Joint Causal Inference from Multiple Contexts

Mooij, J.M.; Magliacane, S.; Claassen, T.

Publication date
2020
Document Version
Final published version
Published in
Journal of Machine Learning Research

Journal of Machine Learning Research 21 (2020) 1-108 Submitted 3/17; Revised 1/20; Published 3/20

Joint Causal Inference from Multiple Contexts

Joris M. Mooij* J.M.MOOIJ@UVA.NL
Korteweg-De Vries Institute, University of Amsterdam
Postbox 94248, 1090 GE Amsterdam, The Netherlands

Sara Magliacane SARA.MAGLIACANE@IBM.COM
MIT-IBM Watson AI Lab, IBM Research
75 Binney St, Cambridge, MA 02142, USA

Tom Claassen TOMC@CS.RU.NL
Institute for Computing and Information Sciences, Radboud University Nijmegen
Postbox 9010, 6500 GL Nijmegen, The Netherlands

Editor: Peter Spirtes

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 from different contexts that elegantly unifies both approaches. JCI is a causal modeling framework rather than a specific algorithm, and it can be implemented using any causal discovery algorithm that can take into account certain background knowledge. JCI can deal with different types of interventions (e.g., perfect, imperfect, stochastic, etc.) in a unified fashion, and does not require knowledge of intervention targets or types in case of interventional data. We explain how several well-known causal discovery algorithms can be seen as addressing special cases of the JCI framework, and we also propose novel implementations that extend existing causal discovery methods for purely observational data to the JCI setting. We evaluate different JCI implementations on synthetic data and on flow cytometry protein expression data and conclude that JCI implementations can considerably outperform state-of-the-art causal discovery algorithms.

Keywords: causal discovery, causal modeling, causal inference, observational and experimental data, interventions, randomized controlled trials

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 pupils at 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 (see also Figure 1). This in itself does not yet imply a causal relation between the two in either direction. Indeed, an alternative explanation of

∗. Part of this work was done while the authors were with the Informatics Institute of the University of Amsterdam.

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 by means of *experimentation*, for example by 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. 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 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, just from *combining* multiple readily available data sets, *without* the need for an impractical experiment.[1] In this paper, we propose a simple and general way to combine and analyze 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., 1999; Zhang, 2008a), *score-based* causal discovery methods (e.g., Cooper and Herskovits, 1992; 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 methods were designed to estimate the causal graph of the system from a single data set corresponding to a single (purely observational) context.

More recently, various causal discovery methods have been proposed that extend these techniques to deal with multiple data sets from different contexts. As an example, the data sets may correspond with a baseline of purely observational data consisting of measurements concerning the "natural" state of the system, and data consisting of measurements under

1. One can show that the conditional dependence $C_\alpha \not\perp\!\!\!\perp X_2 \,|\, C_\beta$ and conditional independence $C_\alpha \perp\!\!\!\perp X_2 \,|\, \{X_1, C_\beta\}$ in the pooled data that are entailed by the causal graph, together with the assumption that neither C_α nor C_β is caused by X_1 or X_2, suffice to arrive at this conclusion.

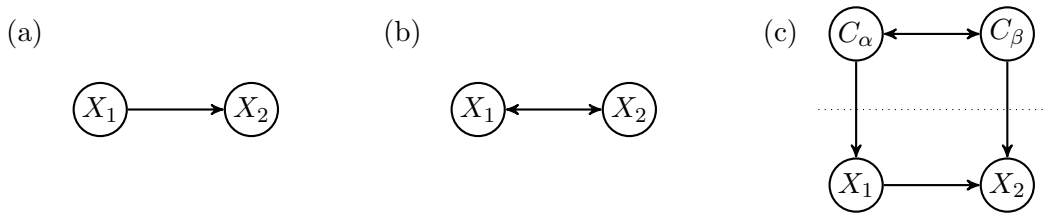


Figure 1: Different causal graphs relating X_1, the daily amount of hours spent on playing violent computer games, and X_2, a measure of aggressive behavior. (a) Playing violent computer games causes aggressive behavior; (b) The observed correlation between X_1 and X_2 is explained by a latent confounder, e.g., a genetic predisposition towards violence. (c) Hypothetical causal graph also involving context variables C_α, which indicates whether ultra-violent games have been banned by the government, and C_β, which represents 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_α and C_β are not caused by system variables X_1 and X_2).

different perturbations of the system due to external interventions on the system.[2] 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, but never directly compare data from different contexts (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).

In this paper, we propose *Joint Causal Inference (JCI)*, a framework for causal modeling of a system in different contexts and for causal discovery from multiple data sets consisting of measurements obtained in different contexts, which takes the latter approach. As will be discussed in more detail in Section 4.3, JCI is the most generally applicable of those approaches—for example, it allows for the presence of latent confounders and cyclic causal

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

relationships—and also offers most flexibility in terms of its implementation. While the ingredients of the JCI framework are not novel, the added value of the framework is that on the one hand it arrives at a unifying description of a diverse spectrum of existing approaches, while on the other hand it serves to inspire new implementations, such as the adaptations of FCI that we propose in this work. Technically, this is achieved by formulating the problem in terms of a (standard) Structural Causal Model that considers system and environment as subsystems of one joint system, rather than other types of representations in which the system is modeled conditionally on its environment (Dawid, 2002; Bareinboim and Pearl, 2013; Oates et al., 2016a; Yang et al., 2018; Forré and Mooij, 2019). This allows us to apply the standard notion of statistical independence in the same ways as is commonly done in the purely observational setting. As we observed in our experiments (that are reported in Section 5), the novel algorithms proposed in this work compare favorably with the state-of-the-art in causal discovery on synthetic data in many settings.

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, into a single data set, 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. On the other hand, if such background knowledge is available, it can be exploited.

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. For example, we will show in this work how FCI (Spirtes et al., 1999; Zhang, 2008a) can easily be adapted to the JCI setting. At the same time, JCI accommodates various well-known causal discovery methods as special cases, such as the standard randomized controlled trial setting (Fisher, 1935), Local Causal Discovery (LCD) (Cooper, 1997) and Invariant Causal Prediction (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 to achieve a causal discovery framework that is more powerful than either of the two separately.

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 and modeling assumptions. In Section 4, we show how JCI can be implemented using various causal discovery methods, and compare it with related work. In Section 5 we report experimental results on synthetic and flow cytometry data. We conclude in Section 6 with some promising directions for future developments.

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 review more recent extensions to the cyclic setting (Bongers et al., 2020). Because the cyclic setting is quite similar to the acyclic one that is mostly considered in the literature, we decided to present both cases in parallel rather than first explaining the acyclic setting and then explaining how everything generalizes to the cyclic setting.[3] 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. (2020).

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 $i \to j$ or $j \leftarrow i$, and call i a *parent* of j and j a *child* of i. We denote all parents of j in the graph $\mathcal{G}$ as $\text{PA}_{\mathcal{G}}(j) := \{i \in \mathcal{V} : i \to j \in \mathcal{E}\}$, and all children of i in $\mathcal{G}$ as $\text{CH}_{\mathcal{G}}(i) := \{j \in \mathcal{V} : i \to j \in \mathcal{E}\}$. We allow for self-cycles $i \to i$, 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 $i \to j \in \mathcal{E}$ 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 nodes in $\mathcal{W}$.

A *walk between* $i, j \in \mathcal{V}$ is a tuple $\langle i_0, e_1, i_1, e_2, i_3, \ldots, e_n, i_n \rangle$ of alternating nodes and edges in $\mathcal{G}$ ($n \geq 0$), such that all $i_0, \ldots, i_n \in \mathcal{V}$, all $e_1, \ldots, e_n \in \mathcal{E} \cup \mathcal{F}$, starting with node $i_0 = i$ and ending with node $i_n = j$, and such that for all $k = 1, \ldots, n$, 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 walk (path) from* $i \in \mathcal{V}$ *to* $j \in \mathcal{V}$ is a walk (path) between i and j such that every edge e_k on the walk (path) is of the form $i_{k-1} \to i_k$, i.e., every edge is directed and points away from i. By repeatedly taking parents, we obtain the *ancestors* of j: $\text{AN}_{\mathcal{G}}(j) := \{i \in \mathcal{V} : i = i_0 \to i_1 \to i_2 \to \cdots \to i_n = j \text{ in } \mathcal{G}\}$. Similarly, we define the *descendants* of i: $\text{DE}_{\mathcal{G}}(i) := \{j \in \mathcal{V} : i = i_0 \to i_1 \to i_2 \to \cdots \to i_n = j \text{ in } \mathcal{G}\}$. In particular, each node is ancestor and descendant of itself. A *directed cycle* is a directed path from i to j such that in addition, $j \to i \in \mathcal{E}$. An *almost directed cycle* is a directed path from

3. The disadvantage is that our notation and definitions deviate somewhat from those commonly used in the acyclic causal discovery literature. Therefore, we recommend reading this section also to those readers that are already familiar with the theory of acyclic structural causal models.

i to j such that in addition, $j \leftrightarrow i \in \mathcal{F}$. All nodes on directed cycles passing through $i \in \mathcal{V}$ together form the *strongly-connected component* $\text{SC}_{\mathcal{G}}(i) := \text{AN}_{\mathcal{G}}(i) \cap \text{DE}_{\mathcal{G}}(i)$ of 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 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 node i_k on a walk (path) $\pi = \langle i_0, e_1, i_1, e_2, i_3, \ldots, e_n, i_n \rangle$ in $\mathcal{G}$ is said to form a *collider on* π if it is a non-endpoint node ($1 \leq k < n$) and 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 (that is, if it is an endpoint node, i.e., $k = 0$ or $k = n$, or 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 $i_{k-1} \leftarrow i_k \leftrightarrow i_{k+1}$), i_k is called a *non-collider on* π. We will denote the colliders on a walk π as $\text{COL}(\pi)$ and the non-colliders on π (including the endpoints of π) as $\text{NCOL}(\pi)$. A triple of nodes $\langle i, j, k \rangle$ in $\mathcal{G}$ is called an *unshielded triple* if i is adjacent to j, j is adjacent to k and i is not adjacent to k in $\mathcal{G}$.

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. (2020) that is most convenient for our purposes. The reason we use SCMs to formulate JCI (rather than for example the more well-known causal Bayesian networks) is that SCMs are expressive enough to model both latent common causes and cyclic causal relationships.

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

(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*[4] *spaces* $\boldsymbol{\mathcal{X}} = \prod_{i \in \mathcal{I}} \mathcal{X}_i$*, which define the domains of the endogenous variables;*

(v) *a product of Borel spaces* $\boldsymbol{\mathcal{E}} = \prod_{j \in \mathcal{J}} \mathcal{E}_j$*, which define the domains of the exogenous variables;*

(vi) *a product probability measure* $\mathbb{P}_{\boldsymbol{\mathcal{E}}} = \prod_{j \in \mathcal{J}} \mathbb{P}_{\mathcal{E}_j}$ *on* $\boldsymbol{\mathcal{E}}$ *specifying the* exogenous distribution*;*

(vii) *a measurable function* $\boldsymbol{f} : \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{E}} \rightarrow \boldsymbol{\mathcal{X}}$*, the* causal mechanism*, such that each of its components* f_i *only depends on a particular subset of the variables, as specified by the*

4. 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 (isomorphic to) Borel spaces.

