subseteq \text{AN}_{\mathcal{G}}(\{X, Y\})$ . Hence  $Z^* \subseteq \text{AN}_{\mathcal{G}}(\{X, Y\})$ . ■

The following result strengthens the results in Peters et al. (2016):

**Corollary 20** *Consider the JCI setting with a single context variable  $C$  and multiple system variables  $\{X_i\}_{i \in \mathcal{I}}$ . Under JCI Assumptions 0 and 1 and faithfulness, the ICP estimator for target  $i \in \mathcal{I}$ :*

$$J_i^* := \bigcap \{I \subseteq \mathcal{I} \setminus \{i\} : C \perp\!\!\!\perp_{\mathbb{P}_{\mathcal{M}}(C, \mathbf{X})} X_i \mid \mathbf{X}_I\}$$

*satisfies  $J_i^* \subseteq \text{AN}_{\mathcal{G}(\mathcal{M})}(i)$ , i.e., the set  $J_i^*$  consists only of potential (possibly indirect) causes of  $i$ .*

Note that this means that asymptotically, ICP outputs a subset of ancestors of the target variable, even in the presence of confounders and linear or nonlinear cycles. This means that we can interpret ICP as a particular causal discovery algorithm implementing the JCI framework.

### 3.7.3 ASD-JCI

Another JCI implementation is easily obtained by using the algorithm by Hyttinen et al. (2014) and its generalization to  $\sigma$ -separation (Forré and Mooij, 2018). Hyttinen et al. (2014) proposed formulating causal discovery as an optimization problem over possible causal graphs, where the loss function sums the weights of all the conditional (in)dependences present in the data that would be violated for a certain underlying causal graph, assuming Markov and faithfulness properties. The input consists of a list of weighted conditional independence statements. The weights  $\lambda$  encode the confidence in the conditional (in)dependence, where a weight of  $\lambda = \infty$  corresponds to a “hard constraint” (absolute certainty) and a weight of  $\lambda = 0$  corresponds to “no evidence at all”. Hyttinen et al. (2014) provide an encoding of the notion of  $d$ -separation in ASP, while Forré and Mooij (2018) generalize the encoding to  $\sigma$ -separation. The optimization problem is solved by making use of an off-the-shelf ASP solver.

There may be multiple optimal solutions to the optimization problem, because the underlying causal graph may not be identifiable from the inputs. Nonetheless, some of the features of the causal graph (e.g., the presence or absence of a certain directed edge) may still be identifiable. We employ the method proposed by Magliacane et al. (2016b) for scoring the confidence that a certain feature is present by calculating the difference between the optimal losses under the additional hard constraints that the feature is present vs. that the feature is absent. Magliacane et al. (2016b) showed that this algorithm for scoring features is sound for oracle inputs and asymptotically consistent under reasonable assumptions.

We will make use of the weights proposed in Magliacane et al. (2016b):  $\lambda_j = \log p_j - \log \alpha$ , where  $p_j$  is the  $p$ -value of a statistical test for the  $j$ 'th conditional independence statement, with independence as null hypothesis, and  $\alpha$  is a significance level (e.g., 1%) that should decrease with sample size at a suitable rate. These weights have the desirable property that independences get a lower weight than strong dependences.

As we will need an acronym for this algorithm later, we will henceforth refer to it as ASD (Accounting for Strong Dependences), as it basically tries to explain the observed dependences in the data, taking into account the strength of these dependences. This is

fundamentally different from other constraint-based algorithms such as PC or FCI, which give priority to observed *independences* and do not take into account the strength of dependences.

Taking into account the JCI background knowledge (a subset of JCI Assumptions 1, 2 and 3), and possible background knowledge on intervention targets, is trivial thanks to the expressive power of ASP, and can be done with a few lines of ASP code. The resulting algorithm is very accurate but scales only up to a few variables due to the combinatoric explosion. Incorporating JCI Assumption 3 considerably reduces computation time.

### 3.7.4 FCI-JCI

By slightly extending the constraint-based causal discovery algorithm FCI (Spirtes et al., 1995; Zhang, 2008), it can be used in a JCI setting. The FCI algorithm assumes that the data was generated by an acyclic SCM. Since interventions and selection bias interact non-trivially, we postpone treatment of selection bias to future work, and focus here on the simpler version of FCI that assumes no selection bias is present. The modifications to FCI to handle the JCI background knowledge are also most straightforward when assuming no selection bias, because then we can ignore some of the orientation rules (Zhang, 2008), and in addition, the PAG cannot contain undirected edges, so any invariant tail mark that is found implies an arrowhead on the opposite edge end.

The FCI algorithm consists of two main phases: an adjacency search phase leading to the undirected skeleton, followed by an edge orientation phase. In the adjacency search the algorithm searches for conditional independences to eliminate edges from the graph. The subsequent orientation stage consists of a set of graphical rules that allow invariant edge marks, signifying either causal (tail marks) or non-causal (arrowhead marks) relations, to be added to the skeleton. For a single observational data set the final result is a so-called Partial Ancestral Graph (PAG) that is a concise representation of ancestral relations and conditional independences. The PAG represents a set of Maximal Ancestral Graphs (MAGs) (Richardson and Spirtes, 2002), and each MAG represents a set of ADMGs (Triantafyllou and Tsamardinos, 2015), and each ADMG represents an infinite set of DAGs (with arbitrary number of latent variables). In the case of purely observational data, the PAG output by FCI gives a complete representation of the Markov equivalence class (Zhang, 2008).

Extending FCI such that it can take into account the additional JCI background knowledge on the adjacency and causal relations between the combined set of context and system variables (see also Section 3.4) is straightforward:

- If JCI Assumption 3 is made, all context variables are connected by bidirected edges, and the adjacency phase of FCI has to be adapted accordingly by not removing any edges between context variables; afterwards, all edges between context variables can be oriented as  $k \leftrightarrow k'$  for  $k \neq k' \in \mathcal{K}$ . In the subsequent phase of orienting v-structures, only system variables can take on the role of the collider in the “v”.
- If JCI Assumption 1 is made, then (since we are assuming no selection bias) any adjacent pair of a context variable  $k \in \mathcal{K}$  and a system variable  $i \in \mathcal{I}$  must be connected in the MAG by an edge with an arrowhead at  $i$ . Therefore, after the adjacency phase, all edges between a context and a system variable can be oriented as  $k \ast \rightarrow i$ , pointing from the context variable  $k \in \mathcal{K}$  to the system variable  $i \in \mathcal{I}$ .



- If both JCI Assumptions 1 and 2 are made, any adjacent pair of a context variable  $k \in \mathcal{K}$  and a system variable  $i \in \mathcal{I}$  must be connected by a directed edge  $k \rightarrow i$  (since we are assuming no selection bias) in the MAG. Hence, after the adjacency phase, all edges between a context and a system variable can be oriented as  $k \rightarrow i$ , pointing from the context variable  $k \in \mathcal{K}$  to the system variable  $i \in \mathcal{I}$ .

The subsequent orientation phase of the FCI algorithm does not need to be adapted. This is a consequence of our assumption that selection bias is absent.

Under the assumptions of the JCI framework, this slightly modified FCI algorithm (that we will refer to as FCI-JCI) is sound. We do not know whether it is complete, i.e., its output might not be a PAG that completely characterizes the equivalence classes of ADMGs in the JCI setting. To read off the presence and absence of (identifiable) ancestral relations from the PAG output by FCI-JCI, one can use the results of Roumpelaki et al. (2016).

### 3.8 Related work

In this section we discuss related work. Since the pioneering work by Fisher (1935), many different causal discovery methods that can deal with data from different contexts have been proposed. Table 4 provides an overview of some of these methods and the features they offer. Note that JCI offers most features of all methods. By implementing the JCI framework using sophisticated causal discovery methods for observational data (plus background knowledge) one obtains versatile and powerful causal discovery algorithms for multiple contexts. We will now discuss in detail some of the aspects of the related work.

#### 3.8.1 IRREDUCIBLE LATENT COMMON CAUSES

Most score-based methods (like the ones by Cooper and Yoo, 1999; Tian and Pearl, 2001; Sachs et al., 2005; Eaton and Murphy, 2007; Hauser and Bühlmann, 2012; Mooij and Heskes, 2013; Oates et al., 2016a) and some constraint-based methods (Zhang et al., 2017) assume *causal sufficiency*, i.e., that no irreducible latent common causes are present. This assumption is likely to be violated in practice and may lead to wrong conclusions. An example is provided in Figure 12. The true causal graph with confounded system variables is identifiable from conditional independences in the data by JCI. Due to the presence of an irreducible latent common cause of  $X_0$  and  $X_1$ , score-based methods will infer the wrong causal graph(s).

#### 3.8.2 CYCLES

As we have seen in Proposition 9, the method by Fisher (1935) can handle cycles. Less well-known is that also LCD (Cooper, 1997) and Trigger (Chen et al., 2007) can handle cycles. Hyttinen et al. (2012) provide an algorithm for linear SCMs with cycles and confounders that deals with perfect interventions. The methods by Hyttinen et al. (2014) and Mooij and Heskes (2013) can deal with cycles in a linear (or approximately linear) setting. The method by Hyttinen et al. (2014) relies on  $d$ -separation, which only applies in certain settings (see Theorem 8). The method can be modified to use  $\sigma$ -separation instead (Forré and Mooij, 2018). The way Mooij and Heskes (2013) handle cycles is not as straightforward. Generally, their method could handle nonlinear cyclic models, but for computational reasons,

	Irreducible latent common causes	Nonlinear mechanisms	Cycles	Perfect interventions	Mechanism changes	Activity interventions	Side effects	Other context changes	Unknown intervention/context targets	Learns intervention/context targets	Multiple system variables	Different variables per context	Combination strategy
(Cooper and Yoo, 1999)	-	+	-	+	-	-	-	-	-	-	+	-	b
(Tian and Pearl, 2001)	-	+	-	-	+	-	-	+	-	-	+	-	b
(Sachs et al., 2005)	-	+	-	+	-	-	-	-	-	-	+	-	b
(Eaton and Murphy, 2007)	-	+	-	+	+	+	+	+	+	+	+	-	b
(Claassen and Heskes, 2010)	+	+	-	-	+	+	+	+	+	-	+	+	a
(Tillman and Spirtes, 2011)	+	+	-	-	+	+	+	+	+	-	+	+	a
(Hauser and Bühlmann, 2012)	-	+	-	+	-	-	-	-	-	-	+	-	b
(Hyttinen et al., 2012)	+	-	+	+	-	-	-	-	-	-	+	-	a
(Mooij and Heskes, 2013)	-	±	±	+	+	+	-	+	-	-	+	-	b
(Hyttinen et al., 2014)	+	+	±	+	-	-	-	-	-	-	+	+	a
(Triantafillou and Tsamardinos, 2015)	+	+	-	+	-	-	-	-	-	-	+	+	a
(Rothenhäusler et al., 2015)	+	-	±	-	-	-	-	+	+	+	+	-	a
(Oates et al., 2016a)	-	-	-	-	-	-	-	+	-	-	+	-	b
(Zhang et al., 2017)	-	+	-	+	+	+	+	+	+	+	+	-	b
(Forré and Mooij, 2018)	+	+	+	+	-	-	-	-	-	-	+	+	a
Joint Causal Inference:	+	+	+	+	+	+	+	+	+	+	+	±	b
(Fisher, 1935)	+	+	+	+	+	+	+	+	+	+	-	-	b
LCD (Cooper, 1997)	+	+	+	+	+	+	+	+	+	-	+	-	b
Trigger (Chen et al., 2007)	+	+	+	+	+	+	+	+	+	+	+	-	b
ICP (Peters et al., 2016)	+	+	+	+	+	+	+	+	+	-	+	-	b
FCI-JCI	+	+	-	+	+	+	+	+	+	+	+	-	b
ASD-JCI	+	+	+	+	+	+	+	+	+	+	+	-	b

Table 4: Overview of causal discovery methods that can combine data from multiple contexts. Features offered by the original implementations of these methods are indicated. Combination strategies are: (a) obtain statistics or constraints from each context separately and then construct a single causal graph based on the combined statistics, (b) pool all data and construct a single causal graph directly from the pooled data. When a feature is offered only under additional restrictive assumptions, it is indicated with a  $\pm$  sign.

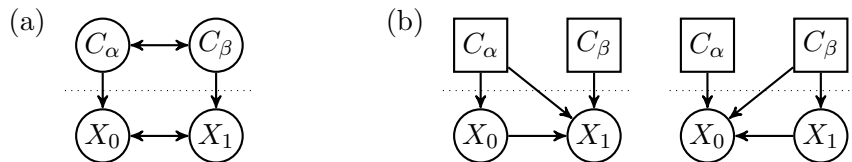


Figure 12: Example that shows that incorrectly assuming causal sufficiency can lead to wrong conclusions. (a) True causal graph, with confounded system variables, which can be completely identified by JCI; (b) optimally scoring influence diagrams according to the method of Eaton and Murphy (2007).

their implementation linearizes the SCMs around each (context-dependent) equilibrium, thereby basically assuming that  $d$ -separation holds *within each context*. The method by Rothenhäusler et al. (2015) assumes linearity and can deal with cycles in that case, under a certain condition that suffices to prove identifiability of the method. The method by Peters et al. (2016) can handle cycles, as our Corollary 20 shows. The JCI framework in general can deal with cycles, but requires its implementation to support this. For example, FCI-JCI was not designed for this.

### 3.8.3 SELECTION BIAS

The only method that claims to be able to deal with selection bias (i.e., conditioning on a latent variable that is a common effect of one or more of the observed variables), at least to some extent, is the IOD algorithm (Tillman and Spirtes, 2011). It allows for different sets of observed (system) variables in each context and for different distributions in each context, while assuming that each context can be described by a MAG that is the marginal of a common PAG defined on the union of all system variables. Under the assumption that hidden variables are selection variables, both with respect to the union of all system variables, and with respect to each context’s set of system variables separately, this framework can handle selection bias. It performs conditional independence tests in each dataset separately, and merges the  $p$ -values of the test results using Fisher’s method. It then constructs the PAG that represents simultaneously all contexts. Since it doesn’t assume invariance of the distribution across contexts, it can deal with a single (latent) context variable that models mechanism changes or other “soft” interventions that do not change the conditional independences in the distribution.

### 3.8.4 PERFECT INTERVENTIONS

Even though JCI allows for contexts to correspond with perfect (surgical) interventions, it does not fully exploit all available information when contexts correspond with perfect interventions with *known* targets. Other methods (e.g., Hyttinen et al., 2014; Triantafillou and Tsamardinos, 2015), that assume perfect interventions with known targets, can sometimes identify a larger part of the causal graph in such situations. An example is provided in Figure 16. It features two system variables that are dependent, both observationally, as in a stochastic surgical intervention on  $X_0$ . If the marginal distribution of  $X_0$  differs between

the two contexts, the graph is identifiable by JCI and by the methods of Hyttinen et al. (2014); Triantafillou and Tsamardinos (2015). If the marginal distribution of  $X_0$  is identical for both contexts, then JCI can no longer identify the causal graph, whereas the methods by Hyttinen et al. (2014); Triantafillou and Tsamardinos (2015) still can.

### 3.8.5 IMPERFECT INTERVENTIONS AND OTHER CONTEXT CHANGES

Cooper and Yoo (1999) provided the first score-based causal discovery algorithm that could deal with data from multiple contexts, focussing on perfect interventions with known targets. They describe in detail how to handle perfect interventions and introduced the idea of adding explicit context variables to deal with imperfect changes, which was later refined by Eaton and Murphy (2007), who provide an algorithm that can handle (stochastic) perfect interventions with unknown targets, soft interventions, and mechanism changes. Also Sachs et al. (2005) use a score-based causal discovery algorithm based on the ideas of Cooper and Yoo (1999) that uses a greedy search strategy through the space of DAGs.

Tian and Pearl (2001) were the first to consider mechanism changes. They deal with sequences of subsequent mechanism changes, exploiting changes in the distribution to infer descendants of the changed mechanism. This is followed by a constraint-based approach from observational data that uses the background knowledge. A similar approach using the differences between data from experimental conditions and an observational baseline as background knowledge for a constraint-based approach was applied by Magliacane et al. (2016b) on the data of Sachs et al. (2005).

Claassen and Heskes (2010) handle certain *environment changes*: direct causal relations between system variables are assumed to be invariant across contexts, but latent confounding (and more generally, the exogenous distribution) may differ. Rothenhäusler et al. (2015) assume stochastic *shift interventions* in which the mean of a target variable is shifted by an (independent) random amount. Various multi-task “structure learning” (i.e., Bayesian network learning) approaches that put a prior on the similarity of the DAGs in multiple contexts which encourages them to be similar have been proposed (e.g., Oates et al., 2014, 2016b).

Some methods which allow for a single context variable have been applied in settings on time-series data, by using time as the context variable (Friedman et al., 2000; Zhang et al., 2017). This extends the more usual approach of treating time-series data by assuming *invariance* of the causal structure across time as in dynamic Bayesian networks (DBNs) (Murphy, 2002), methods based on Granger causality (Granger, 1969), or constraint-based approaches (Entner and Hoyer, 2010).

JCI allows one to handle all interventions and context changes discussed above in a unified way.