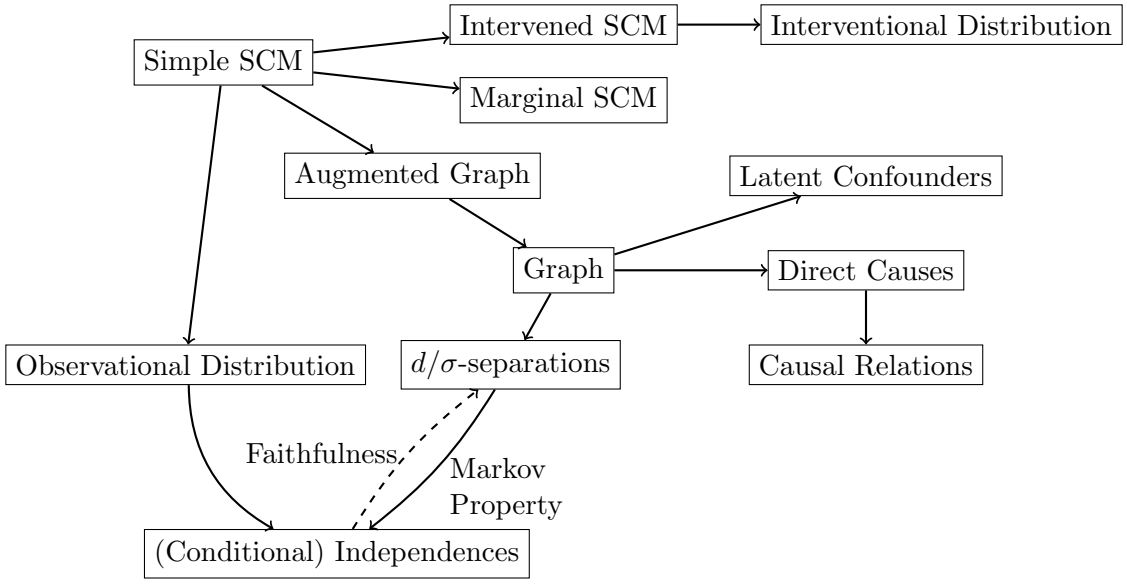


Figure 2: Relationships between various representations of simple SCMs. Directed edges represent mappings. Intervened and marginal SCMs are always defined and are also simple.

directed graph $\mathcal{H}$:

$$f_i : \boldsymbol{\mathcal{X}}_{\mathrm{PA}_{\mathcal{H}}(i)\cap\mathcal{I}} \times \boldsymbol{\mathcal{E}}_{\mathrm{PA}_{\mathcal{H}}(i)\cap\mathcal{J}} \to \mathcal{X}_i, \qquad i \in \mathcal{I}.$$

In discussing the concepts and 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 basic 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 $\mathcal{H}$ in Definition 1(iii) as the *augmented graph* of $\mathcal{M}$. In contrast, the *graph* of $\mathcal{M}$, denoted $\mathcal{G}(\mathcal{M})$, is the directed mixed graph with nodes $\mathcal{I}$, directed edges $i_1 \to i_2$ iff $i_1 \to i_2 \in \mathcal{H}$, and bidirected edges $i_1 \leftrightarrow i_2$ iff there exists $j \in \mathrm{PA}_{\mathcal{H}}(i_1) \cap \mathrm{PA}_{\mathcal{H}}(i_2) \cap \mathcal{J}$.[5] While the augmented graph $\mathcal{H}$ shows in detail the functional dependence of endogenous variables on the (independent) exogenous variables, the graph $\mathcal{G}(\mathcal{M})$ provides an abstraction by not including the exogenous variables explicitly, but using bidirected edges to represent any shared dependence of pairs of endogenous variables on a common exogenous parent. If $\mathcal{G}(\mathcal{M})$ is acyclic, we call the SCM $\mathcal{M}$ *acyclic*, otherwise we call the SCM *cyclic*. If $\mathcal{G}(\mathcal{M})$ contains no bidirected edges, we call the endogenous variables in the SCM $\mathcal{M}$ *causally sufficient*.

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

5. This definition of 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 $1 \to 2 \in \mathcal{H}$, but this edge would not appear in $\mathcal{G}(\mathcal{M})$. For the rigorous version of this definition, see Bongers et al. (2020).

is equal to the exogenous distribution $\mathbb{P}_{\boldsymbol{\varepsilon}}$, and the *structural equations*:

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

hold for all $i \in \mathcal{I}$. An SCM is often specified informally by specifying only the structural equations and the density[6] of the exogenous distribution with respect to some product measure, for example:

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

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 these could have no solution at all, or could have multiple solutions with different distributions (Bongers et al., 2020).

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 $\boldsymbol{g}_{\mathcal{O}} : \boldsymbol{\mathcal{X}}_{(\mathrm{PA}_{\mathcal{H}}(\mathcal{O})\setminus\mathcal{O})\cap\mathcal{I}} \times \boldsymbol{\mathcal{E}}_{\mathrm{PA}_{\mathcal{H}}(\mathcal{O})\cap\mathcal{J}} \to \boldsymbol{\mathcal{X}}_{\mathcal{O}}$ such that for $\mathbb{P}_{\boldsymbol{\varepsilon}}$-almost every $\boldsymbol{e}$ for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$:*

$$\boldsymbol{x}_{\mathcal{O}} = \boldsymbol{g}_{\mathcal{O}}(\boldsymbol{x}_{(\mathrm{PA}_{\mathcal{H}}(\mathcal{O})\setminus\mathcal{O})\cap\mathcal{I}}, \boldsymbol{e}_{\mathrm{PA}_{\mathcal{H}}(\mathcal{O})\cap\mathcal{J}}) \quad \Longleftrightarrow \quad \boldsymbol{x}_{\mathcal{O}} = \boldsymbol{f}_{\mathcal{O}}(\boldsymbol{x}, \boldsymbol{e})\,.$$

(Loosely speaking: the structural equations for $\mathcal{O}$ have a unique solution for $\boldsymbol{X}_{\mathcal{O}}$ in terms of the other variables appearing in those equations.)

If $\mathcal{M}$ is uniquely solvable with respect to $\mathcal{I}$ (in particular, this holds if $\mathcal{M}$ is acyclic), then it induces a unique *observational distribution* $\mathbb{P}_{\mathcal{M}}(\boldsymbol{X})$.

Given an SCM that models a certain system, we can model the system after an idealized intervention in which an external influence enforces a subset of endogenous variables to take on certain values, while leaving the rest of the system untouched.

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

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

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 ξ_i. Other types of interventions are possible as well (see also Section 3.3). If the intervened SCM $\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}$ induces a unique observational distribution, this is denoted as $\mathbb{P}_{\mathcal{M}}\big(\boldsymbol{X} \,|\, \mathrm{do}(I, \boldsymbol{\xi}_I)\big)$ and referred to as the *interventional distribution of $\mathcal{M}$ under the perfect intervention* $\mathrm{do}(I, \boldsymbol{\xi}_I)$. Pearl (2009) derived the *do-calculus* for acyclic SCMs, consisting of three rules that express relationships between interventional distributions of an SCM.

6. We denote a probability measure (or distribution) of a random variable $\boldsymbol{X}$ by $\mathbb{P}(\boldsymbol{X})$, and a density of $\boldsymbol{X}$ with respect to some fixed product measure by $p(\boldsymbol{X})$.

2.1.3. Simple Structural Causal Models

The theory of general cyclic Structural Causal Models is rather involved (Bongers et al., 2020). In this work, for simplicity of exposition, we will focus on a certain subclass of SCMs that has many convenient properties and for which the theory simplifies considerably:

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

All acyclic SCMs are simple. Simple SCMs provide a special case of the more general class of *modular* SCMs (Forré and Mooij, 2017). The class of simple SCMs can be thought of as a generalization of acyclic SCMs that allows for (weak) cyclic causal relations, but preserves many of the convenient properties that acyclic SCMs have.

Indeed, a simple SCM induces a unique observational distribution. Its marginalizations are always defined (Bongers et al., 2020), and are also simple; in other words, the class of simple SCMs is closed under marginalizations. The class of simple SCMs is also closed under perfect interventions, and hence, all perfect interventional distributions of a simple SCM are uniquely defined. Without loss of generality, one can assume that simple SCMs have no self-cycles. The causal interpretation of the graph of an SCM with cycles and/or bidirected edges can be rather subtle in general. However, for graphs of simple SCMs there is a straightforward causal interpretation:

Definition 5 *Let $\mathcal{M}$ be a simple SCM. If $i \to j \in \mathcal{G}(\mathcal{M})$ we call i a* direct cause of j according to $\mathcal{M}$. *If there exists a directed path $i \to \cdots \to j \in \mathcal{G}(\mathcal{M})$, i.e., if $i \in \mathrm{AN}_{\mathcal{G}(\mathcal{M})}(j)$, then we call i a* cause of j according to $\mathcal{M}$. *If there exists a bidirected edge $i \leftrightarrow j \in \mathcal{G}(\mathcal{M})$, then we call i and j* confounded according to $\mathcal{M}$.

We conclude that the graph $\mathcal{G}(\mathcal{M})$ of a simple SCM can be interpreted as its *causal graph*. In the next subsection, we will discuss how the same graph $\mathcal{G}(\mathcal{M})$ of a simple SCM $\mathcal{M}$ also represents the conditional independences that must hold in the observational distribution of $\mathcal{M}$.

2.1.4. Structural Causal Models: Markov Properties

Under certain conditions, the 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 $\mathbb{P}_{\mathcal{M}}(\boldsymbol{X})$. 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:[7]

Definition 6 (d-separation) *We say that a walk $\langle i_0 \dots i_n \rangle$ 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 $i_n \in C$, or
(ii) it contains a collider $i_k \notin \mathrm{AN}_{\mathcal{G}}(C)$, or
(iii) it contains a non-collider $i_k \in C$.
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 $A \perp^d_{\mathcal{G}} B \,|\, C$.*

7. It is also sometimes called "m-separation" in the ADMG literature.

In the general cyclic case, however, the notion of d-separation is too strong, as was already pointed out by Spirtes (1994). A solution is to replace it with a non-trivial generalization of d-separation, known as σ-separation (Forré and Mooij, 2017):

Definition 7 (σ-separation) *We say that a walk $\langle i_0 \dots i_n \rangle$ in DMG $\mathcal{G} = \langle \mathcal{V}, \mathcal{E}, \mathcal{F} \rangle$ is* σ-blocked *by $C \subseteq \mathcal{V}$ if:*

(i) its first node $i_0 \in C$ or its last node $i_n \in C$, or

(ii) it contains a collider $i_k \notin \text{AN}_{\mathcal{G}}(C)$, or

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

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 σ-blocked by a set $C \subseteq \mathcal{V}$, we say that A is σ-separated *from B by C, and we write $A \perp^{\sigma}_{\mathcal{G}} B \,|\, C$.*

Forré and Mooij (2017) proved the following fundamental result for modular SCMs, which we formulate here only for the special case of simple SCMs:

Theorem 8 (Generalized Directed Global Markov Property) *Any solution $(\boldsymbol{X}, \boldsymbol{E})$ of a simple SCM $\mathcal{M}$ obeys the* Generalized Directed Global Markov Property *with respect to the graph $\mathcal{G}(\mathcal{M})$:*

$$A \overset{\sigma}{\underset{\mathcal{G}(\mathcal{M})}{\perp}} B \,|\, C \implies \boldsymbol{X}_A \underset{\mathbb{P}_{\mathcal{M}}(\boldsymbol{X})}{\perp\!\!\!\perp} \boldsymbol{X}_B \,|\, \boldsymbol{X}_C \qquad \forall A, B, C \subseteq \mathcal{I}.$$

The following stronger Markov properties, in which σ-separation is replaced by the more familiar notion of d-separation, have been derived for special cases by Forré and Mooij (2017) (where again we consider only the special case of simple SCMs):

Theorem 9 (Directed Global Markov Property) *Let $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \mathcal{H}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ be a simple SCM. If $\mathcal{M}$ satisfies at least one of the following three conditions:*

(i) $\mathcal{M}$ is acyclic;

(ii) all endogenous spaces $\mathcal{X}_i$ are discrete;

(iii) $\mathcal{M}$ is linear (i.e., $\mathcal{X}_i = \mathbb{R}$ for each $i \in \mathcal{I}$, $\mathcal{E}_j = \mathbb{R}$ for each $j \in \mathcal{J}$, and each causal mechanism $f_i : \boldsymbol{\mathcal{X}}_{\text{PA}_{\mathcal{H}}(i) \cap \mathcal{I}} \times \mathcal{E}_{\text{PA}_{\mathcal{H}}(i) \cap \mathcal{J}} \to \mathcal{X}_i$ is linear), each causal mechanism f_i depends non-trivially on some exogenous variable(s), and its exogenous distribution has a density $p(\boldsymbol{E})$ with respect to Lebesgue measure;

then any solution $(\boldsymbol{X}, \boldsymbol{E})$ of $\mathcal{M}$ obeys the Directed Global Markov Property *with respect to the graph $\mathcal{G}(\mathcal{M})$:*

$$A \overset{d}{\underset{\mathcal{G}(\mathcal{M})}{\perp}} B \,|\, C \implies \boldsymbol{X}_A \underset{\mathbb{P}_{\mathcal{M}}(\boldsymbol{X})}{\perp\!\!\!\perp} \boldsymbol{X}_B \,|\, \boldsymbol{X}_C \qquad \forall A, B, C \subseteq \mathcal{I}.$$

Of these cases, the acyclic and linear cases are well-known.[8]

8. The acyclic case was first shown in the context of linear-Gaussian structural equation models (Spirtes et al., 1998; Koster, 1999). The discrete case fixes the erroneous theorem by Pearl and Dechter (1996), for which a counterexample was found by Neal (2000), by adding the unique solvability condition, and extends it to allow for latent common causes. The linear case extends existing results for the linear-Gaussian setting without latent common causes (Spirtes, 1994, 1995; Koster, 1996) to a linear (possibly non-Gaussian) setting with latent common causes.

We conclude that simple SCMs also have convenient Markov properties. A simple SCM induces a unique observational distribution that satisfies the Generalized Directed Global Markov Property; under additional conditions, it satisfies even the Directed Global Markov Property. Similarly, for any perfect intervention, a simple SCM induces a unique interventional distribution that satisfies the (Generalized) Directed Global Markov Property with respect to the intervened graph. We conclude that the graph $\mathcal{G}(\mathcal{M})$ of a simple SCM has two interpretations: it expresses both the causal structure between the variables as well as the conditional independence structure of the solutions. These two interpretations of the graph $\mathcal{G}(\mathcal{M})$ of a simple SCM can be combined into a causal do-calculus (Forré and Mooij, 2019) that extends the acyclic do-calculus of Pearl (2009) to the class of simple (or more generally, modular) SCMs.

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 $\mathbb{P}_{\mathcal{M}}(\boldsymbol{X})$ exists and satisfies a Markov property with respect to its 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 graph explains *all* conditional independences present in the observational distribution. For the cases in which the d-separation criterion Theorem 9 applies, this amounts to assuming the following implication:

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

Meek (1995) has shown completeness properties of d-separation. More specifically, Meek (1995) showed that faithfulness 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 in particular provides some justification for assuming faithfulness. On the other hand, no completeness results are known yet for the general cyclic case in which the σ-separation criterion Theorem 8 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:

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

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.

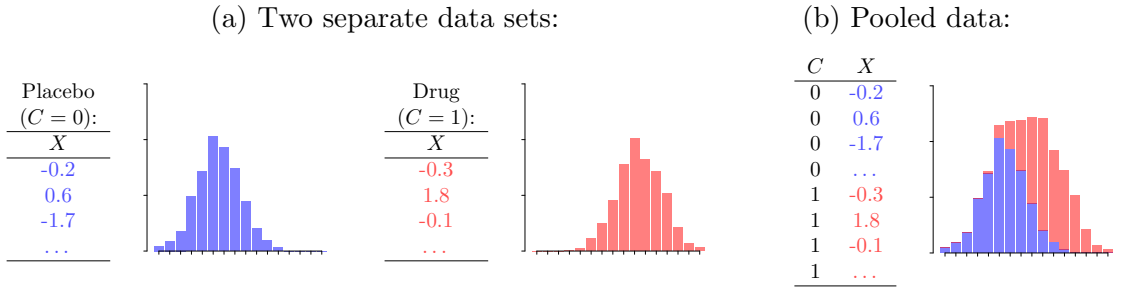


Figure 3: Illustration of the data from an example randomized controlled trial. The data can either be interpreted as (a) two separate data sets, one for the treatment and one for the control group, or (b) as a single data set including a context variable indicating treatment/control. Note that in this 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.

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.[9] Patients are treated depending on the assigned value of C, i.e., patients in the treatment group are treated with the drug and patients in the control group are treated with a placebo. Some time after treatment, the outcome X is measured for each patient. This yields a data set $(C_n, X_n)_{n=1}^N$ with two measurements (C_n, X_n) for the n^{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 cause of treatment C (which is commonly deemed justified if the outcome is an event that occurs later in time than the treatment event);

(ii) there is no latent confounder 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, no data is missing; for example, if only those patients that did not suffer from certain treatment side effects are included in the data set, then this assumption will be violated).

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, at least in this population of patients (see Proposition 10). There are two conceptually slightly different ways of testing

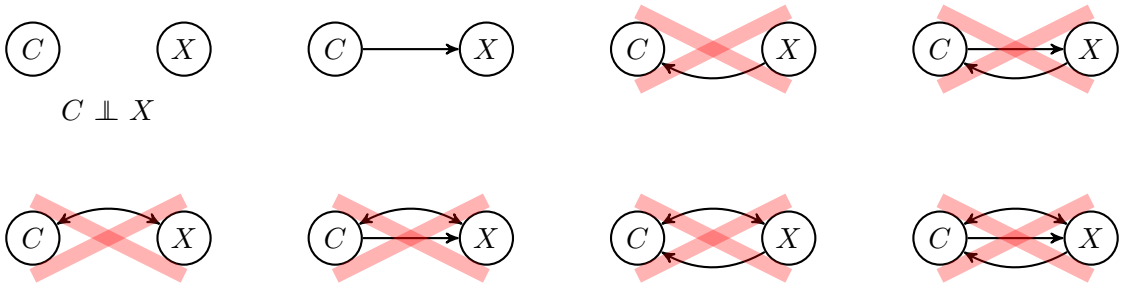
9. Usually this is done in a double-blind way, so that neither the patient nor the doctor knows which group a patient has been assigned to.

this in the data, depending on whether we treat the data as a single pooled data set, or rather as two separate data sets (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* $C \not\perp\!\!\!\perp X$ 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 10 *Suppose that the data-generating process on context variable C and outcome variable X can be modeled by a simple SCM $\mathcal{M}$ and no selection bias is present.*[10] *Under the randomized controlled trial assumptions:*
(i) $C \leftarrow X \notin \mathcal{G}(\mathcal{M})$ ("outcome X is not a cause of treatment C")
(ii) $C \leftrightarrow X \notin \mathcal{G}(\mathcal{M})$ ("there is no latent confounder of treatment C and outcome X"),
a dependence $C \not\perp\!\!\!\perp X$ in the joint distribution $\mathbb{P}(C, X)$ implies that C causes X. Furthermore, the causal effect of C on X is given by:

$$\mathbb{P}_{\mathcal{M}}\big(X \,|\, \mathrm{do}(C = c)\big) = \mathbb{P}_{\mathcal{M}}(X \,|\, C = c). \tag{1}$$

Proof Out of the eight possible graphs $\mathcal{G}(\mathcal{M})$, only two satisfy the assumptions:



By the Markov property (Theorem 8), if the edge $C \to X$ were absent in $\mathcal{G}(\mathcal{M})$, then C would be independent of X. Therefore, if $C \not\perp\!\!\!\perp X$, the edge $C \to X$ must be in $\mathcal{G}(\mathcal{M})$. In both cases, the causal do-calculus applied to $\mathcal{G}(\mathcal{M})$ yields the identity (1). ■

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

10. The context variable C is here considered as an *endogenous* variable in the SCM, as explained in Section 3.1.

idea of our approach is precisely this: *reducing an apparently complicated causal discovery problem with multiple data sets 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 an 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 $\mathbb{P}(X \mid C = 0)$ and $\mathbb{P}(X \mid C = 1)$ 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 Heskes, 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 $C \not\perp\!\!\!\perp X$ that *can only be seen* 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 gold standard 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 of high relevance, since experiments may be impossible, infeasible, 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., 1999; Zhang, 2008a), 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 latent confounders), 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 data set 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 ($A \to B$ vs. $B \to A$) 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 (to be made precise in Section 3.4).

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 11 *Suppose that the data-generating process on four variables X_1, X_2, X_3, X_4 can be modeled by a simple SCM $\mathcal{M}$. Assume that the sampling procedure is not subject to selection bias, and that faithfulness holds. If the following conditional (in)dependencies hold in the observational distribution $\mathbb{P}_{\mathcal{M}}(\boldsymbol{X})$:*

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

then X_3 is a direct cause of X_4 according to $\mathcal{M}$. Furthermore, X_3 and X_4 are unconfounded according to $\mathcal{M}$ and the causal effect of X_3 on X_4 is given by:

$$\mathbb{P}_{\mathcal{M}}\big(X_4 \,|\, \mathrm{do}(X_3 = x_3)\big) = \mathbb{P}_{\mathcal{M}}(X_4 \,|\, X_3 = x_3). \tag{2}$$

Proof By the assumed Markov and faithfulness properties, one can check that the only (cyclic or acyclic) graphs that are compatible with the observed conditional independences are the ones in Figure 4 (left), 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 direct cause of X_4 according to $\mathcal{M}$, but X_4 is not a direct cause of X_3 according to $\mathcal{M}$. Also, X_3 and X_4 cannot be confounded according to $\mathcal{M}$. 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 (equivalence classes of) causal graphs.

Typically, the 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 cause of X_3, or by X_1 and X_3 having a latent confounder, or both. However, under the assumption of faithfulness, one can deduce the Markov equivalence class of the graph from the conditional independences in the observational data, i.e., the class of all DMGs that induce the same separations. 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.

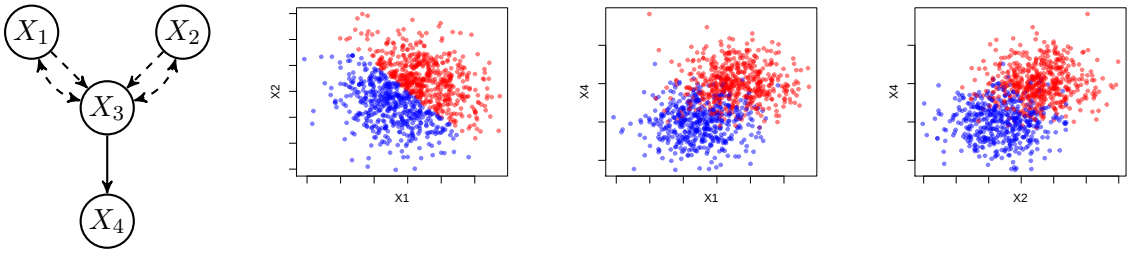


Figure 4: Left: Causal graphs satisfying the "Y-structure" pattern on four variables (X_1,X_2,X_3,X_4). Right: Scatter plots illustrating the Y-structure pattern in purely observational data, where X_3 is discrete-valued and its value is indicated by color (red/blue).

3. Joint Causal Inference

In this section we present Joint Causal Inference (JCI), a novel framework for causal discovery from multiple data sets corresponding to measurements that have been performed in different contexts. JCI combines the existing approaches towards causal discovery that we discussed in Sections 2.2 and 2.3.

3.1. The Distinction between System and Context

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. An observation that will turn out to be crucial in what follows is that 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 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 (Markowetz 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. Like system variables, context variables can be discrete or continuous (or more generally, take values in some Borel space).