### 3.8.6 MULTIPLE CONTEXT VARIABLES

Some causal discovery methods for combining data from different contexts that explicitly consider a context variable, allow for a single context variable only, for example, LCD, ICP, and the method by Zhang et al. (2017). There is an important advantage to allowing multiple context variables, as JCI does generally. One might argue that the case of multiple context variables can always be reduced to a case with a single context variable, by

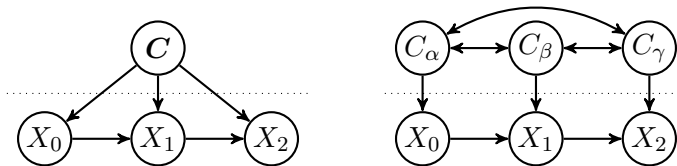


Figure 13: Example that shows that allowing multiple context variables (right) has advantages over considering a single context variable only (left). The causal graph on the right can be identified by JCI from conditional independences in the data, whereas the causal graph on the left is not identifiable.

simply combining all context variables  $\{C_k\}_{k \in \mathcal{K}}$  into a single tuple  $\mathbf{C} = (C_k)_{k \in \mathcal{K}}$ . However, this reduction to a single context variable typically loses important information. This is illustrated in Figure 13. When using only a single context variable in that case, the DMG cannot be identified from conditional independences in the data. On the other hand, when using all three context variables with JCI, the complete DMG can be identified, even when the causal relations between context and system variables are unknown.

### 3.8.7 DEPENDENT CONTEXT VARIABLES

If one allows for multiple context variables and considers the joint distribution on context and system variables, as we do in JCI, one should account for possible dependencies between the context variables. Indeed, incorrectly assuming the context variables to be independent *a priori* may lead to wrong conclusions. An example is provided in Figure 14. In that example, incorrectly assuming the contexts to be independent leads to the wrong conclusion that context variable  $C_\beta$  causes the system variable  $X_0$ , at least for causal discovery algorithms that are tolerant to faithfulness violations. Another example is provided in Figure 15, where incorrectly assuming the contexts to be independent leads to the wrong conclusion that system variables  $X_0$  and  $X_1$  are confounded, at least when assuming known intervention targets.

This issue was recognized and addressed in recent work (Oates et al., 2016a) by introducing a novel graphical modeling framework, Conditional DAGs (CDAGs), which bears some similarity with our approach. However, a disadvantage of the CDAG framework is that existing causal discovery methods cannot be directly applied to learn a CDAG from data, and the wealth of results on causal modeling with SCMs cannot be used directly. One of the key advantages of the JCI framework is that it utilizes existing theory and methods, as it reduces a causal discovery problem from multiple contexts to a purely observational one with background knowledge. This is one of the reasons why JCI offers many more features than the approach by Oates et al. (2016a). We also note that CDAGs can be dealt with as a special case of the JCI framework.

### 3.8.8 DATA MISSING AT RANDOM

A particular case of data missing at random that has been addressed by some of the methods is when the set of observed variables differ between datasets, while still having some overlap.



Figure 14: Example that shows that incorrectly assuming independent context variables can lead to wrong conclusions when using causal discovery algorithms that are tolerant to faithfulness violations. Left: true causal graph, with dependent context variables, which is identifiable by JCI. Right: causal graph that reproduces all (conditional) dependences and minimizes the number of faithfulness violations, when (incorrectly) assuming that context variables are independent.



Figure 15: Example that shows that incorrectly assuming independent context variables can lead to wrong conclusions when assuming known intervention targets. Left: true causal graph, with dependent context variables. Right: wrong causal graph inferred under JCI Assumptions 1 and 2 and known intervention targets, when (incorrectly) assuming that context variables are independent.



Figure 16: Example that shows that JCI does not always allow to fully exploit available background knowledge for (stochastic) surgical interventions. Left: true causal graph, where  $C_\alpha \in \{0, 1\}$  with  $C_\alpha = 0$  corresponding to an observational setting and  $C_\alpha = 1$  to a stochastic surgical intervention on  $X_0$ . This graph is identifiable by JCI. Right: special case where  $\mathbb{P}(X_0 | C_\alpha = 1) = \mathbb{P}(X_0 | C_\alpha = 0)$ , which is not identifiable by JCI, but is still identifiable by the method of Hyttinen et al. (2014).

The first one to address this using constraint-based causal discovery was Tillman (2009), and several other methods have been proposed over the years (Claassen and Heskes, 2010; Tillman and Spirtes, 2011; Hyttinen et al., 2014; Triantafillou and Tsamardinos, 2015). JCI can only deal with this when strengthening its faithfulness assumption: one would need to assume that the context variables are discrete, and that every conditional distribution  $\mathbb{P}(\mathbf{X} | \mathbf{C} = \mathbf{c})$  for  $\mathbf{c} \in \mathbf{C}$  with  $\mathbb{P}(\mathbf{C} = \mathbf{c}) > 0$  is faithful with respect to the same DMG  $\mathcal{G}_{\mathcal{I}}$ . However, in doing so we would give up the ability to handle perfect (“surgical”) interventions.

### 3.8.9 INFLUENCE DIAGRAMS

Our representation of a system within a context imposed by its environment bears strong similarities with influence diagrams (Dawid, 2002). A formal difference is that we consider the context variables to be random variables that reflect the empirical distribution of the experimental design, whereas in influence diagrams they are interpreted as non-random decision variables. The advantage of treating context variables as random variables is that this allows one to apply standard causal discovery techniques (designed for random variables) *jointly* on system and context variables. In particular, the notion of (statistical) conditional independence (Dawid, 1979) suffices. If one would like to treat the context variables as decision (i.e., non-random) variables, extended notions of conditional independence would be necessary (such as in Constantinou and Dawid (2017)), but this would only work under additional assumptions, for example, that context variables are discrete-valued. Since we can always view the context variables as random variables in the *empirical distribution* of the experimental design, this allows us to stick with the standard notion of conditional independence and in addition, handle non-discrete context variables as well.

### 3.8.10 SELECTION DIAGRAMS

Our representation also bears some similarities with selection diagrams (Bareinboim and Pearl, 2013), but one crucial difference is that we are modeling the *joint* distribution on the intervention and system variables, whereas a selection diagram represents the *conditional* distribution of the system variables given the intervention (“selection”) variables. Because

we are modeling the joint distribution and not only the conditional one, we can apply standard causal discovery techniques directly on pooled data, something that would not be as trivial when using selection diagrams instead.

Bareinboim and Pearl (2013) define a selection diagram as follows:

**Definition 21** *Let  $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \mathcal{H}, \mathcal{X}, \mathcal{E}, \mathbf{f}, \mathbb{P}_{\mathcal{E}} \rangle$  and  $\mathcal{M}^* = \langle \mathcal{I}, \mathcal{J}, \mathcal{H}, \mathcal{X}, \mathcal{E}, \mathbf{f}^*, \mathbb{P}_{\mathcal{E}^*} \rangle$  be two acyclic SCMs corresponding to two different contexts, that only differ with respect to their causal mechanisms and exogenous distributions. In particular, they share the same functional graph  $\mathcal{H}$  and hence also their causal graphs are identical,  $\mathcal{G}(\mathcal{M}) = \mathcal{G}(\mathcal{M}^*)$ . The selection diagram  $\overline{\mathcal{G}}$  induced by  $\langle \mathcal{M}, \mathcal{M}^* \rangle$  is the acyclic directed mixed graph with nodes  $\mathcal{I} \dot{\cup} \mathcal{I}'$ , where  $\mathcal{I}'$  is a copy of  $\mathcal{I}$  of selection variable indices (one  $i' \in \mathcal{I}'$  for each endogenous variable index  $i \in \mathcal{I}$ ), such that*

1. *the induced subgraph of  $\overline{\mathcal{G}}$  on  $\mathcal{I}$  equals the causal graph  $\overline{\mathcal{G}}_{\mathcal{I}} = \mathcal{G}(\mathcal{M})$ , and*
2. *for each  $i \in \mathcal{I}$  such that  $f_i \neq f_i^*$  or  $\mathbb{P}_{\mathcal{E}_{\text{pa}_{\mathcal{G}}(i)}} \neq \mathbb{P}_{\mathcal{E}^*_{\text{pa}_{\mathcal{G}}(i)}}$  there is an edge  $i' \rightarrow i$  in  $\overline{\mathcal{G}}$ .*

Consider a JCI model of the form (7) with a single binary context variable  $C$ . The joint SCM can be split into two context-specific SCMs,  $\mathcal{M}^0$  and  $\mathcal{M}^1$ , and the induced selection diagram  $\mathcal{D}$  can be obtained from the causal graph  $\mathcal{G}(\mathcal{M})$  as follows: (i) each edge  $i_1 \rightarrow i_2$  or  $i_1 \leftrightarrow i_2$  in  $\mathcal{G}(\mathcal{M})$  between system variables  $i_1, i_2 \in \mathcal{I}$  is also in  $\mathcal{D}$ ; (ii) if  $C \rightarrow i$  in  $\mathcal{G}(\mathcal{M})$  for  $i \in \mathcal{I}$  then  $i' \rightarrow i$  is in  $\mathcal{D}$ . Since the JCI framework can be used to learn (features of the) causal graph  $\mathcal{G}(\mathcal{M})$  from data, this means that we can thereby learn the selection diagram from data. It is not clear how a selection diagram could be used to represent the same information that a JCI SCM with multiple context variables can represent. Indeed, even though the selection diagram has multiple selection variables, it is still modeling only two contexts, corresponding with just a single binary context variable in the JCI framework.

## 4. Experiments

In this section we report on the experiments we performed with JCI, comparing various implementations of the approach with several baselines and state-of-the-art causal discovery methods. We experimented both with simulated data with perfectly known ground truth and with real-world data where the ground truth is only known approximately. We will make the source code to reproduce our experiments available under a free and open source license.

### 4.1 Methods and baselines

In our experiments we study different implementations of JCI, based on two existing causal discovery algorithms: ASD (Hyttinen et al., 2014; Magliacane et al., 2016b; Forré and Mooij, 2018) and FCI (Spirtes et al., 1995; Zhang, 2008). The ASD algorithm is accurate but slow, while FCI is faster but less accurate due to its “greedy” approach. Another difference between both methods is that ASD can deal with partial inputs, while for FCI it is necessary to provide all independence test results it asks for. Although FCI (and possibly ASD) can deal with selection bias, we ignore this additional complication here and use simplified implementations that assume that there is no selection bias. For the cyclic case,



we used the adaptation of the ASD algorithm proposed by Forré and Mooij (2018) that replaces  $d$ -separation with its general cyclic generalization,  $\sigma$ -separation. Adapting FCI to the cyclic case seems less straightforward and is beyond the scope of this paper.

#### 4.1.1 ASD VARIANTS

For ASD, we implemented different variants as described in Table 5. The variants `ASD-obs`, `ASD-pooled` and `ASD-pikt` use the original implementation of (Hyttinen et al., 2014), but with the weights and query method of (Magliacane et al., 2016b). `ASD-obs` only uses the observational context and ignores data from the other contexts, `ASD-pooled` pools data from all contexts but does *not* add context variables, and `ASD-pikt` uses data from all contexts and assumes that the contexts correspond to perfect surgical interventions with known targets. Inspired by the approach of Tillman (2009) and Tillman and Spirtes (2011), we also implemented a variant `ASD-meta` that uses Fisher’s method as a “meta-analysis” method to combine the  $p$ -values of conditional independence tests performed with data from each context separately into a single overall  $p$ -value, which was then used as input for ASD. We use these implementations as state-of-the-art causal discovery methods for comparison.

For the JCI approach, we study several variants that differ in terms of which JCI assumptions they make, whether they also test conditional independences between context variables, whether all context variables are used or whether they are first merged into a single context variable, and whether the intervention targets of the context variables are considered to be known or not. The different implementations are described in detail in Table 5. The “CI Tests” column describes what conditional independence test are performed, and how, and can have the following values:

- A** use all variables, including context variables; the conditional independence tests performed are of the form  $\tilde{X}_a \perp\!\!\!\perp \tilde{X}_b \mid \tilde{\mathbf{X}}_S$  with  $\{a\} \cup \{b\} \cup S \subseteq \mathcal{I} \cup \mathcal{K}$  and  $\{a\}, \{b\}, S$  mutually disjoint.
- S** use only system variables; the conditional independence tests performed are of the form  $X_a \perp\!\!\!\perp X_b \mid \mathbf{X}_S$  with  $\{a\} \cup \{b\} \cup S \subseteq \mathcal{I}$  and  $\{a\}, \{b\}, S$  mutually disjoint.
- SS** system variables only, separately for each context; the conditional independence tests performed are of the form  $X_a \perp\!\!\!\perp X_b \mid \mathbf{X}_S, \mathbf{C} = \mathbf{c}$  for  $\{a\} \cup \{b\} \cup S \subseteq \mathcal{I}$  and  $\{a\}, \{b\}, S$  mutually disjoint, for all values  $\mathbf{c} \in \mathcal{C}$  with  $\mathbb{P}(\mathbf{c}) > 0$ . Note that this method assumes that the context domain  $\mathcal{C}$  is discrete.
- SF** system variables only, separately for each context, using Fisher’s method;<sup>10</sup> the conditional independence tests performed are of the form  $X_a \perp\!\!\!\perp X_b \mid \mathbf{X}_S, \mathbf{C} = \mathbf{c}$  for  $\{a\} \cup \{b\} \cup S \subseteq \mathcal{I}$  and  $\{a\}, \{b\}, S$  mutually disjoint, for all values  $\mathbf{c} \in \mathcal{C}$  with  $\mathbb{P}(\mathbf{c}) > 0$ . Note that this method assumes that the context domain  $\mathcal{C}$  is discrete.
- NC** test all variables, except for (conditional) independences between context variables; the conditional independence tests performed are of the form  $X_a \perp\!\!\!\perp \tilde{X}_b \mid \tilde{\mathbf{X}}_S$  with  $a \in \mathcal{I}$ ,  $\{b\} \cup S \subseteq \mathcal{I} \cup \mathcal{K}$  and  $\{a\}, \{b\}, S$  mutually disjoint.

10. Fisher’s method (Fisher, 1925) aggregates  $N$  independent  $p$ -values  $\{p_i\}_{i=1}^N$  by computing the  $p$ -value for the statistic  $F := -2 \sum_{i=1}^N \log p_i$ , which has a  $\chi^2$  distribution with  $2N$  degrees of freedom if all  $N$   $p$ -values  $p_i$  are independent.

**PF** test all pairs of context and system variables, conditioning on the remaining context variables; the conditional independence tests performed are of the form  $X_i \perp\!\!\!\perp C_k \mid \mathbf{C}_{\mathcal{K} \setminus \{k\}}$  for  $i \in \mathcal{I}$ ,  $k \in \mathcal{K}$ .

**ASD-JCI123sc** and **ASD-JCI1sc** both use a single merged context variable  $\mathbf{C} = (C_k)_{k \in \mathcal{K}}$ , whereas the other ASD-JCI variants use all context variables  $\{C_k\}_{k \in \mathcal{K}}$  as separate variables. The background knowledge for all ASD-JCI variants is a subset of the three JCI Assumptions 1, 2 and 3. The only ASD-JCI variant that uses background knowledge on intervention targets is **ASD-JCI123kt** (but it does not make any assumptions on the *type* of the intervention).

#### 4.1.2 FCI VARIANTS

We implemented different variants of the FCI algorithm by adapting the implementation in the R package `pcaIlg` (Kalisch et al., 2012). We used the default configuration, i.e., an order-independent (“stable”) skeleton phase, and no conservative or majority rule modifications (Colombo and Maathuis, 2014). For simplicity, we assumed that no selection bias is present, which means that the rules  $\mathcal{R}5$ – $\mathcal{R}7$  in Zhang (2008) can be ignored in the FCI algorithm, and only PAGs without undirected edges need to be considered. We consider two variants of FCI as the current state-of-the-art: **FCI-obs**, which uses only the observational context, and **FCI-pooled**, which uses pooled data from all contexts but does *not* add context variables. We also implemented a “meta-analysis” approach **FCI-meta** that uses Fisher’s method to combine the  $p$ -values from separate contexts into overall  $p$ -values that are used as input for the FCI algorithm. Finally, we have three variants (**FCI-JCI123**, **FCI-JCI1** and **FCI-JCI0**) referring to the JCI adaptation of the FCI algorithm as described in Section 3.7.4, for three different combinations of JCI Assumptions.

Note that the output of FCI is a PAG. Here, we will not evaluate the PAG itself, but rather the presence or absence of ancestral relations that can be identified from the PAG. This also allows us to bootstrap FCI predictions.

#### 4.1.3 LCD VARIANTS

The LCD implementation simply iterates over all context variables and ordered pairs of system variables and tests for the LCD pattern. As conditional independence test we test whether the partial correlation vanishes. As confidence measure for an LCD pattern  $\langle C, X, Y \rangle$ , we use  $-\log p_{C \perp\!\!\!\perp Y}$ . Note that LCD predicts the presence of an ancestral relation, the absence of a confounder between  $X$  and  $Y$ , and the absence of a direct causal effect of  $C$  on  $Y$ .<sup>11</sup>

#### 4.1.4 ICP VARIANTS

We also compare with ICP using the ICP function in the R package `InvariantCausalPrediction` (Peters et al., 2016).<sup>12</sup> Under the assumption of causal sufficiency, ICP returns direct causal

11. The absence of a bidirected edge  $X \leftrightarrow Y$  in the marginalization of the functional graph  $\mathcal{G}$  on  $\{C, X, Y\}$  implies also the absence of the bidirected edge  $X \leftrightarrow Y$  in the full graph  $\mathcal{G}$ . Furthermore, the absence of the direct edge  $C \rightarrow Y$  in this marginalization implies the absence of the direct edge  $C \rightarrow Y$  in the full graph  $\mathcal{G}$ .