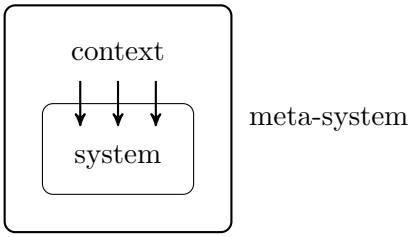


Figure 5: JCI reduces modeling a system in its environment to modeling the meta-system consisting of the system *and* its environment.

The idea of explicitly considering context variables is not novel: they have been discussed in the literature under various names, such as "policy variables" (Spirtes et al., 2000), "force variables" (Pearl, 1993), "decision variables" in influence diagrams (Dawid, 2002), "regime indicators" (Didelez et al., 2006), "selection variables" in selection diagrams (Bareinboim and Pearl, 2013), and "environment variable" (Peters et al., 2016). Their use for causal discovery was already suggested by Cooper and Yoo (1999). Formal aspects in how these variables are treated vary across accounts, however. For example, Dawid (2002) treats system variables as random variables and chooses 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, which expresses their assumed exogeneity with respect to the system.

Conceptually, context variables provide a more general notion 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 is more natural than considering it to be an intervention variable that encodes the result of a gender-changing operation on the patient. Furthermore, interventions usually come with an "observational baseline" of "doing nothing", but this is not always naturally available for more general context variables (e.g., "male" and "female" could both qualify as a baseline, while neither of the two would provide a more natural "observational" baseline than the other). When considering context variables, we do not have to specify such a baseline, whereas if we consider them as intervention variables, one can always ask "which value of the variable corresponds with no intervention?". Ultimately, though, both interpretations can be treated equally from a mathematical modeling perspective. Henceforth, we will use the term "context variable" in general, but "intervention variable" specifically for context variables that model an external intervention on the system.

That being said, the approach we take in JCI is simple (see also Figure 5): 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 probability of the system 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 $\boldsymbol{X} = (X_i)_{i \in \mathcal{I}}$ with values $\boldsymbol{x} \in \boldsymbol{\mathcal{X}} = \prod_{i \in \mathcal{I}} \mathcal{X}_i$ and the context variables $\boldsymbol{C} = (C_k)_{k \in \mathcal{K}}$ with values $\boldsymbol{c} \in \boldsymbol{\mathcal{C}} = \prod_{k \in \mathcal{K}} \mathcal{C}_k$. The latent exogenous variables of the SCM are denoted $\boldsymbol{E} = (E_j)_{j \in \mathcal{J}}$ with values $\boldsymbol{e} \in \boldsymbol{\mathcal{E}} = \prod_{j \in \mathcal{J}} \mathcal{E}_j$. The SCM modeling the meta-system is then assumed to be of the following form:

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

The system variables $\boldsymbol{X}$ and context variables $\boldsymbol{C}$ are all treated as endogenous variables of the meta-system, and the exogenous variables $\boldsymbol{E}$ are independent latent variables that are assumed not to be caused by the system variables $\boldsymbol{X}$ or the context variables $\boldsymbol{C}$.[11] The augmented 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 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 direct causal relations and latent confounders.

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 graph for each context. Representing the contexts jointly, as in (3), we simply obtain the union of those 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 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 6. 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 graph.

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

11. At this stage, we have not yet incorporated the assumption that context variables are exogenous to the system, and they are still treated equally to system variables in (3).

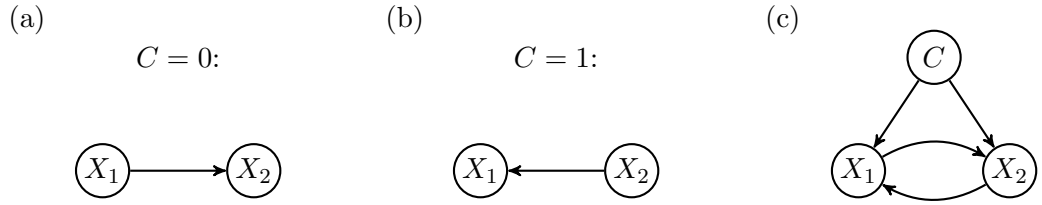


Figure 6: The graph of a mixture of two acyclic SCMs can be cyclic. (a) X_1 causes X_2 in context $C = 0$; (b) X_2 causes X_1 in context $C = 1$; (c) X_1 and X_2 cause each other in the joint model.

fraction of the total number of samples that have been measured in that context. In case the context variables are used to model interventions, for example, the context distribution is determined by the experimental design. 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 data set, 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. In other words, the behavior of the system is invariant of the context distribution.

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 intervention in the usual way (Pearl, 2009). Specifically, the perfect intervention that forces $\boldsymbol{X}_I$ to take on the value $\boldsymbol{\xi}_I$ ("do($\boldsymbol{X}_I = \boldsymbol{\xi}_I$)") 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:

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

while leaving the rest of the model invariant.[12]

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 $\text{CH}(k) = I$, no parents or spouses, and domain $\mathcal{C}_k = \{\emptyset\} \cup \prod_{i \in I} \mathcal{X}_i$, by taking $\boldsymbol{f}_I$ to be

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

of the following form:

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

for $i \in I$. Here, $C_k = \emptyset$ corresponds to no intervention (i.e., the observational baseline). Modeling a perfect intervention in this way is similar to the concept of "force variables" introduced by Pearl (1993). The observational distribution of the system variables is then given by the conditional distribution $\mathbb{P}(\boldsymbol{X} \,|\, C_k = \emptyset)$, the interventional distribution corresponding to the perfect intervention $\mathrm{do}(\boldsymbol{X}_I = \boldsymbol{\xi}_I)$ is given by the conditional distribution $\mathbb{P}(\boldsymbol{X} \,|\, C_k = \boldsymbol{\xi}_I)$, and the marginal distribution $\mathbb{P}(\boldsymbol{X})$ represents a mixture of those. This is illustrated in Figure 7.

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 $\mathrm{CH}(k) = \{i\}$ and domain $\mathcal{C}_k = \{A, B\}$:

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

As another example, a stochastic perfect 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 $j \in \mathrm{PA}(i)$ determine whether the intervention was successful:

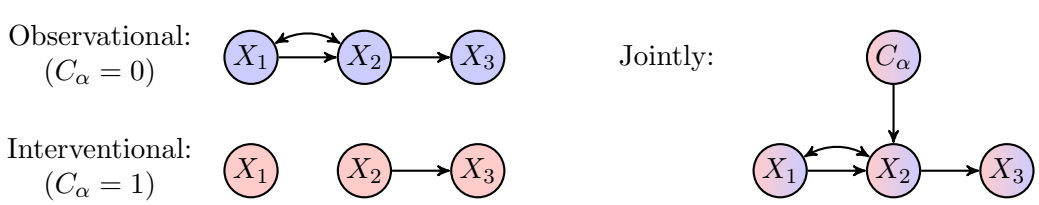
$$f_i(\boldsymbol{X}_{\mathrm{PA}(i)\cap\mathcal{I}}, \boldsymbol{C}_{\mathrm{PA}(i)\cap\mathcal{K}}, \boldsymbol{E}_{\mathrm{PA}(i)\cap\mathcal{J}}) \\ = \begin{cases} \tilde{f}_i(\boldsymbol{X}_{\mathrm{PA}(i)\cap\mathcal{I}}, \boldsymbol{C}_{\mathrm{PA}(i)\cap\mathcal{K}\setminus\{k\}}, \boldsymbol{E}_{\mathrm{PA}(i)\cap\mathcal{J}\setminus\{j\}})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

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. For example, it can deal with perfect interventions (Pearl, 2009), mechanism changes (Tian and Pearl, 2001), soft interventions (Markowetz 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 the context distribution $\mathbb{P}(\boldsymbol{C})$ (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.

(a)



(b)

Observational: $(C_\alpha = 0)$

X_1	X_2	X_3
-0.2	-0.4	0.6
0.6	0.8	1.3
-1.7	0.1	0.3

Interventional: $(C_\alpha = 1)$

X_1	X_2	X_3
-0.3	1.8	-0.1
1.8	-2.2	-0.2

Pooled:

C_α	X_1	X_2	X_3
0	-0.2	-0.4	0.6
0	0.6	0.8	1.3
0	-1.7	0.1	0.3
1	-0.3	1.8	-0.1
1	1.8	-2.2	-0.2

Figure 7: Two ways of representing interventions, either through modeling contexts separately (left), or by modeling system and context jointly (right). In this example, we consider a perfect intervention on X_2, though the same idea applies to other types of interventions. (a) shows the corresponding causal graphs, as separate graphs for each context (left), or as a single joint graph that includes a context variable (right); (b) shows different ways of grouping the data: as separate data sets for each context (left), or as a single joint data set after pooling (right).

3.4.1. JCI Assumption 0

First, we restate formally our basic modeling assumption:

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

$$\mathcal{M}: \begin{cases} C_k = f_k(\boldsymbol{X}_{\mathrm{PA}_{\mathcal{H}}(k)\cap\mathcal{I}}, \boldsymbol{C}_{\mathrm{PA}_{\mathcal{H}}(k)\cap\mathcal{K}}, \boldsymbol{E}_{\mathrm{PA}_{\mathcal{H}}(k)\cap\mathcal{J}}), & k \in \mathcal{K}, \\ X_i = f_i(\boldsymbol{X}_{\mathrm{PA}_{\mathcal{H}}(i)\cap\mathcal{I}}, \boldsymbol{C}_{\mathrm{PA}_{\mathcal{H}}(i)\cap\mathcal{K}}, \boldsymbol{E}_{\mathrm{PA}_{\mathcal{H}}(i)\cap\mathcal{J}}), & i \in \mathcal{I}, \\ \mathbb{P}(\boldsymbol{E}) = \prod_{j\in\mathcal{J}} \mathbb{P}(E_j), \end{cases} \tag{5}$$

that jointly models the system and the context. Its 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 causes any context variable, i.e.,*

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

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 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 was measured before an intervention was performed on the system. For example, a doctor typically first diagnoses a patient *before* deciding on treatment. For system variables containing the results of the medical examination used for the diagnosis and intervention variables describing the treatment that was decided *after*—and based upon—the medical examination, JCI Assumption 1 would not apply.

3.4.3. JCI Assumption 2

The second JCI assumption generalizes the randomization assumption for randomized controlled trials:

Assumption 2 *("Complete randomized context", optional) No context variable is confounded with a system variable, 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 knows 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) (see also Section 5.8), 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 size 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 the system variables reflect the measured properties of the individual cells, and the context variables 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, the lab assistant that 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 8. The following example describes a situation in which this phenomenon may occur.

Example 1 *Consider a randomized trial setup for establishing whether sugar causes plants to grow. Context variable C_α denotes the coin flip result, C_β indicates whether sugar is administered to the plant, and C_γ 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:*

$\mathbb{P}(\boldsymbol{C} = \boldsymbol{c})$	C_α *(coin flip)*	C_β *(sugar)*	C_γ *(water)*
$\frac{1}{2}$	*0*	*0*	*0*
$\frac{1}{2}$	*1*	*1*	*1*

If one would only take context variable C_β (did the plant get sugar?) into account and would treat C_α and C_γ as latent, as in Figure 8(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 must cause plants to grow.

A simple remedy to avoid the wrong conclusion if only C_β is observed would be to drop JCI Assumption 2: then it is no longer identifiable whether C_β causes X_1, or whether C_β and X_1 are just confounded.

3.4.4. JCI Assumption 3

We have seen that JCI Assumption 1 is often easily justifiable, but the applicability of JCI Assumption 2 may be less obvious in practice. We will now state JCI Assumption 3, which can be useful whenever both JCI Assumptions 1 and 2 have been made as well.

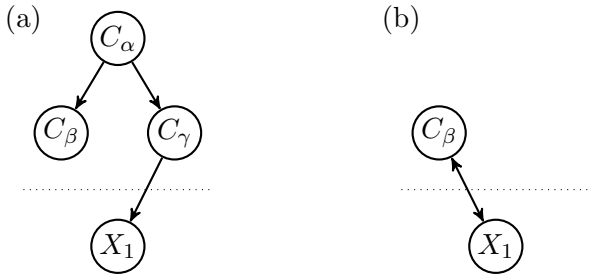


Figure 8: Confounding between system and context variables due to unobserved context variables. (a) If all three context variables $C_\alpha, C_\beta, C_\gamma$ are observed, JCI Assumption 2 would be valid. (b) After marginalizing out C_α and C_γ, leaving only context variable C_β as observed, JCI Assumption 2 is no longer valid.

Assumption 3 *("Generic context model", optional) The context graph*[13] $\mathcal{G}(\mathcal{M})_{\mathcal{K}}$ *is of the following special form:*

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

In Figure 9(b), this assumption is satisfied, while in Figure 9(a), it is not. We will show that JCI Assumption 3 seems stronger than it is, since it can be made without loss of generality in many cases occurring in practice.

In order to precisely formulate and prove that claim, the following definition is needed.

Definition 12 *Given an SCM $\mathcal{M}$ satisfying JCI Assumption 0, define the* conditional system graph $\mathcal{G}(\mathcal{M})_{\mathrm{do}(\mathcal{K})}$ *as the DMG with context nodes $\mathcal{K}$ and system nodes $\mathcal{I}$, and as directed and bidirected edges those edges in $\mathcal{G}(\mathcal{M})$ that contain at least one system node in $\mathcal{I}$ (i.e., excluding edges between context nodes). We will graphically represent the system nodes $\mathcal{I}$ of $\mathcal{G}(\mathcal{M})_{\mathrm{do}(\mathcal{K})}$ by ellipses and the context nodes $\mathcal{K}$ of $\mathcal{G}(\mathcal{M})_{\mathrm{do}(\mathcal{K})}$ by squares.*

Figure 9(c) shows the common conditional system graph for SCMs with graph as given in Figure 9(a) and for SCMs with graph as given in Figure 9(b). The conditional system graph provides a particular graphical representation for an SCM with context and system variables that is less expressive than its graph. This representation is useful when we are not interested in describing relationships between context variables, but only in describing the relationships between system variables and how the context affects the system.

The following key result essentially states that when one is only interested in modeling the causal relations involving the system variables (under JCI Assumptions 1 and 2), one does not need to care about the *causal relations* between the context variables, as long as one correctly models the context *distribution*.

13. Remember that $\mathcal{G}(\mathcal{M})_{\mathcal{K}}$ denotes the subgraph on the context variables $\mathcal{K}$ induced by the causal graph $\mathcal{G}(\mathcal{M})$.

Theorem 13 *Assume that JCI Assumptions 0, 1 and 2 hold for SCM $\mathcal{M}$:*

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

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

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

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

(i) the conditional system graphs coincide: $\mathcal{G}(\mathcal{M})_{\mathrm{do}(\mathcal{K})} = \mathcal{G}(\tilde{\mathcal{M}})_{\mathrm{do}(\mathcal{K})}$;

(ii) if $\tilde{\mathcal{M}}$ and $\mathcal{M}$ induce the same context distribution, i.e., $\mathbb{P}_{\mathcal{M}}(\boldsymbol{C}) = \mathbb{P}_{\tilde{\mathcal{M}}}(\boldsymbol{C})$, then for any perfect intervention on the system variables $\mathrm{do}(I, \boldsymbol{\xi}_I)$ with $I \subseteq \mathcal{I}$ (including the non-intervention $I = \emptyset$), $\tilde{\mathcal{M}}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}$ is observationally equivalent to $\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}$.

(iii) if the context graphs $\mathcal{G}(\tilde{\mathcal{M}})_{\mathcal{K}}$ 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 σ-separations).

Proof See Appendix A. ■

The following corollary of Theorem 13 states that JCI Assumption 3 can be made without loss of generality for the purposes of constraint-based causal discovery if the context distribution contains no conditional independences:

Corollary 14 *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 conditional system graphs coincide: $\mathcal{G}(\mathcal{M})_{\mathrm{do}(\mathcal{K})} = \mathcal{G}(\tilde{\mathcal{M}})_{\mathrm{do}(\mathcal{K})}$;

(ii) for any perfect intervention on the system variables $\mathrm{do}(I, \boldsymbol{\xi}_I)$ with $I \subseteq \mathcal{I}$ (including the non-intervention $I = \emptyset$), $\tilde{\mathcal{M}}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}$ is observationally equivalent to $\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}$;

(iii) if the context distribution $\mathbb{P}_{\mathcal{M}}(\boldsymbol{C})$ contains no conditional or marginal independences, then the same σ-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 This follows from Theorem 13 by showing that there exists an $\tilde{\mathcal{M}}$ that satisfies all requirements in Theorem 13 and JCI Assumption 3 by construction, and that induces the same context distribution as $\mathcal{M}$ does. For a detailed proof, see Appendix A. ■

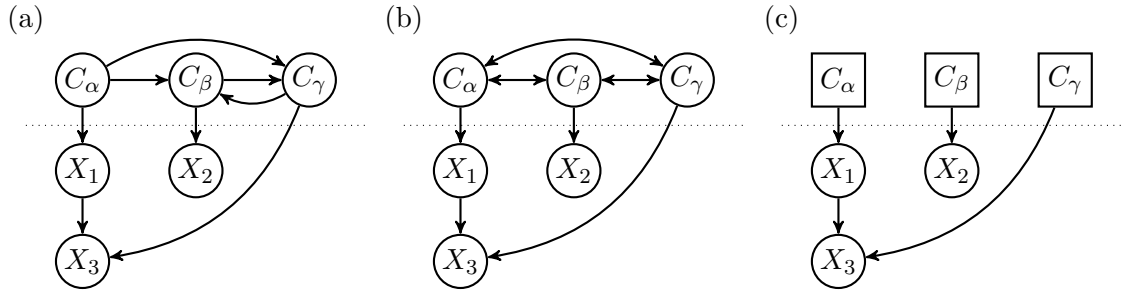


Figure 9: Example graphs of (a) a true SCM $\mathcal{M}$ and (b) the modified SCM $\tilde{\mathcal{M}}$ constructed in the proof of Corollary 14 that satisfies JCI Assumption 3, and (c) their corresponding conditional system graph $\mathcal{G}(\mathcal{M})_{\mathrm{do}(\mathcal{K})}$. Corollary 14 gives sufficient conditions for $\tilde{\mathcal{M}}$ and $\mathcal{M}$ to be equivalent for our purposes.

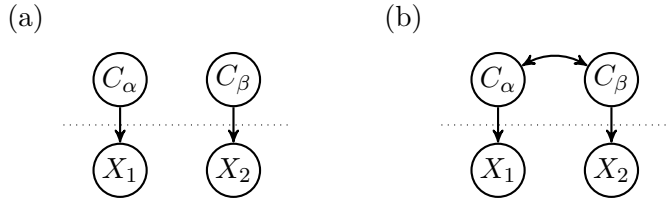


Figure 10: Example that illustrates that the genericity assumption in statement (iii) of Corollary 14 is necessary. Graphs of (a) the true SCM $\mathcal{M}$ and (b) the modified SCM $\tilde{\mathcal{M}}$ constructed in the proof of Corollary 14 that are not Markov equivalent. The graph in (a) is identifiable under JCI Assumptions 0–2. The joint distribution $\mathbb{P}_{\mathcal{M}}(\boldsymbol{X}, \boldsymbol{C})$ is not faithful with respect to the graph in (b), which is the minimal one that also satisfies JCI Assumption 3 such that $\mathbb{P}_{\mathcal{M}}(\boldsymbol{X}, \boldsymbol{C})$ is Markov with respect to it.

An example illustrating this corollary is provided in Figure 9.

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 14 shows that we may assume JCI Assumption 3 without loss of generality if JCI Assumptions 1 and 2 are made and the context distribution contains no (conditional) independences. The causal discovery algorithm then does not need to 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 14 (i.e., $\mathbb{P}_{\mathcal{M}}(\boldsymbol{C})$ containing no conditional independences) is necessary, as the simple counterexample in Figure 10 shows. Depending on how well the causal discovery algorithm can handle faithfulness violations, model misspecification due to incorrectly assuming JCI Assumption 3 even though $\mathbb{P}(\boldsymbol{C})$ contains conditional independences might prevent successful identifica-

C_α	C_β	C_γ	C_δ	C_ϵ	possible interpretation
0	0	0	0	0	observational
1	0	0	0	0	intervention α
0	1	0	0	0	intervention β
0	0	1	0	0	intervention γ
0	0	0	1	0	intervention δ
0	0	0	0	1	intervention ϵ

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.

tion of the causal relationships between system variables. Therefore, it is prudent to check that the empirical context distribution $\hat{\mathbb{P}}(\boldsymbol{C})$ 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 $\boldsymbol{Z}$ if $\mathbb{P}(X = 1, Y = 1 \mid \boldsymbol{Z} = \boldsymbol{z}) > 0$ for all $\boldsymbol{z}$ with $\mathbb{P}(\boldsymbol{Z} = \boldsymbol{z}) > 0$. 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 $\mathbb{P}(\boldsymbol{C})$ corresponding to any such diagonal design (with non-zero probability for each context) contains no conditional independences.

JCI Assumption 3 can easily 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 that expresses the known conditional independences in the experimental design. Theorem 13 can be applied to these more general situations as well and shows that for the purpose of constraint-based causal discovery, any context graph that implies the

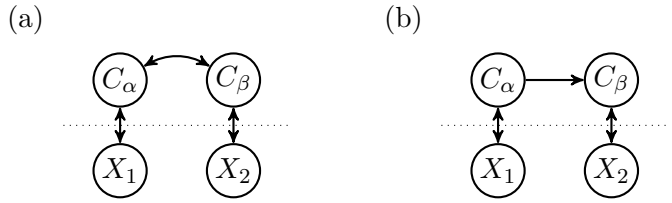


Figure 11: If JCI Assumption 2 does not apply, the causal relations between context variables have testable consequences for the conditional independences in the joint distribution. (a) $X_1 \not\perp\!\!\!\perp X_2 \,|\, \{C_\alpha, C_\beta\}$; (b) $X_1 \perp\!\!\!\perp X_2 \,|\, \{C_\alpha, C_\beta\}$.

observed conditional independences (i.e., any graph that is Markov equivalent to the true context graph) works.

Another alternative is to omit JCI Assumption 3 and instead try to infer the context subgraph $\mathcal{G}(\mathcal{M})_{\mathcal{K}}$ from the data. This would typically be computationally more expensive, but in our experience does not seem to make much of a difference in terms of accuracy in our experiments (as we report in Section 5).

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 of this,[14] see Figure 11. Here, C_α could be "lab", and C_β could be the "temperature" at which an experiment is performed. In this case, we get different conditional independences in the joint distribution $\mathbb{P}(X_1, X_2, C_\alpha, C_\beta)$ if lab causes temperature than when they are confounded (for example, by geographical location). Something similar can happen if context variables are caused by system variables.

3.4.5. Summary of JCI Assumptions and Other Background Knowledge

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 just the standard assumption made throughout the causal discovery literature, but now applied to the meta-system rather than to the system only. 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 causes any context variable) to apply. In some cases, also JCI Assumption 2 (no system variable is confounded with any context variables) applies. If both apply, one can assume JCI Assumption 3 for convenience if the context distribution contains no (conditional) independences. More generally, 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.

14. We are grateful to Thijs van Ommen for pointing this out.

The reader may wonder when one can ever be sure in practice that JCI Assumption 2 applies. There is one very common scenario in which JCI Assumption 2 holds. This is in a scientific experiment in which, in chronological order:

(i) an ensemble of systems is prepared in an initial state;

(ii) the systems are randomly permuted and randomly divided into batches;

(iii) for each batch, all systems in the batch are intervened upon simultaneously in the same way (following an experimental protocol determined in advance);

(iv) measurements of the system variables are performed.