12. We used the default arguments, except that we set `stopIfEmpty` to `TRUE`.

relations. However, in our setting, in which causal sufficiency cannot be assumed, ICP will generally return ancestral causal relations, as we proved in Corollary 20. Note that ICP assumes a single context variable, whereas we typically have multiple context variables in our data. One way to deal with this is to merge all context variables before the pooled data is input to ICP, which we do in ICP-sc. Another way is to run ICP on each context variable separately, hiding the other context variables when feeding the pooled data to ICP, and finally merging all predictions. This is done in ICP-mc.

The conditional independence tests that are performed by ICP are of the form  $\mathbf{C} \perp\!\!\!\perp X_i \mid \mathbf{X}_S$  with  $S \subseteq \mathcal{I} \setminus \{i\}$ , for  $i \in \mathcal{I}$ . The default conditional independence test used in ICP first linearly regresses  $X_i$  on  $S$  for each context  $\mathbf{C} = \mathbf{c}$  individually, and once globally. It then tests whether there exists a context  $\mathbf{c}$  in which the mean or the variance of the regression residuals is different from the global mean or variance. All  $p$ -values are then combined using a Bonferroni correction. Note that this test can detect more conditional dependences than a simple partial correlation test, as it also considers the variation between the variances across contexts.

#### 4.1.5 FISHER’S TEST FOR CAUSALITY

This is a very simple and immensely popular baseline in which we simply go through all pairs  $(i, k)$  of a system variable  $i \in \mathcal{I}$  and a context variable  $k \in \mathcal{K}$ , perform the conditional independence test  $X_i \perp\!\!\!\perp C_k \mid \mathbf{C}_{\mathcal{K} \setminus \{k\}}$  on the pooled data resulting in  $p$ -value  $p_{ik}$ . As confidence value for the ancestral causal relation  $k \in \text{AN}_{\mathcal{G}}(i)$  we report  $-\log p_{ik} + \log \alpha$ , where  $\alpha$  is the threshold for the independence test.

#### 4.1.6 BOOTSTRAPPING

A simple way to improve the stability of causal discovery algorithms is bootstrapping. For a method that outputs a confidence measure for a certain prediction we simply average the confidence measures over bootstrap samples. For FCI, as confidence measure we simply take a  $\{-1, 0, 1\}$ -valued variable encoding the identifiable absence/unidentifiability/identifiable presence of an ancestral relation. For LCD, we average  $-\log p_{\mathbf{C} \perp\!\!\!\perp Y}$  over bootstrap samples. For ICP, we similarly average the negative logarithm of the  $p$ -values for the discovered ancestral causal relations. We do not bootstrap ASD variants because of the high computational complexity. In our experiments, we use 100 bootstrap samples. Bootstrapped methods are indicated with a suffix “-bs”.

#### 4.1.7 CONDITIONAL INDEPENDENCE TESTS

Using an appropriate conditional independence test is important to obtain good causal discovery results. In this work we will use two different conditional independence tests, both relying on the assumption that the context variables are discrete and the system variables have a multivariate Gaussian distribution given the context.

The default conditional independence test that we used is the following. For testing  $\tilde{X}_a \perp\!\!\!\perp \tilde{X}_b \mid \tilde{\mathbf{X}}_S$ , we distinguish two cases:

- $S \cap \mathcal{K} = \emptyset$ : the test then reduces to a standard partial correlations test.

Table 5: Variants of implemented JCI algorithms and baselines used in our experiments. Assumption 0 is always assumed. The meaning of “CI Tests” is (more detailed explanation in main text): A: use all variables, including context variables; S: use only system variables; SS: system variables only, separately for each context; SF: system variables only, separately for each context, using Fisher’s method to combine them into a single  $p$ -value; NC: test all variables, except for (conditional) independences between context variables; PF: test all pairs of context and system variables, conditioning on the remaining context variables. The meaning of the superscript \* is explained in Section 4.1.7. Bootstrapped versions of methods will be indicated with a suffix “-bs” (and have been omitted from this table for clarity).

Name	Data	Context variables	JCI Assumptions			Intervention type	CI Tests
			1	2	3		
Baselines							
ASD-obs	observational	none	-	-	-	none	S
ASD-pooled	pooled	none	-	-	-	any	S
ASD-meta	all	none	-	-	-	any	SF
ASD-pikt	all	none	-	-	-	perfect, known targets	SS
FCI-obs	observational	none	-	-	-	none	S
FCI-pooled	pooled	none	-	-	-	any	S
FCI-meta	all	none	-	-	-	any	SF
JCI Implementations							
ASD-JCI0	pooled	all	-	-	-	any	A
ASD-JCI1	pooled	all	+	-	-	any	A
ASD-JCI12	pooled	all	+	+	-	any	A
ASD-JCI123	pooled	all	+	+	+	any	NC
ASD-JCI123kt	pooled	all	+	+	+	any, known targets	NC
FCI-JCI123	pooled	all	+	+	+	any	NC
FCI-JCI1	pooled	all	+	-	-	any	A
FCI-JCI0	pooled	all	-	-	-	any	A
LCD-sc	pooled	single (merged)	+	-	-	any	A*
ICP-sc	pooled	single (merged)	+	-	-	any	A*
ASD-JCI1-sc	pooled	single (merged)	+	-	-	any	A*
ASD-JCI123-sc	pooled	single (merged)	+	+	+	any	A*
LCD-mc	pooled	all (one-by-one)	+	-	-	any	A
ICP-mc	pooled	all (one-by-one)	+	-	-	any	A*
Fisher	pooled	all (one-by-one)	+	+	-	any	PF

- Otherwise, we go through all observed values  $\mathbf{c}_{S \cap \mathcal{K}}$  of  $\mathbf{C}_{S \cap \mathcal{K}}$ , and use a standard partial correlations test to calculate a  $p$ -value for  $\tilde{X}_a \perp\!\!\!\perp \tilde{X}_b \mid \tilde{\mathbf{X}}_{S \setminus \mathcal{K}}, \mathbf{C}_{S \cap \mathcal{K}} = \mathbf{c}_{S \cap \mathcal{K}}$ . We then aggregate the  $p$ -values corresponding to observed values of  $\mathbf{C}_{S \cap \mathcal{K}}$  into one overall  $p$ -value for  $\tilde{X}_a \perp\!\!\!\perp \tilde{X}_b \mid \tilde{\mathbf{X}}_S$  using Fisher’s method.

Note that for the case with zero context variables (i.e., a single context), this test reduces to a standard partial correlations test.

For the ICP implementations, we make use of the official implementation in the `InvariantCausalPrediction` package. This by default makes use of a conditional independence test that uses more than just partial correlations. We extended this conditional independence test to allow for conditioning on context variables, and make use of this extended test in all the algorithms that assume a single merged context, i.e., the ones marked with a \* in Table 5. It assumes that there is a single context variable, i.e.,  $|\mathcal{K}| = 1$ . For testing  $\tilde{X}_a \perp\!\!\!\perp \tilde{X}_b \mid \tilde{\mathbf{X}}_S$ , it distinguishes the following cases:

- If  $(\{a\} \cup \{b\} \cup S) \cap \mathcal{K} = \emptyset$ , it reduces to a standard partial correlations test.
- If  $\tilde{X}_a = \mathbf{C}$  is the context variable, it uses linear regression to fit  $\tilde{X}_b$  as a linear function of  $\tilde{\mathbf{X}}_S$ , using data pooled over all contexts, and calculates the corresponding residuals. It then goes through all observed values  $\mathbf{c}$  of  $\mathbf{C}$ , and tests whether the residuals in context  $\mathbf{c}$  have a different distribution than the residuals in the other contexts (i.e., with  $\mathbf{C} \neq \mathbf{c}$ ). This two-sample test is performed by comparing the means by a  $t$ -test, and the variance by an  $F$ -test. The two resulting  $p$ -values are combined with a Bonferroni correction. The resulting  $p$ -values, one for each context, are then also combined with a Bonferroni correction.<sup>13</sup>
- If  $\tilde{X}_b = \mathbf{C}$  is the context variable, proceed similarly as in the previous case.
- If the context variable  $\mathbf{C}$  is part of  $\tilde{\mathbf{X}}_S$ , we go through all observed values  $\mathbf{c}_{S \cap \mathcal{K}}$  of  $\mathbf{C}_{S \cap \mathcal{K}}$ , and use a standard partial correlations test to calculate a  $p$ -value for  $X_a \perp\!\!\!\perp X_b \mid \mathbf{X}_{S \setminus \mathcal{K}}, \mathbf{C}_{S \cap \mathcal{K}} = \mathbf{c}_{S \cap \mathcal{K}}$ . We then aggregate the  $p$ -values corresponding to observed values of  $\mathbf{C}_{S \cap \mathcal{K}}$  into one overall  $p$ -value for  $X_a \perp\!\!\!\perp X_b \mid \tilde{\mathbf{X}}_S$  using Fisher’s method.

As a final note regarding the conditional independence tests, we state that `ASD-pikt` uses a standard partial correlation test to calculate a  $p$ -value for each context separately. It internally subsequently combines all these  $p$ -values by adding the  $\log p$ -values, but taking into account how the graph structure is changed through surgical interventions.

The choice of the  $p$ -value threshold  $\alpha$  for rejecting the null hypothesis of independence is an important one. To obtain consistent results, one should let  $\alpha$  decrease to 0 with increasing sample size. In our experiments, we used fixed sample size and simply used a global threshold  $\alpha = 0.01$ .

## 4.2 Simulations

We simulated random linear-Gaussian SCMs with  $p$  system variables and  $q$  context variables. We considered both the acyclic setting and the cyclic one. We simulated stochastic

<sup>13</sup>. This is the default test in the ICP function in `InvariantCausalPrediction` for continuous data.

interventions of two different intervention types: mechanism changes, and perfect (surgical) interventions.

Random causal graphs were simulated by drawing directed edges independently between system variables with probability  $\epsilon$ . For the acyclic models, we only allowed directed edges  $i_1 \rightarrow i_2$  for  $i_1 < i_2$  with  $i_1, i_2 \in \mathcal{I}$ . For cyclic models, we allowed directed edges  $i_1 \rightarrow i_2$  for  $i_1 \neq i_2$  with  $i_1, i_2 \in \mathcal{I}$ , and subsequently selected only the graphs in which at least one cycle exists. We drew bidirected edges independently between all unordered pairs of system variables with probability  $\eta$ , and associated each bidirected edge with a separate latent confounding variable. For each context variable, we randomly selected a single system variable as its target, while ensuring that each system variable has at most one context variable as its direct cause. We sampled all linear coefficients between system variables, context variables and confounders from the uniform distribution on  $[-1.5, -0.5] \cup [0.5, 1.5]$ . The latent variables (“error terms”) were sampled independently from the standard-normal distribution. To ensure that system variables have comparable scales, we rescaled the weight matrix such that each system variable would have variance 1 if all its direct causes would be i.i.d. standard-normal.

We used binary context variables in a “diagonal” design. This means that for each random SCM, we simulated  $q+1$  contexts, with the first context being purely observational (i.e.,  $C_k = 0$  for all  $k \in \{1, \dots, q\}$ ), and the other  $q$  contexts corresponding with one of the context variables turned on (say  $C_{k'} = 1$  for some  $k' \in \{1, \dots, q\}$ ) and the others turned off ( $C_k = 0$  for the other  $k \in \{1, \dots, q\} \setminus \{k'\}$ ). We either took all interventions to be mechanism changes, or all interventions to be surgical. For mechanism changes, we simply add the value of the parent context variable to the structural equation (i.e., this corresponds with adding a constant offset of 1 to the intervention target variable when the intervention is turned on). For perfect interventions, we additionally set the linear coefficients of incoming edges on the intervention target to zero. Finally, we sampled  $N$  observed values of system variables from each context and combined all samples into one pooled data set per random SCM.

### 4.3 Evaluation

In evaluating the results, we consider different prediction tasks: establishing the absence or presence of ancestral causal relations between system variables, the absence or presence of direct causal relations between system variables, and the absence or presence of confounders. In addition, we consider predicting the absence or presence of indirect intervention targets (i.e., whether or not some context variable is ancestor of some system variable) and of direct intervention targets (i.e., whether or not some context variable is parent of some system variable).

The output of the FCI variants is a Partial Ancestral Graph (PAG), which requires further processing to read off ancestral relations, direct relations and confounders present in the Markov equivalence class represented by the PAG. We did not implement methods to calculate the identifiable (presence and absence of) direct relations, nor the (presence and absence of) confounders from the PAG, but focus here on the ancestral relations.

Each method outputs a confidence score for each feature of interest, where positive scores mean that it is more likely that a feature is present in the causal graph  $\mathcal{G}$ , whereas

negative scores mean that it is more likely that a feature is absent. The higher the absolute value of the score, the more likely its presence/absence is. The predictions are pooled both within model instances (e.g., all possible ancestral relations  $i \in \text{AN}_{\mathcal{G}}(j)$  for all ordered pairs of system variables  $i, j \in \mathcal{I}$ ) and across model instances to gather more statistics. The scores are then ranked and turned into ROC curves and PR curves (one PR curve for the presence, and one for the absence of the features) by comparing with the true features.

#### 4.4 Results: Small models

We first present results for small models with  $p = 4$  system variables and  $0 \leq q \leq 4$  (as a default,  $q = 2$ ) context variables. We used  $\epsilon = 0.5$ ,  $\eta = 0.5$ , and sampled  $N_c = 500$  samples for each context, i.e.,  $N = 500(q + 1)$  samples in total.

##### 4.4.1 ASD-JCI VS. BASELINES (CAUSAL MECHANISM CHANGES)

We start by showing off the advantage that JCI can offer over existing methods. We first consider only ASD variants because this most clearly shows the impact of how one merges data from different contexts and how one treats the context variables, since the other aspects of the causal discovery algorithm are the same for all ASD variants. In Figure 17 we present results for several ASD variants for acyclic models with causal mechanism changes. We compare the JCI variants `ASD-JCI123` (unknown intervention targets) and `ASD-JCI123kt` (known intervention targets) with the available baselines, `ASD-obs` (observational data only), `ASD-pooled` (pooled data from all context treated as if they were all observational, context variables not included), `ASD-meta` (using Fisher’s method to combine  $p$ -values from conditional independence tests in separate contexts), and `ASD-pikt` (which assumes that interventions are perfect and uses knowledge of intervention targets). The tasks of predicting ancestral causal relations and direct causal relations show relatively similar ROC and PR curves. Predicting the absence or presence of confounders is a more challenging task.

The three baselines `ASD-obs`, `ASD-pooled` and `ASD-meta` show very similar performance behaviors. In particular, for the tasks of predicting the presence of the features, these baselines perform poorly, not much better than random guessing. `ASD-pikt` even performs poorly on nearly all prediction tasks in this simulation setting because it incorrectly assumes that the interventions are perfect. The two JCI variants, on the other hand, strongly outperform the baselines and obtain very high precisions. In particular, even without knowing the intervention targets, `ASD-JCI123` manages to predict the presence of (direct and indirect) causal relations at maximum precision for low recall. Exploiting knowledge of the intervention targets, `ASD-JCI123-kt` obtains an even more impressive precision for predicting ancestral causal relations for a large range of recalls. This illustrates the big improvement in precision that JCI can yield.

Figure 18 shows a similar picture in the cyclic setting, where all ASD variants make use of  $\sigma$ -separation (Forré and Mooij, 2018). The task of predicting the presence of ancestral relations is easier than in the acyclic setting, because for most pairs of system variables, one is ancestor of the other due to the cycles. The task of predicting the absence of ancestral relations, on the other hand, is more challenging. Detecting the presence or absence of confounders in this setting has become nearly impossible, for any method. For the other

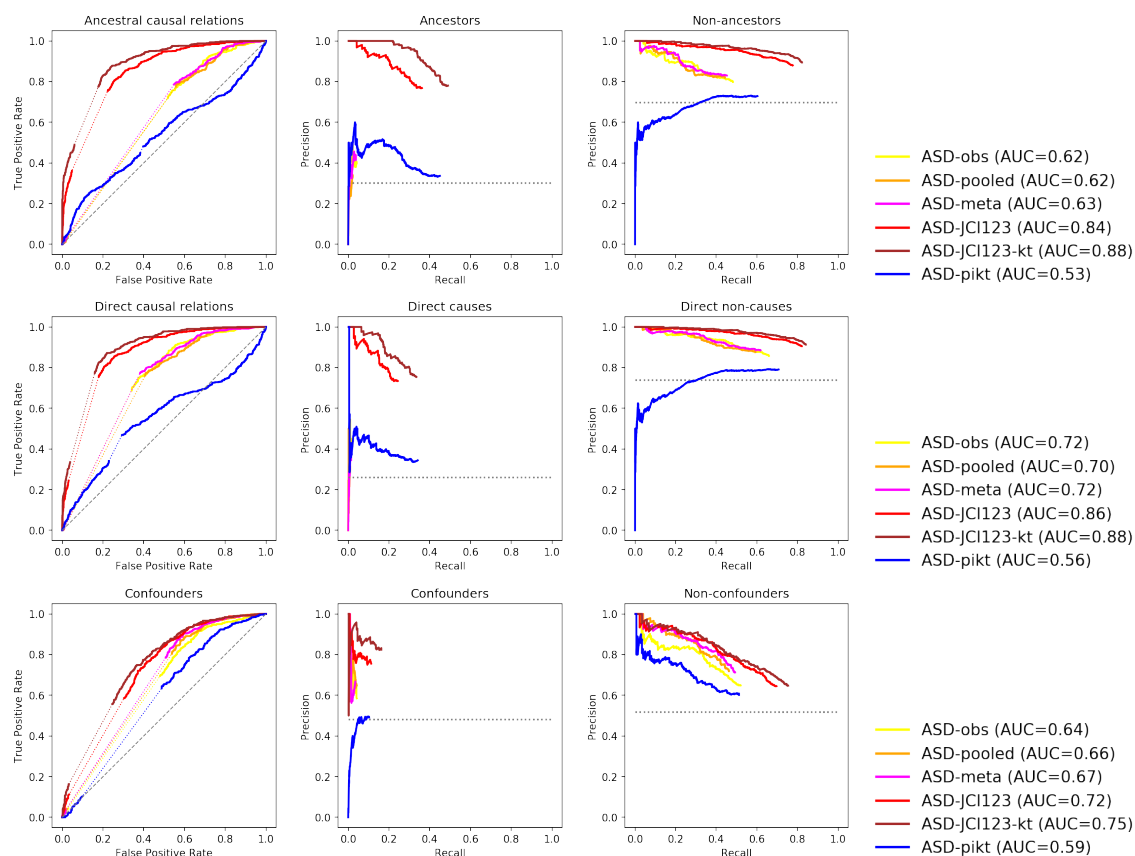


Figure 17: Results of some ASD variants (acyclic, causal mechanism changes). The two JCI variants strongly outperform the baselines in this setting.

tasks, the JCI approach again shows substantially improved precisions compared to the baselines.