The experimental protocol specifies the *intended interventions* for each batch, which should be completely encoded as context variables. Since the intended interventions have been decided *before* system variables are measured, the intended interventions cannot be caused by the system variables. Because the systems were randomly permuted and divided into batches, the assigned batch cannot be caused by prior values of the system variables, or by anything else that may also have an effect on the system variables. Because the intended interventions for each system in each batch are determined completely by the batch, this implies that intended interventions and system variables cannot be confounded. As long as the context variables provide a complete encoding of the intended interventions (i.e., the intended interventions are in one-to-one correspondence to values of the context variables), JCI Assumptions 1 and 2 then apply to the context variables.[15] If additionally, no (conditional) independences hold in the empirical context distribution, we can also make use of JCI Assumption 3 to simplify and speed up the causal discovery procedure.

In more general scenarios, such as the example in the introduction (concerning the question whether playing violent computer games causes aggressive behavior), the validity of JCI Assumption 2 (no confounding between context and system) should not be taken for granted. For example, it could be that precisely the schools with a more violent population of pupils see themselves forced to actively take measures to promote social behavior. In that case, the level of violence in the past would confound C_β (does a school take measures to stimulate social behavior) and X_2 (how violently do the pupils of the school behave). Thus, in scenarios like these, it seems safer not to rely on JCI Assumption 2 as incorrectly assuming it might lead to wrong conclusions (although we do not currently understand the precise impact of such model misspecification).

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 applying JCI on a combination of different interventional data sets, intervention targets can be learnt

15. If the experimenter sticks to the experimental protocol that was fixed before the experiment was performed, any possible influence of the system variables on the *performed interventions* is excluded, and therefore the *performed interventions* will equal the *intended interventions*. This means that the JCI modeling framework (with JCI Assumptions 1 and 2) applies also when interpreting the context variables as the performed interventions. This may explain why it is considered good scientific practice to perform an experiment according to an experimental protocol that was fixed beforehand, and not deviate from it in case of unexpected measurement outcomes, for example.

from data when they are not known (as the direct effects of intervention variables), 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 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.

4. Causal Discovery from Multiple Contexts with JCI

In this section, we discuss how causal discovery from multiple contexts can be performed in the Joint Causal Inference framework. Our starting point is the assumption that some model of the form (5) is an appropriate causal model for the system and its context, and we have obtained samples of all system variables in multiple contexts.[16] Suppose that the exact model $\mathcal{M}$ and in particular, its causal graph $\mathcal{G}(\mathcal{M})$, are 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 data set for context $\boldsymbol{c} \in \boldsymbol{\mathcal{C}}$ as $\mathcal{D}^{(\boldsymbol{c})} = \big((x_{in}^{(\boldsymbol{c})})_{i\in\mathcal{I}}\big)_{n=1}^{N_{\boldsymbol{c}}}$, and for simplicity, assume that no values are missing. The number of samples in each context, given by $N_{\boldsymbol{c}}$, is allowed to depend on the context. As a first step, we *pool* the data, thereby representing it as a single data set $\mathcal{D} = (\boldsymbol{x}_n, \boldsymbol{c}_n)_{n=1}^{N}$ where $N = \sum_{\boldsymbol{c}\in\boldsymbol{\mathcal{C}}} N_{\boldsymbol{c}}$. We then assume that $\mathcal{D}$ is an i.i.d. sample of $\mathbb{P}_{\mathcal{M}}(\boldsymbol{X}, \boldsymbol{C})$, where $(\boldsymbol{X}, \boldsymbol{C}, \boldsymbol{E})$ is a solution of the SCM $\mathcal{M}$ of the form (5).[17]

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 ac-

16. An interesting problem setting considered by several researchers (Claassen and Heskes, 2010; Tillman and Spirtes, 2011; Triantafillou and Tsamardinos, 2015; Hyttinen et al., 2014; Forré and Mooij, 2018) that we do not consider here would be to allow for each context a (possibly context-dependent) subset of system variables to remain unobserved.

17. Although this may sound as an innocuous assumption, it is not necessarily satisfied by the data generating process. For example, suppose that in a randomized controlled trial, it is decided *a priori* that a certain number N_0 of patients will be assigned to the control group, and a number N_1 of patients to the treatment group, but which patients end up in which group is completely randomized. The resulting pooled data is not i.i.d.; indeed, if we repeat this procedure, we will always end up with the same number of patients in each group, whereas for an i.i.d. sample, the numbers would fluctuate around their expected values. Nevertheless, this assumption can be made here without losing much generality. In particular, for the case of binary treatment and binary outcome, Wasserman (2004, Section 15.5) shows that for independence tests based on the (log) odds ratio the i.i.d. assumption can be weakened accordingly. Alternatively, in a bootstrapping procedure (which we will use in practice for most implementations of JCI in Section 5), the resampled pooled data is i.i.d. by construction.

cordance 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 $\mathbb{P}_{\mathcal{M}}(\boldsymbol{X}, \boldsymbol{C})$ is faithful with respect to the graph $\mathcal{G}(\mathcal{M})$, using the appropriate separation criterion (σ-separation in general, or d-separation for specific cases, as discussed in Section 4.1). We will discuss the ramifications of the faithfulness assumption in more detail in Section 4.1.

Definition 15 *We say that a particular feature of $\mathcal{G}(\mathcal{M})$ is* identifiable *from $\mathbb{P}_{\mathcal{M}}(\boldsymbol{X}, \boldsymbol{C})$ and background knowledge if the feature is present in the graph $\mathcal{G}(\tilde{\mathcal{M}})$ of any SCM $\tilde{\mathcal{M}}$ with $\mathbb{P}_{\tilde{\mathcal{M}}}(\boldsymbol{X}, \boldsymbol{C}) = \mathbb{P}_{\mathcal{M}}(\boldsymbol{X}, \boldsymbol{C})$ 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 $\mathbb{P}_{\mathcal{M}}(\boldsymbol{X}, \boldsymbol{C})$, and the available background knowledge.

The key insight of the Joint Causal Inference framework that allows one to deal with data from multiple contexts $(\mathcal{D}^{(\boldsymbol{c})})_{\boldsymbol{c} \in \boldsymbol{\mathcal{C}}}$ 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.

This trick also allows us to easily learn intervention targets from data in a similar way as we usually learn causal effects between variables from data: the intervention targets are simply encoded as the direct effects of the intervention variables.

After discussing the faithfulness assumption in more detail, we will give a few suggestions of how JCI can be implemented in Section 4.2.

4.1. Faithfulness Assumption

In this subsection we will discuss the subtleties of the faithfulness assumption in the JCI setting and compare it with alternative faithfulness assumptions that have been made in the literature.

Given a simple SCM $\mathcal{M}$ of the form (5) with graph $\mathcal{G} := \mathcal{G}(\mathcal{M})$, the joint distribution $\mathbb{P}_{\mathcal{M}}(\boldsymbol{X}, \boldsymbol{C})$ induced by the SCM satisfies the Generalized Directed Global Markov Property (Theorem 8) with respect to the graph $\mathcal{G}$ of the SCM, i.e., any σ-separation $U \perp^{\sigma}_{\mathcal{G}} V \,|\, W$ between sets of nodes $U, V, W \subseteq \mathcal{I} \cup \mathcal{K}$ in the graph $\mathcal{G}$ implies a conditional independence $\tilde{\boldsymbol{X}}_U \perp\!\!\!\perp_{\mathbb{P}_{\mathcal{M}}(\boldsymbol{X}, \boldsymbol{C})} \tilde{\boldsymbol{X}}_V \,|\, \tilde{\boldsymbol{X}}_W$, where we write $\tilde{\boldsymbol{X}} := (\boldsymbol{X}, \boldsymbol{C})$. Under the additional assumptions of Theorem 9, 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 $\mathbb{P}_{\mathcal{M}}(\boldsymbol{X}, \boldsymbol{C})$ is faithful with respect to the graph $\mathcal{G}(\mathcal{M})$ of $\mathcal{M}$. In other words, any conditional independence $\tilde{\boldsymbol{X}}_U \perp\!\!\!\perp_{\mathbb{P}_{\mathcal{M}}(\boldsymbol{X}, \boldsymbol{C})} \tilde{\boldsymbol{X}}_V \,|\, \tilde{\boldsymbol{X}}_W$ for sets of nodes $U, V, W \subseteq \mathcal{I} \cup \mathcal{K}$ is due to the σ-separation $U \perp^{\sigma}_{\mathcal{G}} V \,|\, W$ in $\mathcal{G}$ (or d-separation, if applicable), and no other

conditional independences in $\mathbb{P}_{\mathcal{M}}(\boldsymbol{X}, \boldsymbol{C})$ exist. In particular, this assumption rules out any conditional independence between context variables in case JCI Assumption 3 is made. If JCI Assumption 3 is not made, conditional independences in the context distribution that are faithfully described by any DMG are allowed.

This faithfulness assumption allows us to deal with different types of interventions, including perfect interventions. For example, for the perfect intervention on X_2 illustrated in Figure 7 the causal graphs $\mathcal{G}_{\mathcal{I}}^{(\boldsymbol{c})}$ (restricted to the system variables $\{X_i\}_{i\in\mathcal{I}}$) depend on the context $\boldsymbol{c} \in \boldsymbol{\mathcal{C}}$: in the observational context ($C_\alpha = 0$), $X_1 \to X_2$, 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 $X_1 \perp\!\!\!\perp X_2 \,|\, C_\alpha = 0$, we still have $X_1 \not\perp\!\!\!\perp X_2 \,|\, C_\alpha$ because $X_1 \not\perp\!\!\!\perp X_2 \,|\, C_\alpha = 1$.[18] In other words, the fact that $\mathbb{P}(\boldsymbol{X} \,|\, C_\alpha = 1)$ 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 another faithfulness assumption. In our notation, such approaches assume that $\mathbb{P}(\boldsymbol{X} \,|\, \boldsymbol{C} = \boldsymbol{c})$ is faithful w.r.t. a causal subgraph $\mathcal{G}_{\mathcal{I}}^{(\boldsymbol{c})}$ that may be context-dependent, and must then reason about how these context-dependent subgraphs are related, explicitly relying on knowledge about the type of interventions (typically assuming that the interventions are perfect interventions with known targets). This faithfulness assumption is to a certain extent stronger than ours because it requires faithfulness of the system within *each* context. On the other hand, it is to a certain extent weaker than ours because ours implies restrictions on the context distribution ($\mathbb{P}_{\mathcal{M}}(\boldsymbol{C})$ must be faithful to some DMG, at the very least) that the alternative does not have. We consider the extension of the applicability of JCI (because no knowledge of intervention types or targets is needed) due to the faithfulness assumption we chose here to outweigh the limitations in applicability (because not *every* context distribution can be handled). In the rest of this section, we will discuss some simple workarounds that can be applied in practice when dealing with faithfulness violations in the context distribution $\mathbb{P}_{\mathcal{M}}(\boldsymbol{C})$.

Under JCI Assumption 3, the faithfulness assumption implies that the context distribution $\mathbb{P}(\boldsymbol{C})$ does not contain any conditional independences. In case the empirical context distribution $\hat{\mathbb{P}}(\boldsymbol{C})$ *does* contain conditional independences, one has several options. The first (assuming also JCI Assumptions 1 and 2 are made) is to modify the assumed 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

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

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 in the context distribution 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_α is a (deterministic) function of C_θ and C_ι: $C_\alpha = \neg(C_\theta \vee C_\iota)$. One might naïvely believe that this could be dealt with by simply removing context variable C_α from consideration, leaving only context variables $C_\beta, \ldots, C_\iota$ 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_α, C_θ and C_ι together into a single context variable given by the triple $(C_\alpha, C_\theta, C_\iota)$. Accidentally, in this case this would be mathematically equivalent to the pair (C_θ, C_ι), but the interpretation of $(C_\alpha, C_\theta, C_\iota)$ is different from that of (C_θ, C_ι). 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 $C_\iota = 1$. Then, C_α becomes a constant, and constants can be safely ignored (or, trivially combined with any other context variable).[19]

To wrap up: 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 omitting 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 (or to both combined). Nonetheless, we can still use the inferred causal relations between system variables.