#### 4.4.2 ASD-JCI VS. BASELINES (PERFECT INTERVENTIONS)

Figures 19 and 20 show results for respectively the acyclic and cyclic settings, but now for perfect (surgical) interventions with known targets rather than causal mechanism changes.

In these perfect intervention scenarios, the JCI variants again obtain a much higher precision than any of the baselines, with the sole exception of `ASD-pikt`. In this setting, the latter method successfully exploits the assumed surgical nature of the interventions, thereby outperforming the JCI variants that do not make any assumption about the nature of the intervention. However, a significant disadvantage of `ASD-pikt` in practice is that its assumption of perfect interventions with known targets may not be valid. As we already saw in Section 4.4.1, `ASD-pikt` then breaks down, in contrast to the JCI variants.



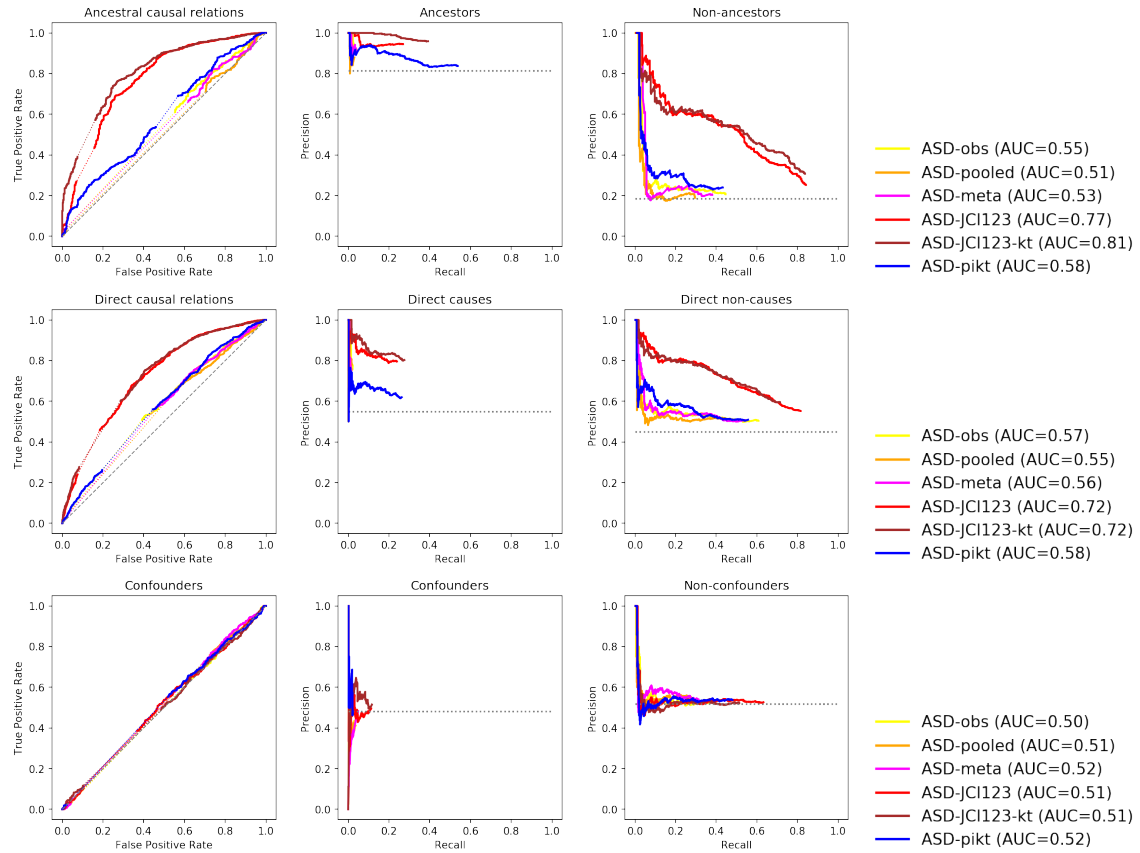


Figure 18: Results of some ASD variants (cyclic, causal mechanism changes). The two JCI variants substantially outperform the baselines in this setting.

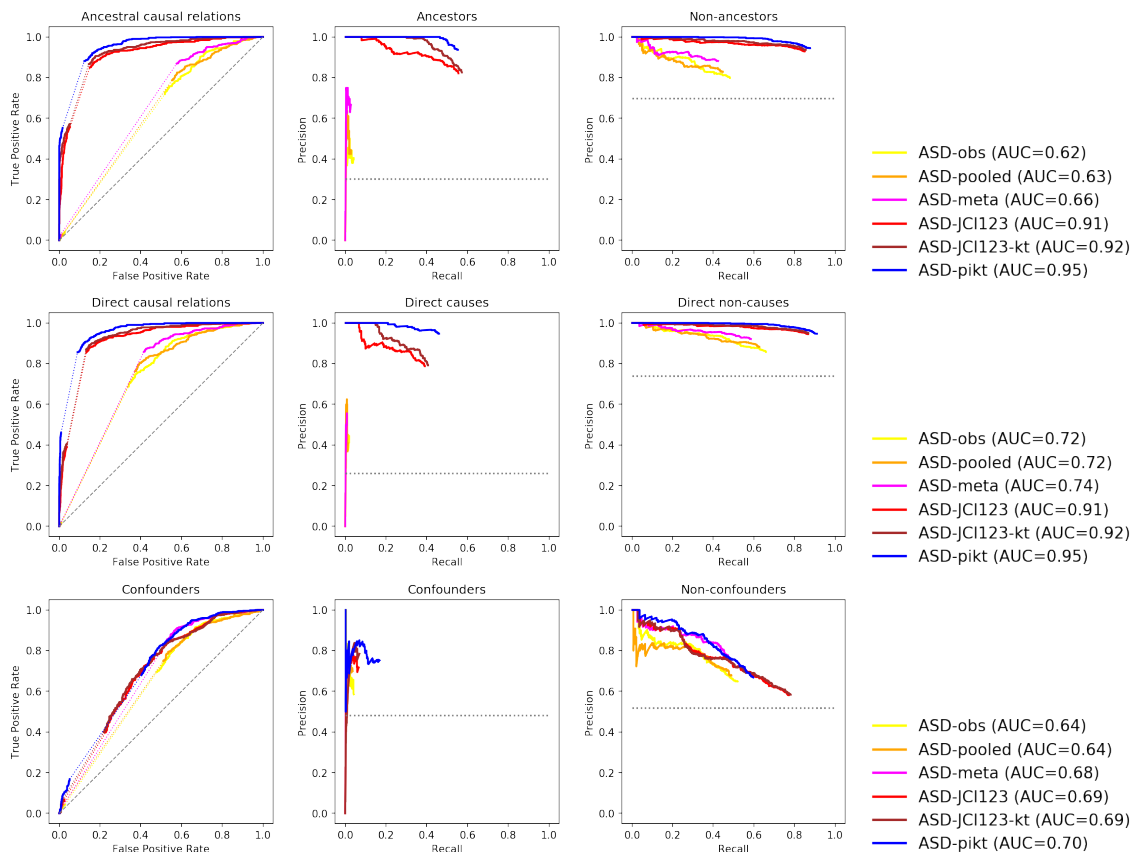


Figure 19: Results of some ASD variants (acyclic, surgical interventions). **ASD-pikt**, which takes into account the surgical nature of the interventions, is the best performing method in this setting. The JCI variants do not assume perfect interventions, but still yield a vast improvement of the precision of the predicted features with respect to the other baselines.

For the cyclic setting with perfect interventions (Figure 20), we observe that predicting the presence of confounders still seems impossible for all methods, but predicting their absence seems at least feasible in principle (although it seems a very challenging task).

#### 4.4.3 INFLUENCE OF JCI ASSUMPTIONS

We now investigate in more detail which JCI assumptions are responsible for the excellent performance of the JCI variants of ASD. Figure 21 shows that, as expected, the more prior knowledge about the context variables is used, the better the predictions become.

However, surprisingly, the biggest boost in precision with respect to the observational baseline is due to simply pooling the data and adding the context variables: **ASD-JCI0** already strongly improves over **ASD-obs**. Adding more background knowledge regarding the nature of the context variables helps to improve the results further. JCI Assumption 1

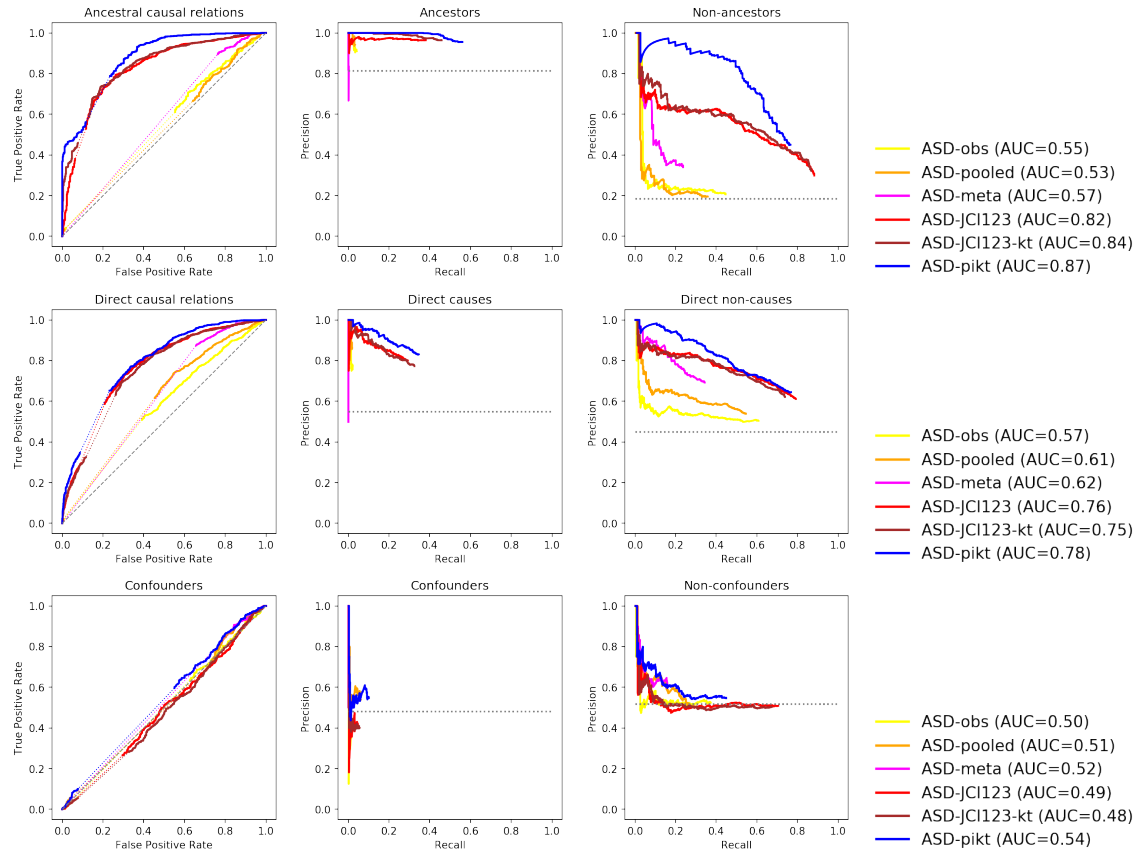


Figure 20: Results of some ASD variants (cyclic, surgical interventions). We see a similar picture as in the acyclic case in Figure 19.

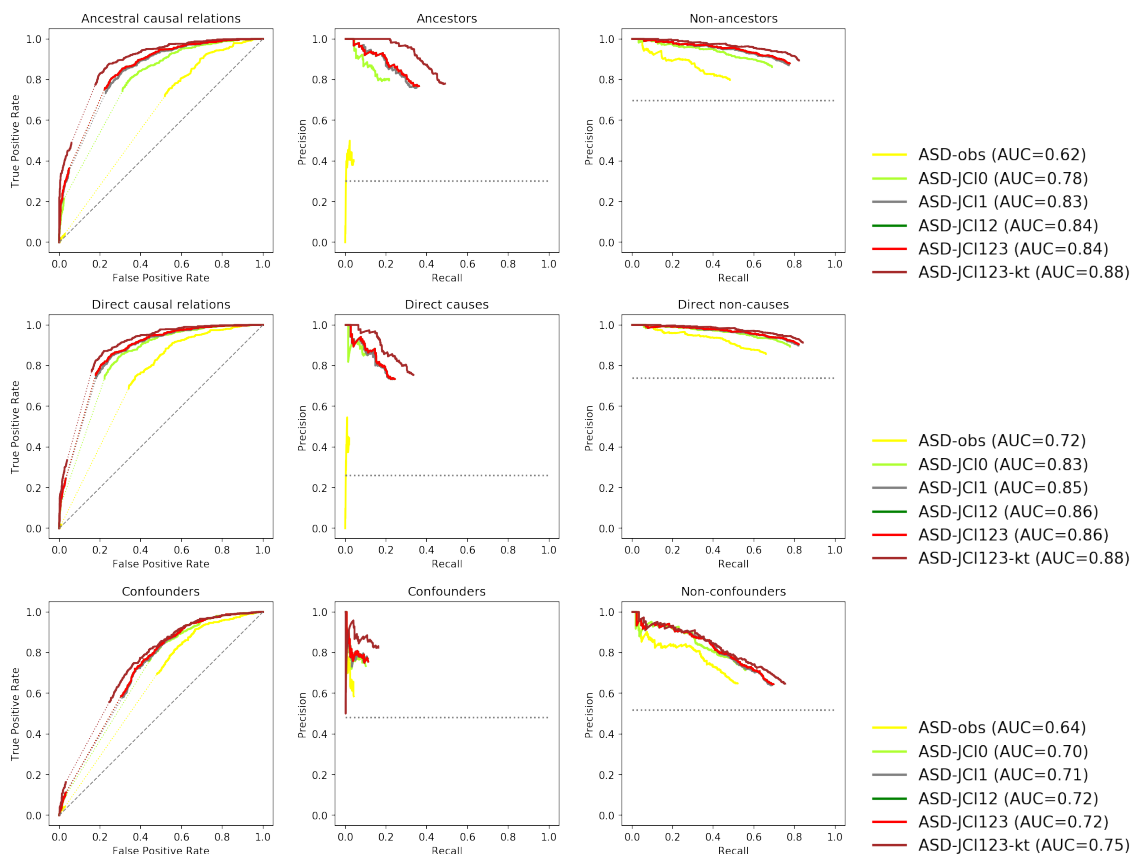


Figure 21: Influence of JCI assumptions: making use of more prior knowledge leads to better results (acyclic, causal mechanism changes). For reference, the non-JCI baseline ASD-obs is shown which only uses observational data.

yields a marginal improvement. JCI Assumptions 2 (and 3) do not lead to any further improvements for discovering the causal relations between system variables in this setting, though. Exploiting knowledge of the intervention targets, on the other hand, turns out to be very helpful for getting highly accurate predictions for ancestral relations between system variables, and also significantly improves the precision of predicting direct causal relations and confounders. We have shown here only the results for the acyclic setting with causal mechanism changes since we obtained similar results for the other simulation settings.

We also investigated variants of ASD-JCI0, ASD-JCI1 and ASD-JCI12 where we did not perform any independence tests on the context variables, i.e., using conditional independence testing scheme “NC” rather than “A”. This is possible because ASD is capable of handling incomplete inputs. We obtained almost identical results (in all simulation settings considered) to the standard variants of those methods in which we do perform conditional independence tests on the context variables themselves (not shown here).