19. 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. 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.

C_α	C_β	C_γ	C_δ	C_ϵ	C_ζ	C_η	C_θ	C_ι	$N_{\boldsymbol{c}}$
1	0	0	0	0	0	0	0	0	853
1	1	0	0	0	0	0	0	0	902
1	0	1	0	0	0	0	0	0	911
1	0	0	1	0	0	0	0	0	723
1	0	0	0	1	0	0	0	0	810
1	0	0	0	0	1	0	0	0	799
1	0	0	0	0	0	1	0	0	848
0	0	0	0	0	0	0	1	0	913
0	0	0	0	0	0	0	0	1	707
1	1	1	0	0	0	0	0	0	899
1	1	0	1	0	0	0	0	0	753
1	1	0	0	1	0	0	0	0	868
1	1	0	0	0	1	0	0	0	759
1	1	0	0	0	0	1	0	0	927

C_β	C_γ	C_δ	C_ϵ	C_ζ	C_η	$(C_\alpha, C_\theta, C_\iota)$	$N_{\boldsymbol{C}}$
0	0	0	0	0	0	(1,0,0)	853
1	0	0	0	0	0	(1,0,0)	902
0	1	0	0	0	0	(1,0,0)	911
0	0	1	0	0	0	(1,0,0)	723
0	0	0	1	0	0	(1,0,0)	810
0	0	0	0	1	0	(1,0,0)	799
0	0	0	0	0	1	(1,0,0)	848
0	0	0	0	0	0	(0,1,0)	913
0	0	0	0	0	0	(0,0,1)	707
1	1	0	0	0	0	(1,0,0)	899
1	0	1	0	0	0	(1,0,0)	753
1	0	0	1	0	0	(1,0,0)	868
1	0	0	0	1	0	(1,0,0)	759
1	0	0	0	0	1	(1,0,0)	927

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

	Reagent	Intervention
C_α	α-CD3, α-CD28	global activator
C_β	ICAM-2	global activator
C_γ	AKT inhibitor	activity of AKT
C_δ	G0076	activity of PKC
C_ϵ	Psitectorigenin	abundance of PIP2
C_ζ	U0126	MEK activity
C_η	LY294002	PIP2, PIP3 mechanism change
C_θ	PMA	PKC activity
C_ι	β2CAMP	PKA activity

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).

4.2. Implementing JCI

Any causal discovery method that is applicable under the assumptions described in Section 4 can be used for Joint Causal Inference. In this section we will describe some concrete examples of JCI implementations. Identifiability may greatly benefit 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., 1999; Zhang, 2008a), 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. Often, simple adaptations of off-the-shelf implementations can be made that do allow

one to benefit from all JCI background knowledge. For example, in Section 4.2.4 we will propose a simple adaptation of FCI that does so.

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, taking into account the appropriate JCI background knowledge.

Any soundness, completeness and consistency results for the causal discovery algorithm that hold for the algorithm in the purely observational setting, but including the 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 13 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 JCI implementations. We will first describe two existing algorithms (LCD and ICP) that can be used off-the-shelf for causal discovery within the JCI framework. Then we propose two adaptations of existing algorithms (ASD and FCI) so that they can be used for causal discovery within the JCI framework. In Section 5, we will provide empirical results on the properties of those four algorithms.

4.2.1. Local Causal Discovery (LCD)

Perhaps the first implementation of JCI (apart from randomized controlled trials) 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 16 *Suppose that the data-generating process on three variables X_1, X_2, X_3 can be represented by a faithful, simple SCM $\mathcal{M}$ with $\mathcal{I} = \{1, 2, 3\}$ and that the sampling procedure is not subject to selection bias. If X_2 is not a cause of X_1 according to $\mathcal{M}$, the following conditional (in)dependencies in the observational distribution $\mathbb{P}_{\mathcal{M}}(X_1, X_2, X_3)$*

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

imply that the graph $\mathcal{G}(\mathcal{M})$ must be one of the three DMGs in Figure 12. In particular,
(i) X_3 is not a cause of X_2 according to $\mathcal{M}$;
(ii) X_2 is a direct cause of X_3 according to $\mathcal{M}$;
(iii) X_2 and X_3 are not confounded according to $\mathcal{M}$;

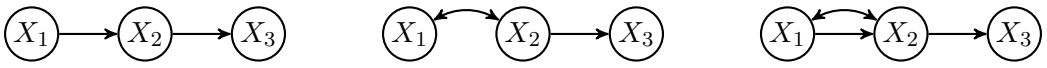


Figure 12: All possible DMGs detected by LCD.

(iv) the causal effect of X_2 on X_3 is given by:

$$\mathbb{P}_{\mathcal{M}}\big(X_3 \,|\, \mathrm{do}(X_2 = x_2)\big) = \mathbb{P}_{\mathcal{M}}(X_3 \,|\, X_2 = x_2). \tag{6}$$

Proof The proof proceeds by enumerating all (possibly cyclic) DMGs on three variables and ruling out the ones that do not satisfy the assumptions. The assumption that X_2 is not a cause of X_1 implies that there is no directed edge $X_2 \to X_1$ in the graph $\mathcal{G}(\mathcal{M})$. If there were an edge between X_1 and X_3, $X_1 \perp\!\!\!\perp X_3 \,|\, X_2$ would not hold (faithfulness). Also, since $X_1 \not\perp\!\!\!\perp X_2$, X_1 and X_2 must be adjacent (Markov property). 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 $X_1 \to X_2$ and $X_1 \leftrightarrow X_2$ (both of which have an arrowhead at X_2), this means that there must be a directed edge $X_2 \to X_3$, 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 graphs are the ones in Figure 12. The causal do-calculus applied to $\mathcal{G}(\mathcal{M})$ yields (6). ■

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.

A conservative version of LCD has been applied by Triantafillou et al. (2017) to the task of inferring signaling networks from mass-cytometry data. A high-dimensional version of LCD has been shown to be successful in predicting the effects of gene knockouts on gene expression levels (Versteeg and Mooij, 2019) from large-scale interventional yeast gene expression data (Kemmeren et al., 2014). 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 single nucleotide polymorphisms (SNPs) as context variables. Since the DNA content cannot be caused by gene expression levels, JCI Assumption 1 is satisfied. Chen et al. (2007) then argue that Mendelian randomization justifies 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.

4.2.2. Invariant Causal Prediction (ICP)

ICP exploits invariance of 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; however, that only works under the assumption of causal sufficiency and acyclicity. 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., 2018). ICP has been successfully applied to predict the effects of gene knockouts on gene expression levels (Meinshausen et al., 2016) from large-scale interventional yeast gene expression data (Kemmeren et al., 2014).

ICP can be interpreted as a particular implementation of the JCI framework, even in the general setting with nonlinear relations between variables and with latent confounders and cycles present (although faithfulness is then required). The following result broadens the conditions under which ICP identifies (possibly indirect) causal relations, strengthening results of Peters et al. (2016):

Corollary 17 *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 $i \in \mathcal{I}$:*

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

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

Proof Follows immediately from Proposition 28 in Appendix A.2. ■

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. Hence, ICP can be interpreted as a particular causal discovery algorithm implementing the JCI framework.

4.2.3. ASD-JCI

Here we introduce a novel JCI implementation that builds on the algorithm by Hyttinen et al. (2014) and its generalization to σ-separation (Forré and Mooij, 2018) and some of the extensions proposed by Magliacane et al. (2016b). Since adapting this algorithm to the JCI setting is straightforward, and the algorithm itself has been described in detail in the cited papers, we here only provide a brief description of how it works. More details can be found in Appendix C.

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)dependencies 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 λ 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 Answer Set Programming (ASP), a declarative programming language that can be used amongst others for solving discrete optimization problems. Forré and Mooij (2018) generalize the encoding to σ-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 or absent 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 α 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 dependencies.

As we will need an acronym for this algorithm later, we will henceforth refer to it as ASD (Accounting for Strong Dependencies), as it essentially tries to explain the observed dependencies in the data, taking into account the statistical strength of these dependencies. 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 dependencies.

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, as it removes the need to learn the causal relations of the context variables.

Since the JCI background knowledge can be completely and exactly encoded as constraints on the possible causal graphs, we can directly extend the known soundness, completeness and consistency results for ASD:

Theorem 18 *ASD-JCI is sound and complete for oracle inputs. It is asymptotically consistent if the weights are asymptotically consistent.*

Proof For the precise meaning of these statements, we refer the reader to Appendix C, and in particular, to Theorem 43 and 44. ■

4.2.4. FCI-JCI

Here we introduce an adaptation of the constraint-based causal discovery algorithm FCI (Spirtes et al., 1999; Zhang, 2008a) that can be used in a JCI setting. The FCI algorithm was designed to work under the assumption that the data was generated by an acyclic SCM. While FCI can deal with selection bias, we will assume here for simplicity that no selection bias is present.

The FCI algorithm consists of two main phases: an adjacency search phase leading to the 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, signify-

ing 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 (Triantafillou 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 provides a complete description of the Markov equivalence class (Zhang, 2008a; Ali et al., 2009). For a more detailed discussion of MAGs and PAGs, we refer the reader to Appendix B.1.

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 is adapted accordingly by not removing any edges between context variables; afterwards, all edges between context variables are oriented as $k \leftrightarrow k'$ for $k \neq k' \in \mathcal{K}$. In the subsequent phase of orienting unshielded triples, only system variables can take on the role of the collider.

- 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 are oriented as $k \ast\!\!\rightarrow i$, with an arrowhead at 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 $k \to i$ (since we are assuming no selection bias) in the MAG. Hence, after the adjacency phase, all edges between a context and a system variable are oriented as $k \to i$, 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.

We will refer to this adaptation of the FCI algorithm as `FCI-JCI`. In particular, we distinguish three variants: `FCI-JCI0` (which only makes JCI Assumption 0), `FCI-JCI1` (also JCI Assumption 1), and `FCI-JCI123` (also JCI Assumptions 1, 2, 3). In Appendix B.4 we discuss how one can read off the identified causal and non-causal relations from the PAG output by FCI or FCI-JCI. Furthermore, in Appendix B.5 we discuss how one can read off the direct targets and direct non-targets of interventions represented by context nodes from the PAG output by `FCI-JCI123`.

4.2.5. Speeding up `FCI-JCI123`

Adding the context nodes may make FCI-JCI considerably slower than FCI on a single context. In this subsection, we propose a further adaptation of `FCI-JCI123`. By exploiting the following observation, we can achieve a considerable speedup:

Lemma 19 *Let $\mathcal{M}$ be an SCM that satisfies JCI Assumptions 0, 1, 2. Then for $X \subseteq \mathcal{I}$ and $Y, Z \subseteq \mathcal{I} \cup \mathcal{K}$:*

$$X \underset{\mathcal{G}(\mathcal{M})}{\perp} Y \mid Z \implies X \underset{\mathcal{G}(\mathcal{M})}{\perp} Y \mid Z \cup (\mathcal{K} \setminus Y)$$

(for both d-separation as well as for σ-separation).