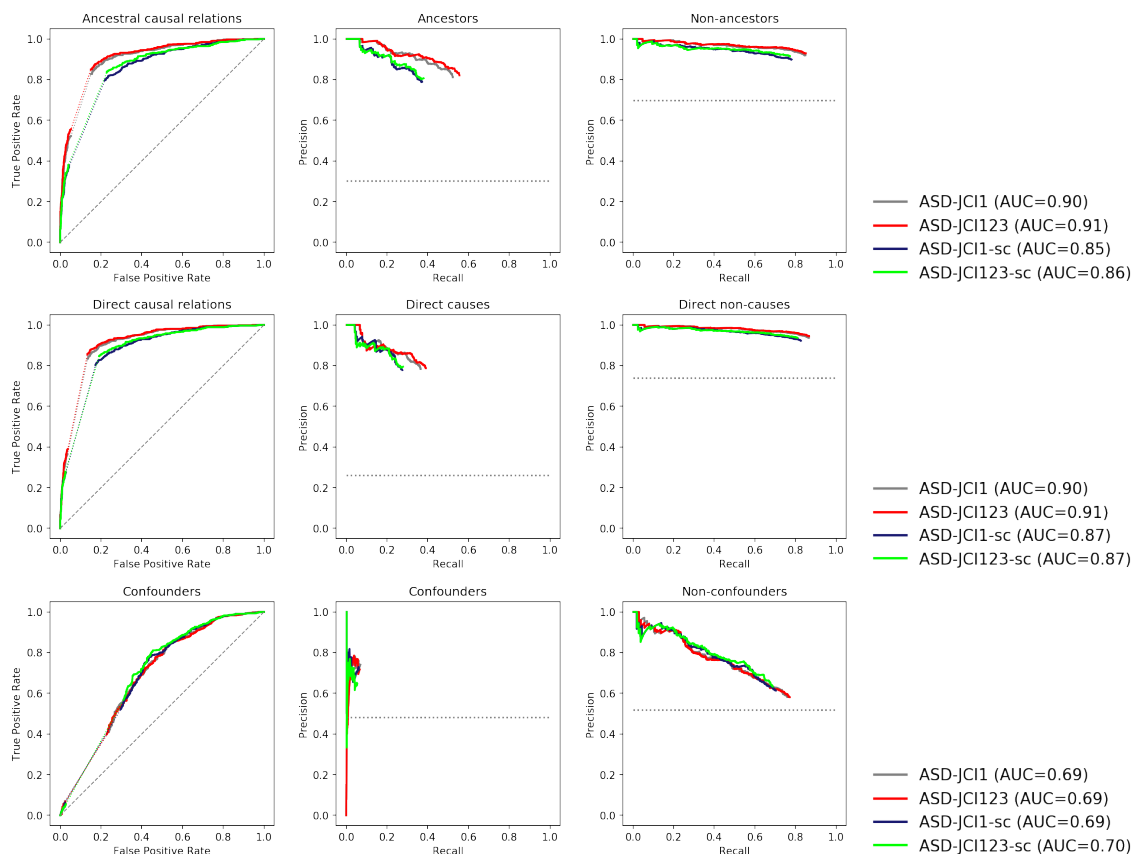


Figure 22: For ASD-JCI variants, exploiting multiple context variables leads to better results than using a single (merged) context variable (acyclic, perfect interventions).

#### 4.4.4 MULTIPLE CONTEXT VARIABLES VS. SINGLE (MERGED) CONTEXT VARIABLE

Figure 22 shows that for ASD-JCI variants, exploiting multiple context variables leads to better results than using a single (merged) context variable, as expected. Similar results hold for the cyclic settings and for causal mechanism changes (not shown).

#### 4.4.5 FCI VARIANTS

Figure 23 shows the results for the various FCI variants. FCI is seen to be somewhat less accurate than ASD, but bootstrapping helps to boost precision for lower recalls. Similarly to ASD, we conclude that the JCI variants of FCI substantially outperform the non-JCI variants. Interestingly, FCI-JCI1 and FCI-JCI123 seem to yield identical results in this setting.

Results for causal mechanism changes are very similar to those for surgical interventions, and therefore are not shown here. The results for the cyclic setting are also not shown, because FCI was designed for the acyclic setting.

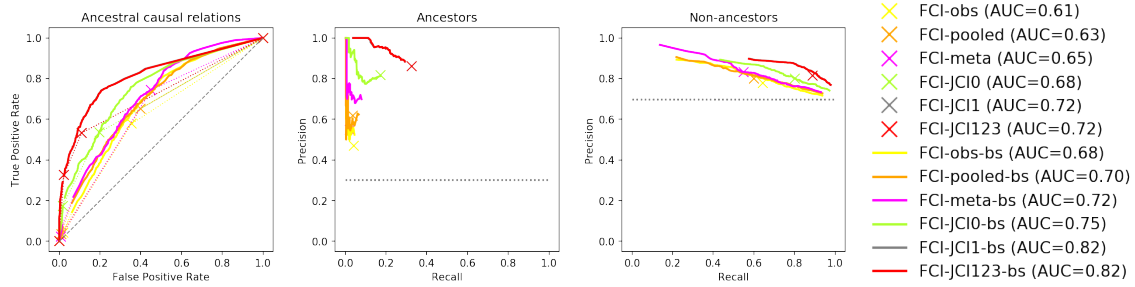


Figure 23: Results of FCI variants (acyclic, surgical interventions).

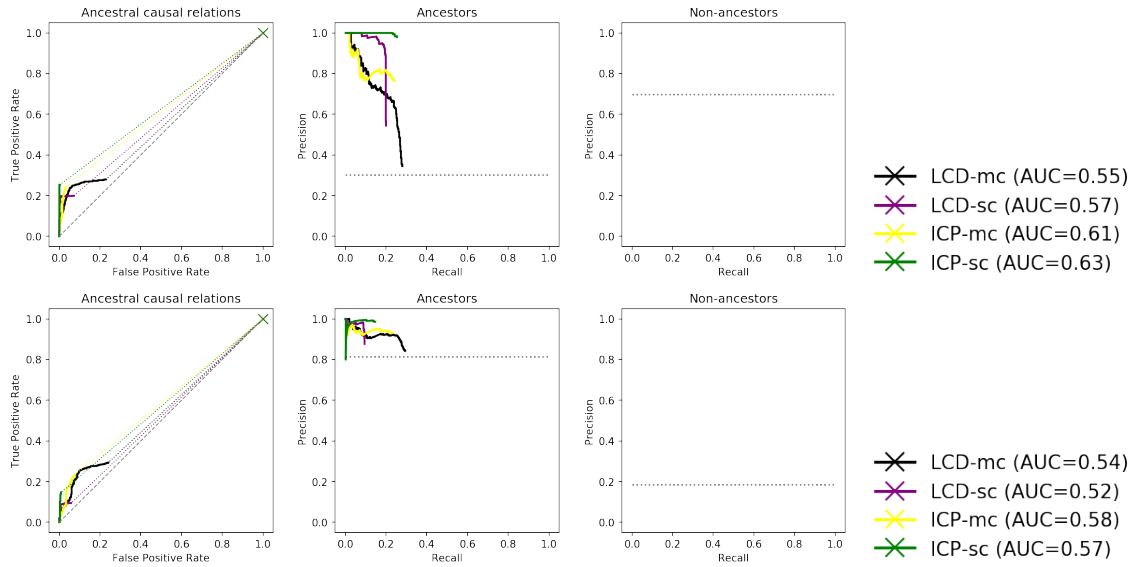


Figure 24: LCD and ICP benefit from merging the context variables into a single one (perfect interventions) for the task of predicting the presence of ancestral relations. Top: acyclic; bottom: cyclic.

#### 4.4.6 LCD AND ICP

Figure 24 shows the results of the LCD and ICP variants for the task of predicting ancestral relations. We only show the results for perfect interventions, as the results for causal mechanism changes are similar. LCD and ICP both only can predict the presence of ancestral relations, not their absence. LCD and ICP apparently benefit from merging the context variables into a single one. We speculate that a possible explanation of this phenomenon could be that a combination of conditional independence tests of the form  $C_k \perp\!\!\!\perp X_{i'} \mid X_i$  with each  $C_k$  binary (and both  $X_i$  and  $X_{i'}$  real-valued) might be less reliable than a single test  $C \perp\!\!\!\perp X_{i'} \mid X_i$  where  $C$  is categorical with  $\gg 2$  states. Another observation we made is that bootstrapping does lead to only marginal improvements for these methods (not shown).

## 4.4.7 VARYING THE NUMBER OF CONTEXT VARIABLES

As we have seen, discovery of causal relations between system variables can benefit strongly from observing the system in multiple contexts. As Figure 25 shows, the more context variables are taken into account, the better the predictions for ASD-JCI123 become. Although not shown here, the same conclusion holds as well for the other JCI variants ASD-JCI0, ASD-JCI1 and ASD-JCI123-kt. It does not hold for ASD baselines in general, but it does for ASD-pikt if interventions are perfect. For the JCI variants of FCI, FCI-JCI0, FCI-JCI1 and FCI-JCI123 we also observed that the more context variables are available, the more accurate the predictions become.

On the other hand, the same conclusion does not hold for ASD-JCI1-sc and ASD-JCI123-sc. This suggests that having multiple contexts is mostly beneficial if each context variable targets only a small subset of system variables, and only for methods that can explicitly take into account multiple context variables. For LCD and ICP, precision also does not improve monotonically with the number of context variables. Although LCD-mc and LCD-mc in principle allow for multiple context variables, they suffer from a drop in recall because they focus on detecting a certain causal pattern that becomes increasingly rare with more context variables. Indeed, for the extreme case  $q = p$ , each system variable is targeted by a single context variable in our simulation setting, and hence one would expect LCD-mc and ICP-mc to make no predictions at all. Any predictions they make must therefore be false positives, resulting in low precision.

## 4.4.8 LEARNING INDIRECT INTERVENTION TARGETS

Figure 26 shows for the acyclic setting with causal mechanism changes how accurately *indirect* intervention targets (i.e., which system variables are descendants of each context variable?) can be discovered by various methods. Baselines ASD-obs, ASD-pooled, ASD-meta and ASD-pikt cannot learn intervention targets (neither direct ones nor indirect ones), since they do not represent context variables explicitly, and are therefore excluded.<sup>14</sup> LCD and ICP also cannot address this task. Although ASD-JCI123-kt makes use of known *direct* intervention targets (i.e., which system variables are children of each context variable?), this means that there is still a non-trivial task of learning the *indirect* ones.

The task of deciding that a system variable is targeted is an easier one than deciding that a system variable is *not* targeted by an intervention. Although Fisher’s test is generally hard to beat when it comes to predicting indirect intervention targets, ASD-JCI123-kt outperforms it in this setting by exploiting the knowledge about *direct* intervention targets. While JCI Assumption 2 turned out to be unimportant for learning the causal relations between system variables, it is seen to be very useful for this task of learning causal relations between context and system variables.

Figure 27 shows a largely similar picture for the cyclic setting with causal mechanism changes. Surprisingly, FCI variants are also performing quite well in the cyclic setting. We do not show the results for perfect interventions here, as we observed that this task is easier,

14. However, it would be trivial to extend ASD-pikt such that it can predict indirect intervention targets, by combining the known direct intervention targets of a certain intervention variable with the descendants of those assumed targets as predicted by the method as a postprocessing step.

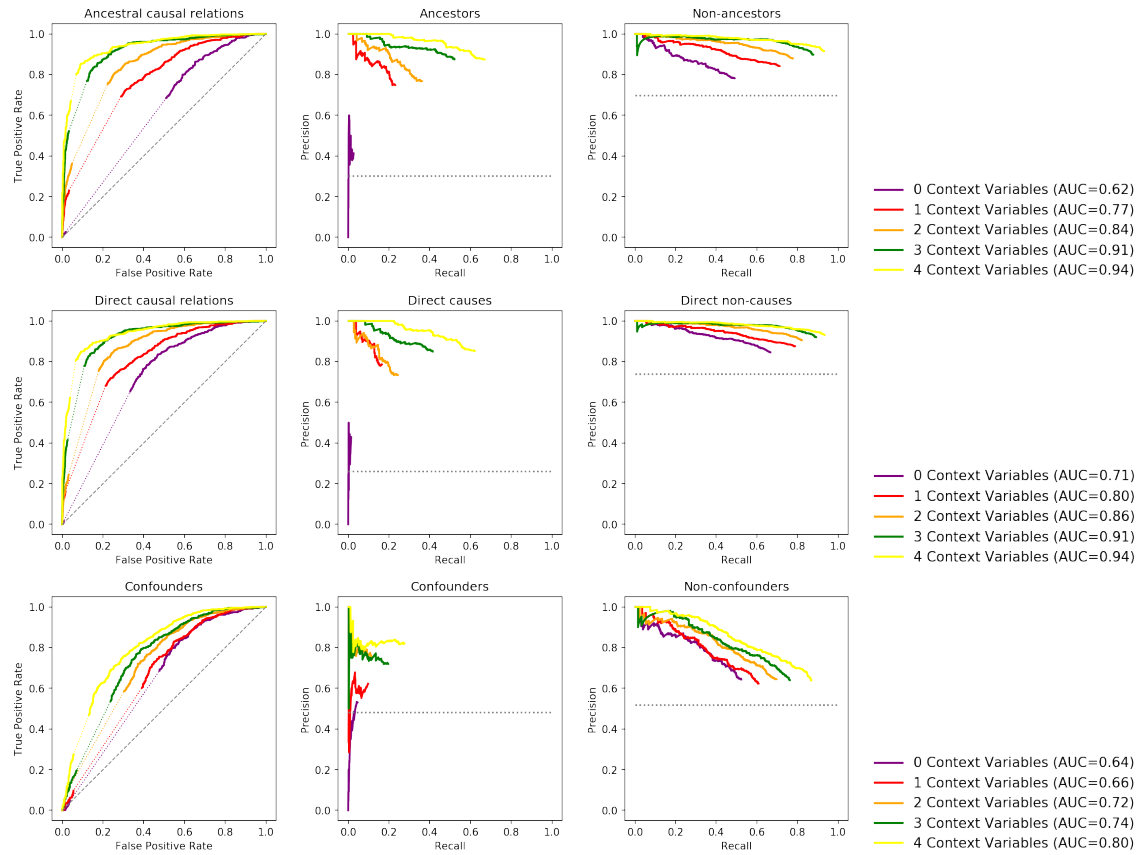


Figure 25: Taking into account more context variables leads to better predictions for ASD-JCI123 (acyclic, causal mechanism changes).



JOINT CAUSAL INFERENCE FROM MULTIPLE CONTEXTS

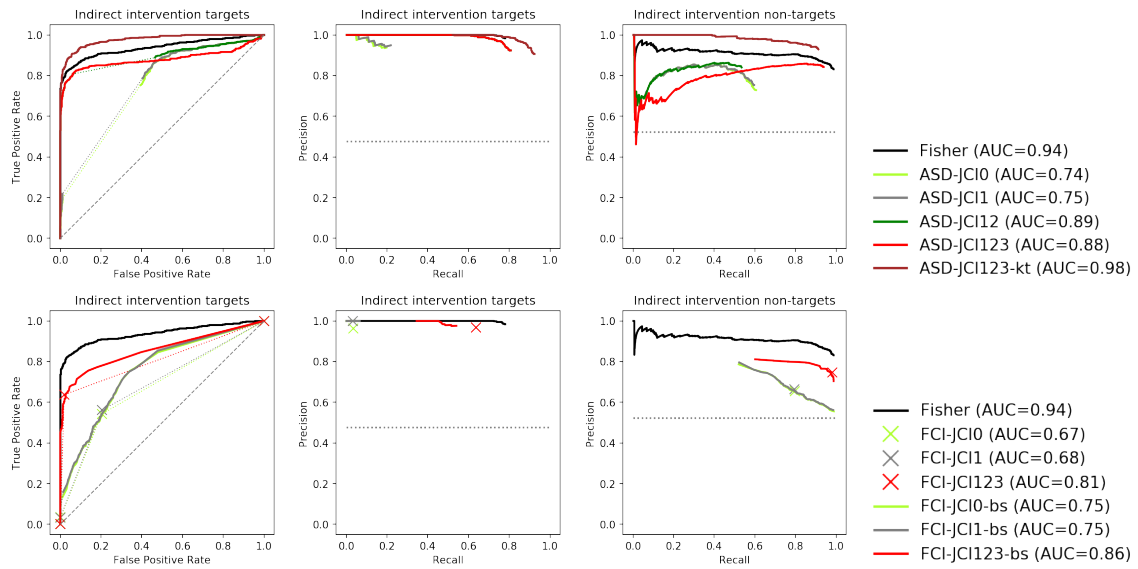


Figure 26: Learning indirect intervention targets (acyclic, causal mechanism changes). Top: ASD variants; Bottom: FCI variants.

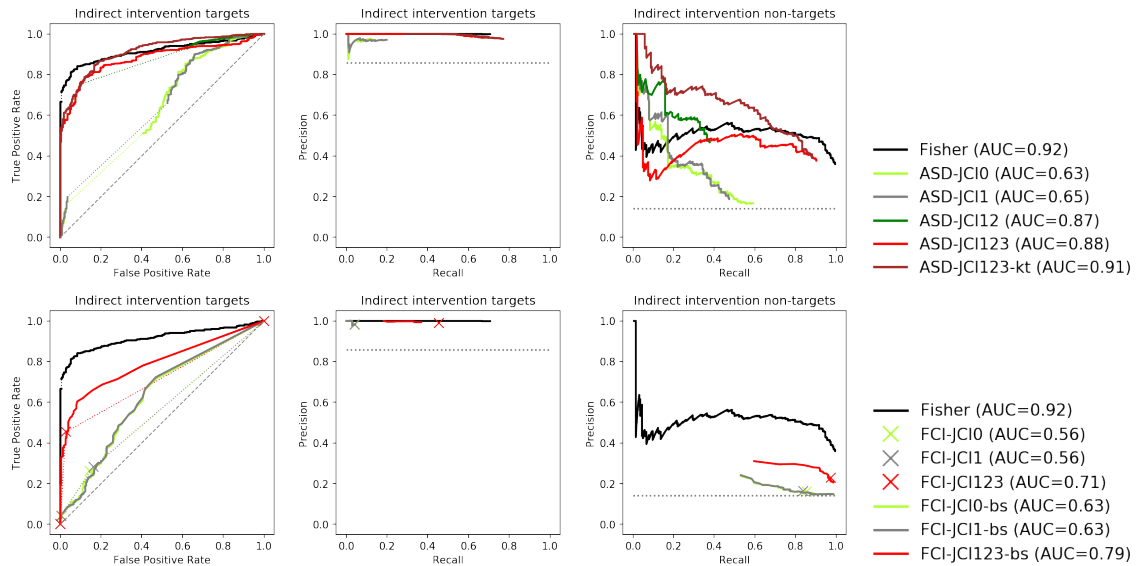


Figure 27: Learning indirect intervention targets (cyclic, causal mechanism changes). Top: ASD variants; Bottom: FCI variants.

and the results are generally better, but otherwise mostly similar conclusions are obtained. The only exception is that for perfect interventions, the best method is Fisher’s approach.

#### 4.4.9 LEARNING DIRECT INTERVENTION TARGETS

Fisher’s test is not able to predict *direct* intervention targets (i.e., which system variables are children of each intervention variable?), but the ASD-JCI variants can.<sup>15</sup> Figure 28 shows the results for the four different simulation settings. The task is considerably easier in the acyclic setting than in the cyclic setting. Having perfect interventions makes it slightly easier than with causal mechanism changes. We again notice that exploiting JCI Assumption 2 considerably improves performance on this task.

#### 4.4.10 COMPUTATION TIME

So far we have only considered the accuracy of the predictions. Another interesting aspect to study is the computation time that various methods need. Figure 29 shows total computation time (for all prediction tasks together) for all methods considered thus far. Note the logarithmic scale on the  $x$ -axis. We only show runtimes for causal mechanism changes since those for perfect interventions are nearly identical. On the other hand, we do see that the cyclic setting is more computationally demanding in general than the acyclic one.

Already for small models of  $p+q = 4+2 = 6$  variables, the ASD algorithms become slow because they are performing an optimization over a large discrete space. The availability of more background knowledge makes the search space considerably smaller, and hence leads to reduced computation time for the ASD variants. Also, the search space is considerably larger in the cyclic setting than in the acyclic one. By design, FCI variants are much faster, but bootstrapping also takes its toll. The fastest methods are Fisher’s test, LCD and ICP. Figure 30 shows how computation time scales with the number of context variables, for three JCI implementations (ASD-JCI123, ASD-JCI1 and FCI-JCI123).

### 4.5 Results: Larger models

We now present results for larger models, with  $p = 11$  system variables and  $q = 9$  context variables. We only consider causal mechanism changes here, but we do distinguish the acyclic and cyclic settings. We again used 500 samples per context. For the acyclic setting, we used  $\epsilon = \eta = 0.25$ , while for the cyclic setting we used  $\epsilon = \eta = 0.15$  to get more or less similarly dense graphs in both scenarios. The motivation for these parameter choices is that they mimic the setting of the real-world data set that we will study in the next subsection.

#### 4.5.1 FCI

Figure 31 shows the accuracy for the task of predicting ancestral causal relations between system variables for various JCI variants of FCI and of FCI baselines, in both the acyclic and the cyclic setting. The conclusions are in line with what we already observed for smaller models. Again, bootstrapping FCI helps considerably to boost the accuracy of its predictions. As before, FCI-obs (which uses only observational data) performs worst. The two baselines FCI-pooled and FCI-meta (that make use of all data) lead to a moderate improvement. The JCI variants (FCI-JCI0, FCI-JCI1 and FCI-JCI123) perform the best, delivering almost maximum precision for a considerable recall range on the task of predicting the presence of an ancestral relation. JCI Assumption 1 does not seem to help much, as

15. With proper postprocessing of the PAG, also FCI-JCI variants could.

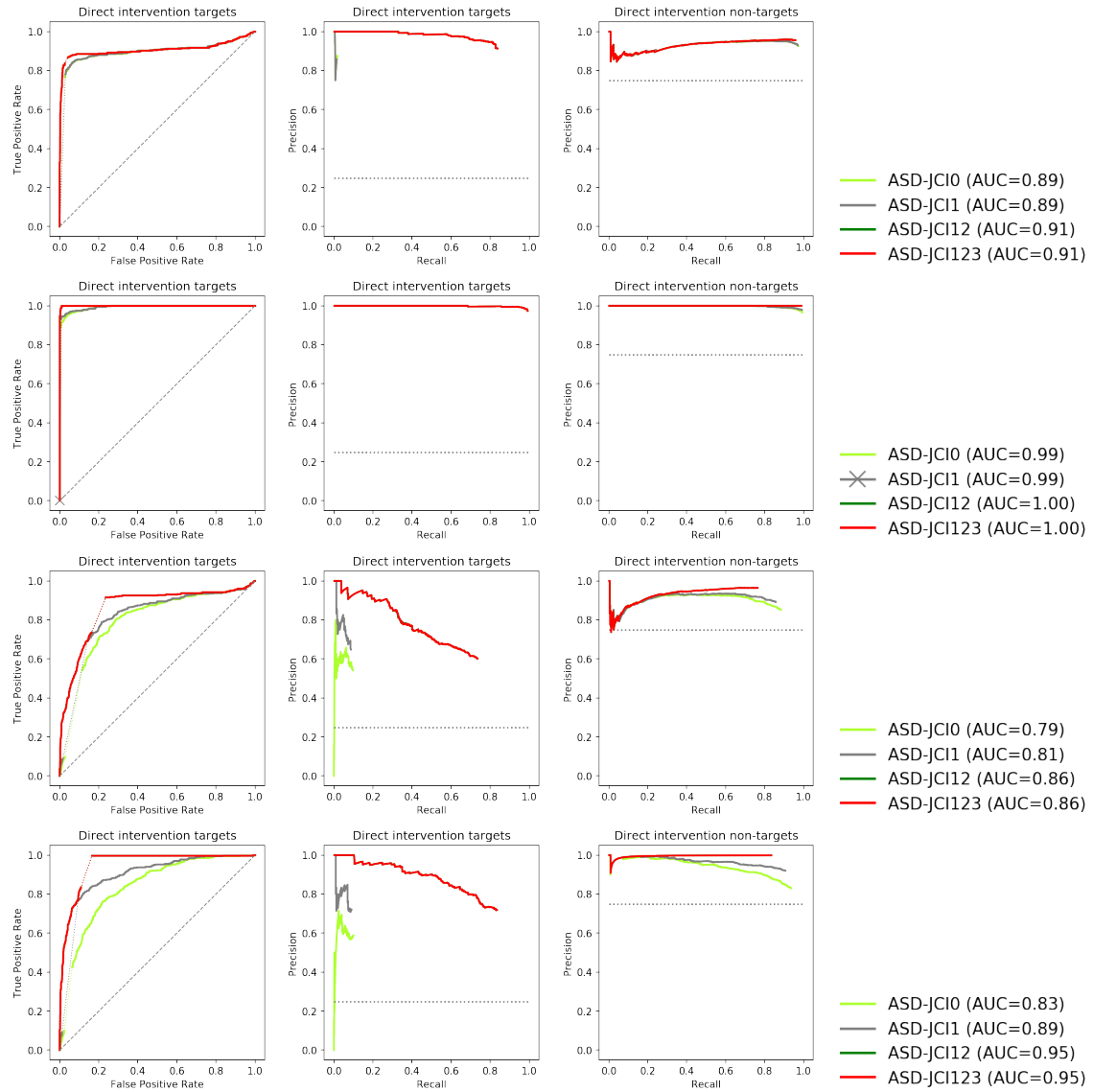


Figure 28: Learning direct intervention targets. From top to bottom: acyclic, causal mechanism changes; acyclic, perfect interventions; cyclic, causal mechanism changes; cyclic, perfect interventions.

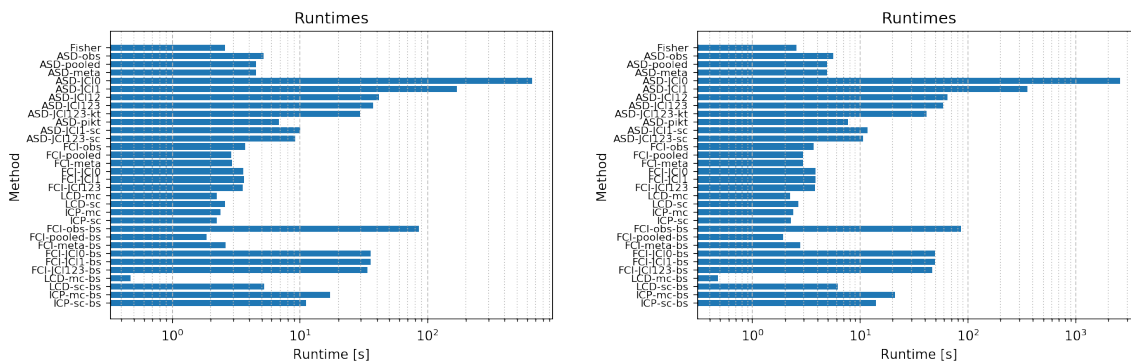


Figure 29: Runtimes for various methods (shown: causal mechanism changes; for perfect interventions, runtimes are similar). Left: acyclic; Right: cyclic.

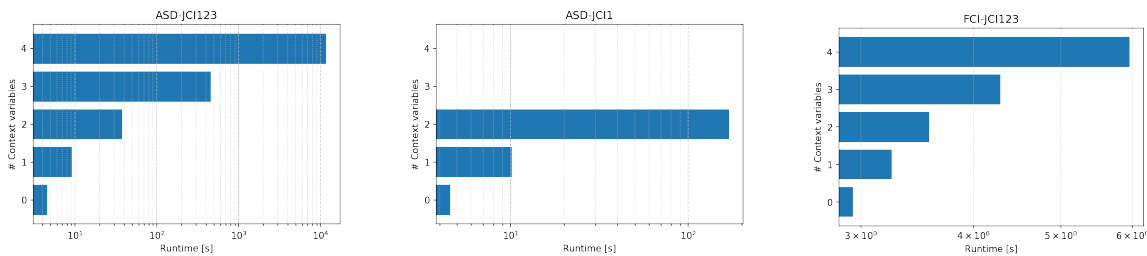


Figure 30: Runtimes for three different algorithms as a function of the number of context variables. Note the considerably different ranges of the (logarithmic)  $x$ -axis. Results are omitted if the computation took too long to finish.

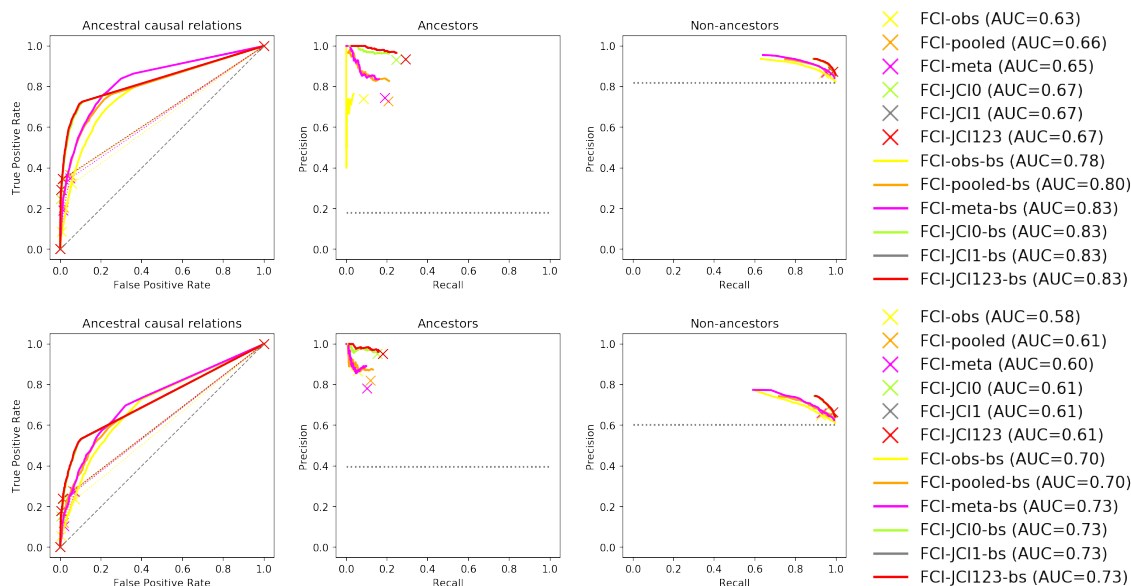


Figure 31: Accuracy of FCI results for discovering ancestral causal relations between system variables for acyclic (top) and cyclic (bottom) large models.

the results for FCI-JCI0 and FCI-JCI1 seem to be identical. Assuming in addition JCI Assumption 2 (and 3) does help to obtain slightly higher precision. The good performance of FCI variants in the cyclic setting is surprising, since FCI relies on the assumption of acyclicity.

#### 4.5.2 LCD AND ICP

In Figure 32, we show the accuracy of (bootstrapped) LCD and ICP for the task of predicting ancestral causal relations between system variables, in both the acyclic and the cyclic setting. As for FCI, bootstrapping improves the accuracy of LCD and ICP results, and we decided to only show the bootstrapped results here. The “multiple context” (“-mc”) versions of both algorithms clearly outperform the versions that use only a single (merged) context (“-sc”) in these settings. Interestingly, the accuracy of both methods is quite similar. The additional complexity of ICP apparently does not lead to substantially better results than the LCD algorithm already offers in these settings. Even more strikingly, the precision of LCD-mc is comparable to that of FCI-JCI123, the most accurate of the JCI variants of FCI (cf. Figure 31), on the task of predicting the presence of ancestral relations.

#### 4.5.3 LEARNING INDIRECT INTERVENTION TARGETS

Figure 33 shows the performance of JCI variants of FCI (and as a baseline, Fisher’s test) on the task of learning indirect intervention targets, for both the acyclic and cyclic settings. Interestingly, JCI Assumption 2 seems necessary to obtain good results on this task. Still, Fisher’s test outperforms FCI-JCI123-bs. Again, we conclude that with all its complexity,

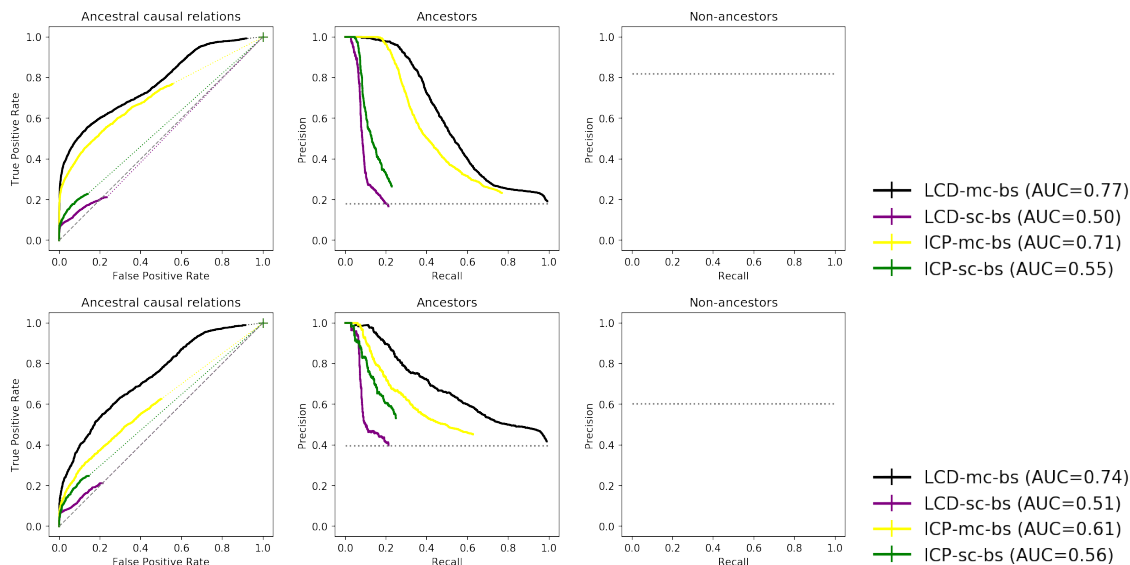


Figure 32: Accuracy of (bootstrapped) LCD and ICP results for discovering ancestral causal relations between system variables for acyclic (top) and cyclic (bottom) large models.

(JCI variants of) the FCI algorithm cannot beat a simple (JCI) baseline on a particular task.

#### 4.5.4 COMPUTATION TIME

Finally, Figure 34 shows runtimes of the methods that we considered. First, we observe no big differences between the runtimes for the cyclic setting with respect to the acyclic one. However, we do observe huge differences in runtime between various methods. LCD variants and Fisher’s test are by far the fastest. ICP variants come second. Bootstrapping puts a large toll on computation time for FCI variants. JCI variants of FCI are much slower than non-JCI variants. This seems to be mostly due to an exponential increase in the number of conditional independence tests. Indeed, we observed that FCI-JCI variants are conditioning on a substantial fraction of all  $2^9$  subsets of all context variables before finding a separating set. Nevertheless, JCI variants of FCI are still computationally feasible in this setting, even with bootstrapping.

## 4.6 Results: Real-world (flow cytometry) data

As an application on real-world data, we study the flow cytometry data of Sachs et al. (2005). This dataset has become a benchmark in causal discovery, even though the reliability and especially the completeness of its ground truth, the “consensus network” in (Sachs et al., 2005), is debated. Most causal discovery methods that have been applied on this data use the background knowledge about the intervention types and targets, which is specified in Table 3, with the notable exception of the method of Eaton and Murphy (2007). Using that

# JOINT CAUSAL INFERENCE FROM MULTIPLE CONTEXTS

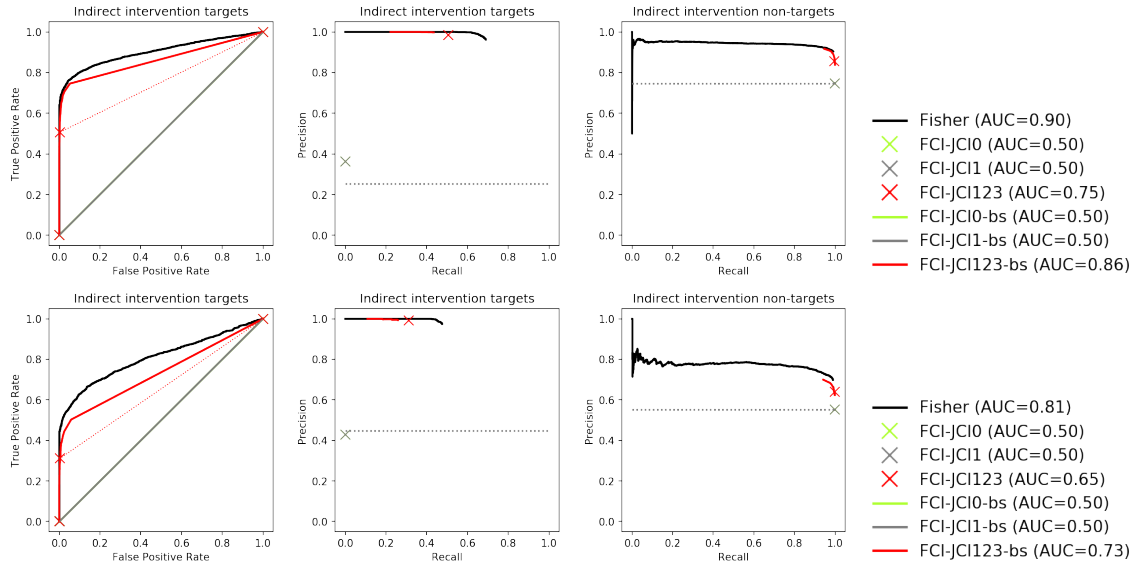


Figure 33: Accuracy of FCI results for discovering indirect intervention targets for acyclic (top) and cyclic (bottom) large models.

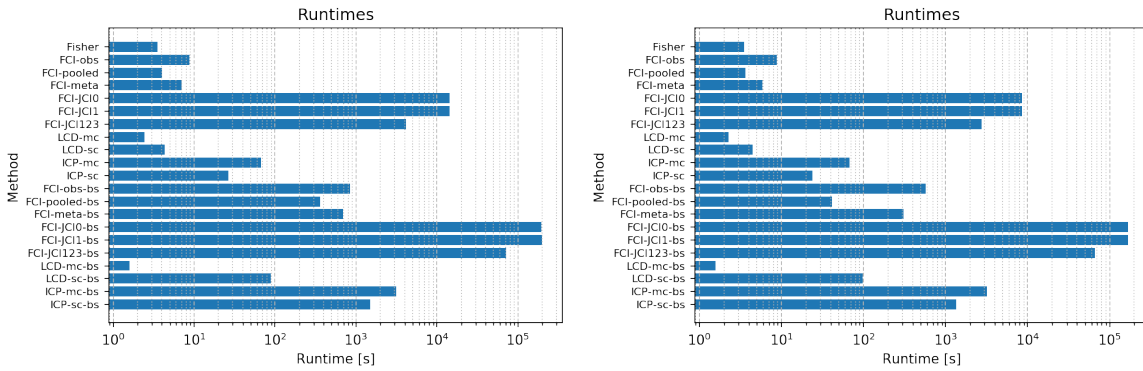


Figure 34: Runtimes for various methods on larger models. Left: acyclic; Right: cyclic.

background knowledge simplifies the causal discovery problem considerably. However, the accuracy of this background knowledge is not universally accepted.

We ran various FCI-JCI variants on a subset of the flow cytometry data.<sup>16</sup> The experimental design is described in Table 2. To enable comparisons with other results reported in the literature, we only used the 8 (out of 14) experimental conditions in the data set in which no ICAM.2 had been administered (i.e., with  $C_\beta = 0$ ), and ignored the others. In order to avoid deterministic relations between context variables, we merged context variables  $C_\alpha$  ( $\alpha$ -CD3/CD28),  $C_\theta$  (PMA) and  $C_\iota$  ( $\beta$ 2CAMP), as discussed in Section 3.6. This means that we must interpret the merged context variable as referring to the addition of PMA or  $\beta$ 2CAMP, combined with *omitting*  $\alpha$ -CD3/CD28.

Similarly to Eaton and Murphy (2007), we do not use background knowledge regarding intervention types or targets, but rather try to learn the intervention targets from the data itself. We only assume that the experimental setting is captured by the JCI framework. JCI Assumption 1 should be true because the intervention is performed some time (approximately 20 minutes) before the measurements are done. We have already discussed the validity of JCI Assumption 2 for this particular experimental setting in Section 3.4. Assuming that the context variables provide a complete causal description of the context (in particular, that there are no unintended batch effects), JCI Assumption 2 may apply. In that case, JCI Assumption 3 also applies since there are no (conditional) independences in the context distribution (after merging  $C_\alpha$ ,  $C_\theta$  and  $C_\iota$ ).

For comparison, we also ran FCI-obs, i.e., standard FCI using only the observational data set (i.e., the one in which only global activators  $\alpha$ -CD3 and  $\alpha$ -CD28 have been administered). We also ran FCI-meta, which uses Fisher’s method to combine  $p$ -values of conditional independence tests in the 8 separate experimental conditions, which are then used as input for standard FCI. Finally, we ran FCI-pooled, i.e., standard FCI on the 8 experimental conditions pooled together (but excluding context variables). In all those FCI variants, we assumed that no selection bias would be present. We additionally compare with other JCI implementations, in particular, multiple variants of LCD and ICP.

The “consensus network” and the PAGs obtained by the FCI baselines are shown in Figure 35. Figure 36 shows PAGs obtained by the FCI-JCI variants. Note that we show the PAGs obtained from using all data (i.e., without bootstrapping), although they are not necessarily stable. Therefore, we show in Figure 37 also the bootstrapped results for the learned ancestral relations between system variables and the learned (indirect) intervention targets. One can see here that most, but not all, of these features are stably predicted.

We draw the following conclusions from these results. First, the reference methods that exploit knowledge on intervention types and targets obtain rather consistent results. Second, the JCI methods that do not assume knowledge on intervention types and targets (and the approach by Eaton and Murphy (2007) show less consistent results. This could indicate that the data contain not enough signal in order to solve this more ambitious task reliably. If the “causal signal” is low, some model misspecification (for example, strong non-linearities or deviations from Gaussianity, which makes a simple partial correlation based conditional independence test inadequate) could lead to inconsistencies between methods. Nonetheless, the performance of the FCI-JCI variants seems a big improvement over the simple FCI

---

16. As preprocessing, we simply took the logarithm of the raw values.



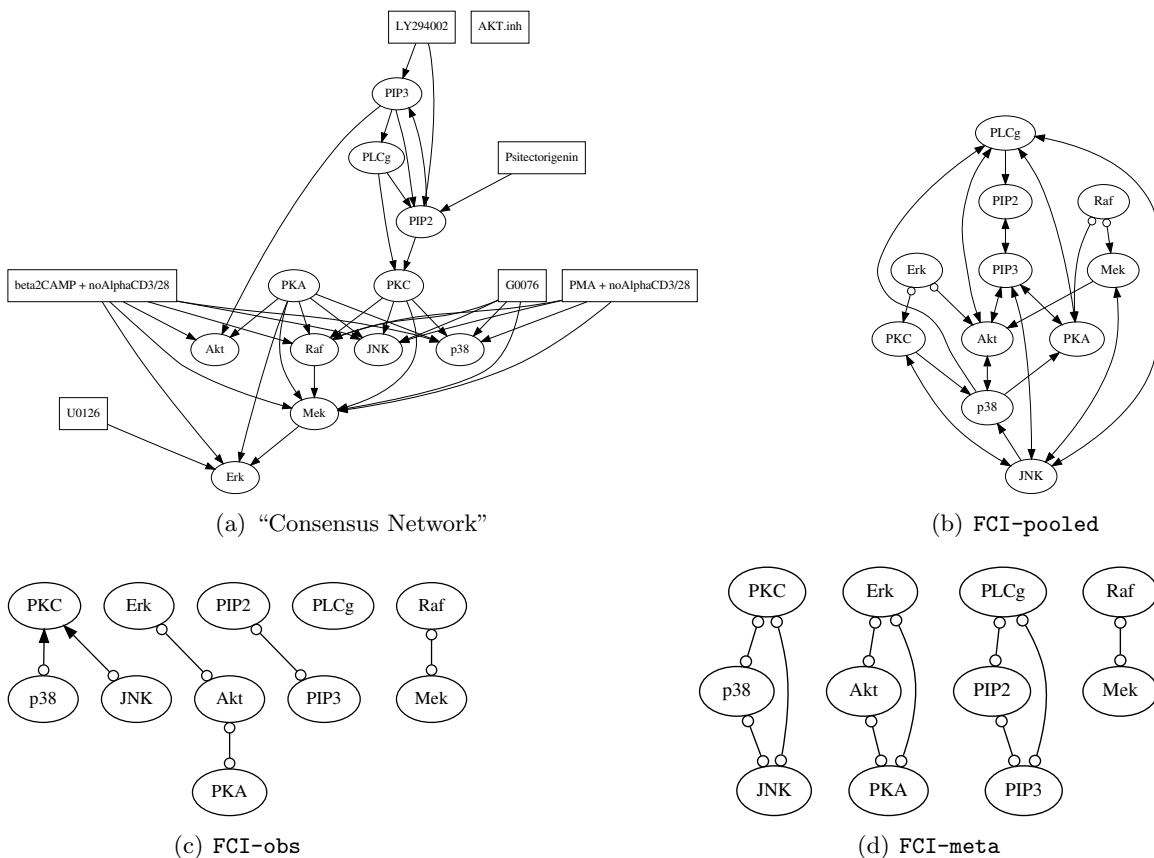
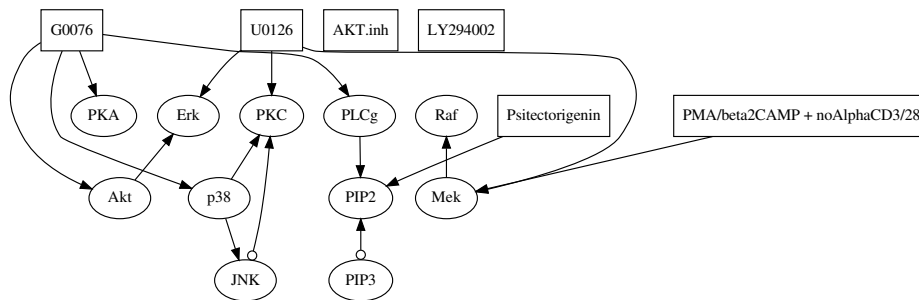


Figure 35: PAGs resulting from various FCI baselines on the flow cytometry data of Sachs et al. (2005). Intervention variables are denoted with rectangles, system variables with circles. From top to bottom: (a) "Consensus network" according to Sachs et al. (2005); (b) FCI on pooled data (without adding the context variables); (c) FCI on the first ("observational") dataset in which only global activators  $\alpha$ -CD3 and  $\alpha$ -CD28 have been administered; (d) FCI with as input the result of Fisher's method for combining conditional independence test results from all datasets.

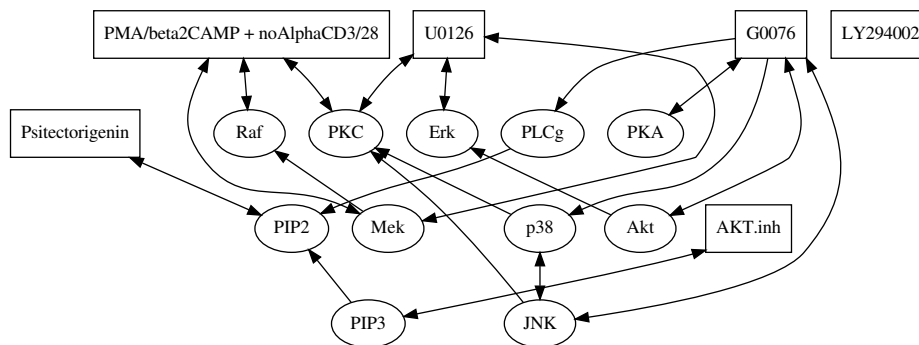
baselines FCI-obs, FCI-meta and FCI-pooled. Remarkably, FCI-JCI123 manages to orient most of the edges. The output also resembles the consensus network, although some of the edges seem to be reversed. Considering that we have not taken into account the available background knowledge on the intervention types and targets (Table 3), this is still an impressive result.

## 5. Conclusions and discussion

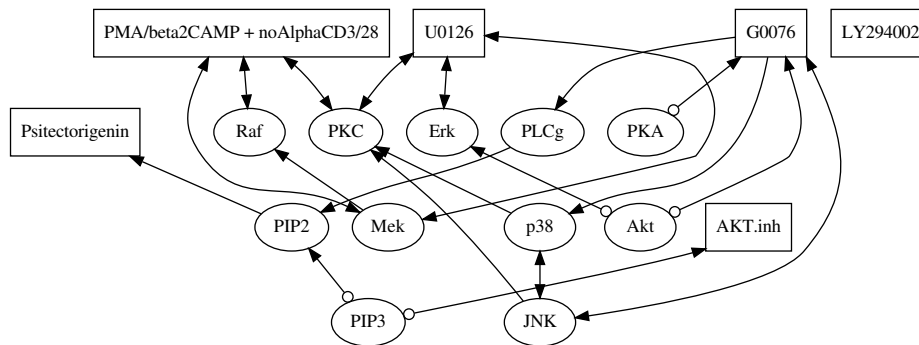
In this work, we proposed Joint Causal Inference (JCI), a powerful and elegant framework for causal discovery from datasets from multiple contexts. JCI generalizes the ideas of



(a) FCI-JCI123



(b) FCI-JCI1



(c) FCI-JCI0

Figure 36: PAGs resulting from various FCI-JCI variants on the flow cytometry data of Sachs et al. (2005). These causal discovery methods do not make use of the biological prior knowledge regarding intervention types and targets, but learn the intervention targets from the data. Intervention variables are denoted with rectangles, system variables with circles. From top to bottom, less JCI Assumptions are made. Note that these are individual PAGs that have not been bootstrapped. To get an idea of the robustness, Figure 37 shows also the corresponding bootstrap estimates for certain features of the PAGs.

# JOINT CAUSAL INFERENCE FROM MULTIPLE CONTEXTS

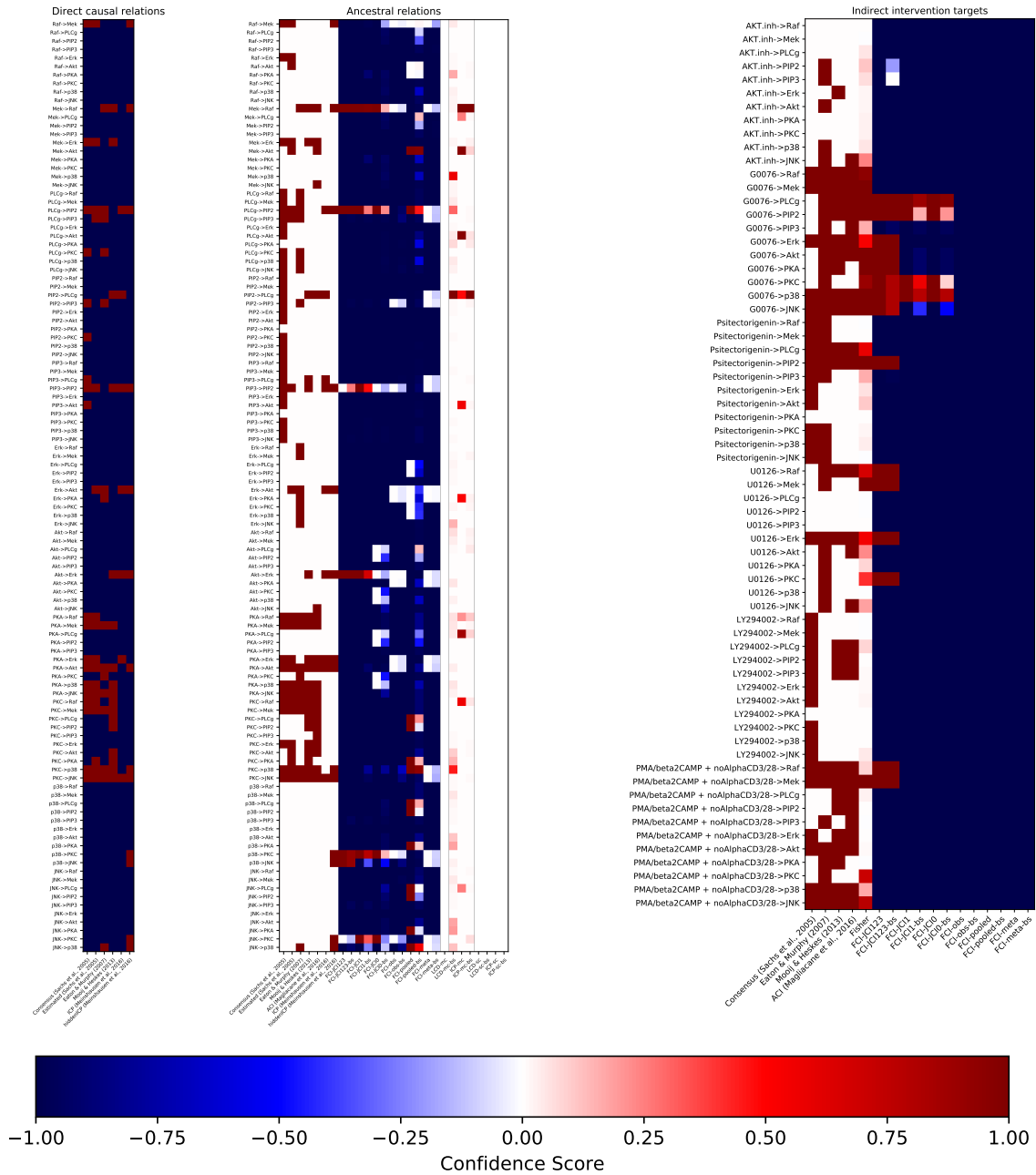


Figure 37: Causal relationships between the biochemical agents in the flow cytometry data of Sachs et al. (2005), according to different causal discovery methods and the “consensus network” according to Sachs et al. (2005). We also included results of causal discovery methods reported in other works.

causal discovery based on experimentation (as in randomized controlled trials and A/B-testing) to multiple context and system variables. Seen from another perspective, it also generalizes the ideas of causal discovery from purely observational data to the setting of datasets from multiple contexts—for example, different interventional regimes—by reducing the latter to a special case of the former, with additional background knowledge on the causal relationships involving the context variables. We proposed different flavours of JCI that differ in the amount of background knowledge that is assumed, some being more conservative than others. JCI can be implemented with any causal discovery method that can take into account the background knowledge. Surprisingly, we saw that one can even apply an off-the-shelf causal discovery algorithm for purely observational data on the pooled data, completely ignoring the background knowledge, and thereby already obtain significant improvements in the accuracy of the discovered causal relations.

We have seen how JCI deals with different types of interventions in a unified fashion, how it reduces learning intervention targets to learning the causal relations between context and system variables, and that it allows one to fully exploit all the information in the joint distribution on system and context variables. JCI was partially inspired on the approach by Eaton and Murphy (2007), but is much more generally applicable, as it allows for latent confounders and cycles, which are both important in many application domains. Especially noteworthy is that more conservative flavours of JCI allow for confounders between system and context variables, which cannot always be excluded, for example because the relevant aspects of the system’s context were only partially observed.

Since JCI is a causal modeling approach rather than a specific algorithm, it can be implemented in many different ways. We have investigated various implementations of JCI, amongst which some existing algorithms (LCD, ICP, and standard estimators for the presence of a causal effect in a randomized controlled trial), and also proposed novel implementations that are adaptations of algorithms for causal discovery from purely observational data to the JCI setting. In particular, we proposed **ASD-JCI**, an adaptation of the approach by Hyttinen et al. (2014) combined with ideas from Magliacane et al. (2016b), which is very flexible and accurate. By replacing d-separation with  $\sigma$ -separation (Forré and Mooij, 2018), **ASD-JCI** can also be used in general nonlinear cyclic settings. A major disadvantage of **ASD-JCI** is that it becomes computationally extremely expensive already for as few as  $\approx 7$  variables. We also proposed **FCI-JCI**, an adaptation of the FCI algorithm that enables it to exploit the applicable JCI background knowledge. This algorithm is less accurate than **ASD-JCI**, but much faster.

We evaluated different implementations of the JCI approach on synthetic data. We saw that JCI implementations outperform other state-of-the-art causal discovery algorithms in most settings. In some cases, the gains were quite extreme; for example, while purely observational causal discovery methods did not perform better than random guessing, JCI variants were able to discover with almost perfect precision ancestral causal relations between system variables. The only case in which all JCI implementations were outperformed by another causal discovery algorithm that combines data from different contexts, was the setting in which the contexts correspond with perfect (surgical) interventions with known targets. The reason is that none of the JCI implementations exploited the surgical nature of the interventions. However, we also saw that if interventions are not perfect (for example, in the case of causal mechanism changes), JCI implementations still perform very

well, while algorithms relying on the surgical nature of interventions may suffer from model misspecification. Another interesting observation we made in the experiments on synthetic data is that for the task of learning indirect (ancestral) causal relations, the classic (and very simple and fast) LCD algorithm can be competitive with many more sophisticated algorithms, like ICP and bootstrapped FCI-JCI.

We further illustrated the use of JCI by analyzing flow cytometry protein expression data (Sachs et al., 2005), a famous “benchmark” in the field of causal discovery. Unfortunately, applying ASD-JCI on the 11 system and 6 context variables would take excessive amounts of computation time, so we had to resort to FCI-JCI instead for causal discovery on a global scale. We compared with LCD and ICP variants that do causal discovery on more local scales. The results of various methods differ considerably, but show also some consistent patterns. This suggests that there is indeed a strong causal signal in the data, but it seems hard to conclude which of the various methods is best equipped to extract this signal most reliably, because the ground truth is only partially known. In future work, we plan to analyze more recent cytometry data sets that will allow for a more principled validation. Because often the true causal structure is not known, while interventional data is available, this requires to extend JCI-based causal discovery with causal prediction techniques, enabling one to predict the results of a particular intervention (Magliacane et al., 2018).

JCI offers increased flexibility when it comes to designing experiments for the purpose of causal discovery, as the JCI framework facilitates analysis of data from almost arbitrary experimental designs. This allows researchers to trade off the number and complexity of experiments to be done with the reliability of the analysis of the data for the purpose of causal discovery. Compared with existing methods, the framework offered by JCI is the most generally applicable, handling various intervention types and other context changes in a unified and non-parametric way, allowing for latent variables and cycles, and also applies when intervention types and targets are unknown, a common situation in causal discovery for complex systems.

As future work, we plan to (i) weaken the faithfulness assumption of JCI with respect to the context variables to allow for even more general experimental designs, (ii) address the problem of learning from datasets with non-identical (but overlapping) sets of observed variables, (iii) address selection bias, (iv) develop algorithms that need less computation time for delivering reliable results, (v) work on more applications on real-world data.

#### ACKNOWLEDGMENTS

We thank Thijs van Ommen for useful discussions and the reviewers for their constructive comments. SM, JMM and TC were supported by NWO, the Netherlands Organization for Scientific Research (VIDI grant 639.072.410). SM was also supported by the Dutch programme COMMIT/ under the Data2Semantics project. TC was also supported by NWO grant 612.001.202 (MoCoCaDi), and EU-FP7 grant agreement n.603016 (MATRICS).

#### References

Elias Bareinboim and Judea Pearl. A general algorithm for deciding transportability of experimental results. *Journal of Causal Inference*, 1:107–134, 2013.

- Tineke Blom, Anna Klimovskaia, Sara Magliacane, and Joris M. Mooij. An upper bound for random measurement error in causal discovery. In *Proceedings of the 34th Annual Conference on Uncertainty in Artificial Intelligence (UAI-18)*, 2018.
- Stephan Bongers, Jonas Peters, Bernhard Schölkopf, and Joris M. Mooij. Theoretical aspects of cyclic structural causal models. *arXiv.org preprint*, arXiv:1611.06221v2 [stat.ME], August 2018. URL <https://arxiv.org/abs/1611.06221v2>.
- Lin S. Chen, Frank Emmert-Streib, and John D. Storey. Harnessing naturally randomized transcription to infer regulatory relationships among genes. *Genome Biology*, 8(10):R219, 2007.
- David M. Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002.
- Tom Claassen and Tom Heskes. Causal discovery in multiple models from different experiments. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems (NIPS 2010)*, pages 415–423, Vancouver, British Columbia, Canada, 2010.
- Tom Claassen and Tom Heskes. A logical characterization of constraint-based causal discovery. In *UAI 2011*, pages 135–144, 2011.
- Diego Colombo and Marloes H Maathuis. Order-independent constraint-based causal structure learning. *The Journal of Machine Learning Research*, 15(1):3741–3782, 2014.
- Panayiota Constantinou and A. Philip Dawid. Extended conditional independence and applications in causal inference. *The Annals of Statistics*, 45(6):2618–2653, 2017.
- Gregory F. Cooper. A simple constraint-based algorithm for efficiently mining observational databases for causal relationships. *Data Mining and Knowledge Discovery*, 1(2):203–224, 1997.
- Gregory F. Cooper and Changwon Yoo. Causal discovery from a mixture of experimental and observational data. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI 1999)*, pages 116–125, Stockholm, Sweden, 1999.
- A. Philip Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society B*, 41(1):1–31, 1979.
- A. Philip Dawid. Influence diagrams for causal modelling and inference. *International Statistical Review*, 70(2):161–189, 2002.
- Vanessa Didelez, A. Philip Dawid, and Sara Geneletti. Direct and indirect effects of sequential treatments. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence (UAI 2006)*, pages 138–146. AUAI Press, 2006.
- Daniel Eaton and Kevin Murphy. Exact Bayesian structure learning from uncertain interventions. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS 2007)*, San Juan, Puerto Rico, 2007.

- D. Entner and P. O. Hoyer. On causal discovery from time series data using FCI. In *Proceedings of the Fifth European Workshop on Probabilistic Graphical Models (PGM-2010)*, pages 121–128, 2010.
- R. A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, 1925.
- R. A. Fisher. *The Design of Experiments*. Hafner, 1935.
- Patrick Forré and Joris M. Mooij. Markov properties for graphical models with cycles and latent variables. *arXiv.org preprint*, arXiv:1710.08775 [math.ST], October 2017. URL <https://arxiv.org/abs/1710.08775>.
- Patrick Forré and Joris M. Mooij. Constraint-based causal discovery for non-linear structural causal models with cycles and latent confounders. In *Proceedings of the 34th Annual Conference on Uncertainty in Artificial Intelligence (UAI-18)*, 2018.
- Patrick Forré and Joris M. Mooij. Causal calculus in the presence of cycles, latent confounders and selection bias. *arXiv.org preprint*, arXiv:1901.00433 [stat.ML], January 2019. URL <https://arxiv.org/abs/1901.00433>.
- Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe’er. Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, 7(3-4):601–620, 2000.
- Dan Geiger, Thomas Verma, and Judea Pearl. Identifying independence in Bayesian networks. *Networks*, 20(5):507–534, 1990.
- Clive W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969.
- Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13:2409–2464, 2012.
- David Heckerman, Dan Geiger, and David M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:197–243, 1995.
- Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *arXiv.org preprint*, arXiv:1706.08576 [stat.ME], June 2017. URL <https://arxiv.org/abs/1706.08576>.
- Antti Hyttinen, Frederick Eberhardt, and Patrick O. Hoyer. Learning linear cyclic causal models with latent variables. *Journal of Machine Learning Research*, 13:3387–3439, 2012.
- Antti Hyttinen, Frederick Eberhardt, and Matti Järvisalo. Constraint-based causal discovery: Conflict resolution with answer set programming. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence (UAI 2014)*, pages 340–349, Quebec City, Quebec, Canada, 2014.
- Markus Kalisch, Martin Mächler, Diego Colombo, Marloes H. Maathuis, and Peter Bühlmann. Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, 47(11):1–26, 2012.

- Yutaka Kano and Shohei Shimizu. Causal inference using nonnormality. In *Proceedings of the International Symposium on Science of Modeling, the 30th Anniversary of the Information Criterion*, pages 261–270, 2003.
- Mikko Koivisto and Kismat Sood. Exact Bayesian structure discovery in Bayesian networks. *Journal of Machine Learning Research*, 5:549–573, 2004.
- Sara Magliacane, Tom Claassen, and Joris M. Mooij. Joint causal inference from observational and interventional datasets. *arXiv.org preprint*, arXiv:1611.10351v1 [cs.LG], November 2016a. URL <http://arxiv.org/abs/1611.10351v1>.
- Sara Magliacane, Tom Claassen, and Joris M. Mooij. Ancestral causal inference. In *Proceedings of Advances in Neural Information Processing Systems (NIPS 2016)*, pages 4466–4474, Barcelona, Spain, 2016b.
- Sara Magliacane, Thijs van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31 (NeurIPS2018)*, pages 10869–10879. Curran Associates, Inc., 2018. URL <http://papers.nips.cc/paper/8282-domain-adaptation-by-using-causal-inference-to-predict-invariant-conditional-distributions.pdf>.
- Subramani Mani. *A Bayesian Local Causal Discovery Framework*. PhD thesis, University of Pittsburg, March 2006. URL <http://d-scholarship.pitt.edu/10181/>.
- Florian Markowetz, Steffen Grossmann, and Rainer Spang. Probabilistic soft interventions in conditional Gaussian networks. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics (AISTATS 2005)*, Bridgetown, Barbados, 2005.
- Christopher Meek. Strong completeness and faithfulness in Bayesian networks. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (UAI 1995)*, 1995.
- Joris M. Mooij and Tom Heskes. Cyclic causal discovery from continuous equilibrium data. In *Proceedings of the 29th Annual Conference on Uncertainty in Artificial Intelligence (UAI 2013)*, pages 431–439, 2013.
- Joris M. Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: Methods and benchmarks. *Journal of Machine Learning Research*, 17(32):1–102, 2016.
- Kevin Murphy. *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, University of Pittsburg, July 2002. URL <http://www.cs.ubc.ca/~murphyk/Thesis/thesis.pdf>.
- Chris J. Oates, Jim Korkola, Joe W. Gray, and Sach Mukherjee. Joint estimation of multiple related biological networks. *Annals of Applied Statistics*, 8(3):1892–1919, 2014.



- Chris J. Oates, Jim Q. Smith, and Sach Mukherjee. Estimating causal structure using conditional dag models. *Journal of Machine Learning Research*, 17(1):1880–1903, 2016a.
- Chris J. Oates, Jim Q. Smith, Sach Mukherjee, and James Cussens. Exact estimation of multiple directed acyclic graphs. *Statistics and Computing*, 26(4):797–811, 2016b.
- J. Pearl. A constraint propagation approach to probabilistic reasoning. In *Proceedings of the First Conference on Uncertainty in Artificial Intelligence (UAI 1985)*, pages 357–370, 1986.
- Judea Pearl. Comment: Graphical models, causality, and intervention. *Statistical Science*, 8:266–269, 1993.
- Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2009.
- J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, USA, 2017.
- Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15:2009–2053, 2014.
- Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society, Series B*, 78(5):947–1012, 2016.
- Thomas S. Richardson. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30:145–157, 2003.
- Thomas S. Richardson and Peter Spirtes. Ancestral graph Markov models. *The Annals of Statistics*, 30(4):962–1030, August 2002.
- Dominik Rothenhäusler, Christina Heinze, Jonas Peters, and Nicolai Meinshausen. BACK-SHIFT: Learning causal cyclic graphs from unknown shift interventions. In *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, pages 1513–1521. Curran Associates, Inc., 2015.
- Anna Roumpelaki, Giorgos Borboudakis, Sofia Triantafillou, and Ioannis Tsamardinos. Marginal causal consistency in constraint-based causal learning. In Frederick Eberhardt, Elias Bareinboim, Marloes Maathuis, Joris Mooij, and Ricardo Silva, editors, *Proceedings of the UAI 2016 Workshop on Causation: Foundation to Application*, number 1792 in CEUR Workshop Proceedings, pages 39–47, Aachen, 2016. URL <http://ceur-ws.org/Vol-1792/paper5.pdf>.
- Karen Sachs, Omar Perez, Dana Pe’er, Douglas A. Lauffenburger, and Garry P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308, 2005.

- Ilya Shpitser, Robin J. Evans, Thomas S. Richardson, and James M. Robins. Introduction to nested Markov models. *Behaviormetrika*, 41(1):3–39, 2014.
- Peter Spirtes, Christopher Meek, and Thomas S. Richardson. Causal inference in the presence of latent variables and selection bias. In *Proceedings of 11th Conference on Uncertainty in Artificial Intelligence (UAI 1995)*, 1995.
- Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2000.
- Jin Tian and Judea Pearl. Causal discovery from changes. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence (UAI 2001)*, Seattle, Washington, USA, 2001.
- Robert E. Tillman. Structure learning with independent non-identically distributed data. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML 2009)*, pages 1041–1048, 2009.
- Robert E. Tillman and Peter Spirtes. Learning equivalence classes of acyclic models with latent and selection variables from multiple datasets with overlapping variables. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS 2011)*, 2011.
- Sofia Triantafillou and Ioannis Tsamardinos. Constraint-based causal discovery from multiple interventions over overlapping variable sets. *Journal of Machine Learning Research*, 16:2147–2205, 2015.
- Sofia Triantafillou, Vincenzo Lagani, Christina Heinze-Deml, Angelika Schmidt, Jesper Tegner, and Ioannis Tsamardinos. Predicting causal relationships from biological data: Applying automated causal discovery on mass cytometry data of human immune cells. *Scientific Reports*, 7:12724, 2017.
- Thijs van Ommen and Joris M. Mooij. Algebraic equivalence of linear structural equation models. In *Proceedings of the 33rd Annual Conference on Uncertainty in Artificial Intelligence (UAI-17)*, 2017.
- S. Wright. Correlation and causation. *Journal of Agricultural Research*, 20:557–585, 1921.
- Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873–1896, 2008.
- Kun Zhang, Biwei Huang, Jiji Zhang, Clark Glymour, and Bernhard Schölkopf. Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 1347–1353, 2017.