Proof By contradiction: Suppose there exists a path $\langle v_0, e_1, v_1, e_2, v_3, \ldots, e_{n-1}, v_n \rangle$ between X and Y that is open given $Z \cup (\mathcal{K} \setminus Y)$ and contains no non-endpoint nodes in $X \cup Y$, but is closed given Z. That can only happen if the path contains a collider v_i $(0 < i < n)$ that is $\mathcal{G}(\mathcal{M})$-ancestor of $\mathcal{K} \setminus (Y \cup Z)$. Then $v_i \in \mathcal{K}$ (because no system node is ancestor of $\mathcal{K}$ by JCI Assumption 1). Let $1 \leq j < i$ be the lowest index such that $v_l \in \mathcal{K}$ for all $l \in \{j, j+1, \ldots, i\}$. Then $v_{j-1} \leftarrow v_j$ on the path (by JCI Assumptions 1, 2), where v_{j-1} is in another strongly-connected component than v_j, and therefore $v_j \in \mathcal{K} \setminus Y$ blocks the path, which is a contradiction. ■

This implies that in the skeleton search of `FCI-JCI12` and `FCI-JCI123`, the search spaces for finding separating sets between pairs of nodes can be reduced. Indeed, instead of testing for each subset B of $A \subseteq \mathcal{I} \cup \mathcal{K}$ whether B d-separates node v from node w, one can test for each subset B of $A \setminus \mathcal{K}$ whether $B \cup \mathcal{K} \setminus \{v, w\}$ d-separates v from w.[20] For the next stages of the FCI algorithm, it does not matter *which* separating set is found (as long as *any* separating set is found if there is one), as follows from Zhang (2006, Lemma 3.2.1 and 3.2.2). This modification to the skeleton phase reduces the worst-case number of conditional independence tests by a factor that is exponential in the number of context variables. We will refer to the adapted version of `FCI-JCI123` that implements this modified version of the skeleton search as `FCI-JCI123r`.

4.2.6. Soundness, Completeness and Consistency Results for FCI-JCI

The FCI algorithm was shown to be sound and complete (Zhang, 2008a) for oracle inputs. In Appendix B, we prove:

Theorem 20 *Let $\mathcal{M}$ be an acyclic SCM that satisfies JCI Assumption 0. Assume that its distribution $\mathbb{P}_{\mathcal{M}}(\boldsymbol{X}, \boldsymbol{C})$ is faithful w.r.t. the graph $\mathcal{G}(\mathcal{M})$. Then, with input $\mathbb{P}_{\mathcal{M}}(\boldsymbol{X}, \boldsymbol{C})$:*

- *`FCI-JCI0` is sound and complete;*
- *`FCI-JCI1` is sound if $\mathcal{M}$ also satisfies JCI Assumption 1;*
- *`FCI-JCI12` is sound if $\mathcal{M}$ also satisfies JCI Assumptions 1 and 2;*
- *`FCI-JCI123` is sound and complete if $\mathcal{M}$ also satisfies JCI Assumptions 1, 2, 3.*

Here, sound *means that the output of the algorithm is a DPAG that contains the true* $\mathrm{DMAG}(\mathcal{M})$*, and* complete *means that all edge marks of the true* $\mathrm{DMAG}(\mathcal{M})$ *that can be identified from the (conditional) independences in $\mathbb{P}_{\mathcal{M}}(\boldsymbol{X}, \boldsymbol{C})$ and the JCI background knowledge have been oriented in the DPAG output by the algorithm.*

20. Of course, the power of the conditional independence test may be reduced when conditioning on many variables. However, if the number of contexts in the pooled data is small, one can design conditional independence tests that do not suffer from this problem.

Proof We refer the reader to Appendix B, and in particular to Theorems 35, 37, and 38. ■

In practice, one often does not have access to the joint distribution $\mathbb{P}_{\mathcal{M}}(\boldsymbol{X}, \boldsymbol{C})$, but only to a finite sample of it. In that case we have:

Corollary 21 *The FCI variants mentioned in Theorem 20 are also asymptotically consistent under the assumptions stated in Theorem 20 if the conditional independence test (including the choice of the threshold to decide between independence and dependence) is consistent.*

Proof Direct application of Lemma 36 in Appendix B. ■

4.3. Related Work

In this section we provide a more detailed comparison with 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.

4.3.1. Latent Confounders

Most score-based methods that combine multiple contexts (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; Yang et al., 2018) assume *causal sufficiency*, i.e., that no latent confounders are present. This simplifies the causal discovery problem considerably, but the assumption is likely violated in practice and may lead to wrong conclusions. This is well-known for causal discovery from a single observational data set, but also applies to the JCI setting.

4.3.2. Cycles

As we have seen in Proposition 10, 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 (see Proposition 16). 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 9). The method can be modified to use σ-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

	Latent confounders	Nonlinear mechanisms	Cycles	Perfect interventions	Mechanism changes	Activity interventions	Other context changes	Unknown intervention/context targets	Learns intervention/context targets	Global causal discovery	Different variables in each context	Combination strategy
(Fisher, 1935)	+	+	+	+	+	+	+	+	+	-	-	b
LCD (Cooper, 1997)	+	+	+	+	+	+	+	+	-	-	-	b
(Cooper and Yoo, 1999)	-	+	-	+	-	-	-	-	-	+	-	b
(Tian and Pearl, 2001)	-	+	-	-	+	-	+	-	-	+	-	b
(Sachs et al., 2005)	-	+	-	+	-	-	-	-	-	+	-	b
(Eaton and Murphy, 2007)	-	+	-	+	+	+	+	+	+	+	-	b
Trigger (Chen et al., 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
ICP (Peters et al., 2016)	+	+	+	+	+	+	+	+	-	-	-	b
(Zhang et al., 2017)	-	+	-	+	+	+	+	+	+	+	-	b
(Yang et al., 2018)	-	+	-	+	+	-	-	-	-	+	-	a/b
(Forré and Mooij, 2018)	+	+	+	+	-	-	-	-	-	+	+	a
Joint Causal Inference (this work)	+	+	+	+	+	+	+	+	+	+	-	b
FCI-JCI (this work)	+	+	?	+	+	+	+	+	+	+	-	b
ASD-JCI (this work)	+	+	+	+	+	+	+	+	+	+	-	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. When a feature is offered only under additional restrictive assumptions, it is indicated with a ± sign. Combination strategies (right-most column) 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.

computational reasons, 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 17 shows. The JCI framework in general allows for cycles, but requires its implementation to support this.

4.3.3. Selection Bias

The only causal discovery method for multiple data sets that is explicitly claimed 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 MAG defined on the union of all system variables. It performs conditional independence tests in each data set 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 does not 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. It can also deal with perfect interventions since Fisher's method is used to test for independence in *all* contexts (see also Figure 7).

4.3.4. 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, focusing 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 mechanism 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.

Another recent approach to constraint-based causal discovery in a JCI setting is the one by Yang et al. (2018); these authors propose the algorithm IGSP that can be seen as an implementation of JCI for causally sufficient, acyclic models with a diagonal experimental design under JCI Assumptions 1, 2, for mechanism changes with intervention targets assumed to be known. An advantage of IGSP over our JCI approach is that essentially no assumptions on the context distribution need to be made (apart from positivity) since it relies on a weaker faithfulness assumption.

Tian and Pearl (2001) were the first to consider *mechanism changes*. They deal with sequences of 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 also takes into account the background knowledge on the causal ordering of the system variables inferred from analysing the interventional data. 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 between contexts. 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.

4.3.5. 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 simply combining all context variables $\{C_k\}_{k\in\mathcal{K}}$ into a single tuple $\boldsymbol{C} = (C_k)_{k\in\mathcal{K}}$. However, this reduction to a single context variable typically loses 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.

4.3.6. 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_β causes the system variable X_0, at least for causal discovery algorithms that are tolerant to faithfulness violations.

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

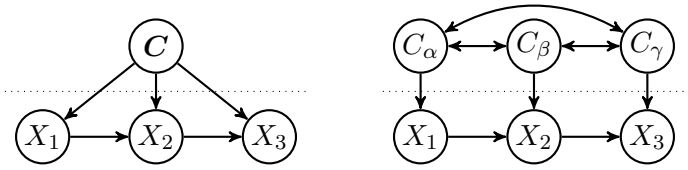


Figure 13: Example that shows that allowing multiple context variables (right) has advantages over considering a single context variable only (left). The causal graph on the right can be identified by JCI (with JCI Assumptions 1, 2, 3) from conditional independences in the pooled data, whereas the causal graph on the left is not identifiable.

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 dependencies (except $C_\alpha \not\perp\!\!\!\perp C_\beta$) and minimizes the number of faithfulness violations (faithfulness here implies $C_\beta \not\perp\!\!\!\perp X_1 \,|\, C_\alpha$), when (incorrectly) assuming that context variables are independent.

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.

4.3.7. Partially Overlapping Sets of Variables

A particular case of missing data that has been addressed by some of the methods is when the set of observed variables differ between data sets, while still having some overlap. 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 $\mathbb{P}(\boldsymbol{X} \,|\, \boldsymbol{C} = \boldsymbol{c})$ for $\boldsymbol{c} \in \boldsymbol{C}$ with $\mathbb{P}(\boldsymbol{C} = \boldsymbol{c}) > 0$ is faithful with respect to the marginalization of the same

DMG $\mathcal{G}_{\mathcal{I}}$ on the system variables that were observed in that context (or, in case of perfect interventions with known targets, the corresponding marginal intervened DMG).

4.3.8. 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 standard 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 (Forré and Mooij, 2019). Since we can always view the context variables as random variables in the *empirical distribution* of the experimental design (see also Footnote 17), this allows us to make use of the standard notion of conditional independence for the purposes of causal discovery.

4.3.9. Selection Diagrams

Our representation of a system within a context imposed by its environment also bears some similarities with selection diagrams (Bareinboim and Pearl, 2013). Selection diagrams have also been used for causal modeling in different contexts, 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 22 *Let* $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \mathcal{H}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ *and* $\mathcal{M}^* = \langle \mathcal{I}, \mathcal{J}, \mathcal{H}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}^*, \mathbb{P}^*_{\boldsymbol{\mathcal{E}}} \rangle$ *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 augmented graph* $\mathcal{H}$ *and hence also their graphs are identical,* $\mathcal{G}(\mathcal{M}) = \mathcal{G}(\mathcal{M}^*)$*. The* selection diagram $\mathcal{S}$ *induced by* $\langle \mathcal{M}, \mathcal{M}^* \rangle$ *is the acyclic directed mixed graph with nodes* $\mathcal{I} \dot{\cup} \overline{\mathcal{I}}$*, where* $\overline{\mathcal{I}} := \{\bar{i} : i \in \mathcal{I}\}$ *is a copy of* $\mathcal{I}$ *of* selection variable *indices, such that*

1. *the induced subgraph of* $\mathcal{S}$ *on* $\mathcal{I}$ *equals the common graph of* $\mathcal{M}$ *and* $\mathcal{M}^*$*, i.e.,* $\mathcal{S}_{\mathcal{I}} = \mathcal{G}(\mathcal{M}) = \mathcal{G}(\mathcal{M}^*)$*, and*
2. *for each* $i \in \mathcal{I}$ *such that* $f_i \neq f_i^*$ *or* $\mathbb{P}_{\boldsymbol{\mathcal{E}}_{\mathrm{PA}_{\mathcal{G}(\mathcal{M})}(i)}} \neq \mathbb{P}^*_{\boldsymbol{\mathcal{E}}_{\mathrm{PA}_{\mathcal{G}(\mathcal{M})}(i)}}$ *there is an edge* $\bar{i} \to i$ *in* $\mathcal{S}$*.*

From the definition, it is apparent that a selection diagram essentially models *two* contexts, and that the selection variables in the selection diagram correspond to the *children of the context variable* in our representation. Indeed, consider a JCI model of the form (5) (p. 21) 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 $i_1 \to i_2$ 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 $C \to i$ in $\mathcal{G}(\mathcal{M})$ for $i \in \mathcal{I}$ then $\bar{i} \to i$ 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 (features of) the selection diagram from data.

It is not clear how a selection diagram could be used to represent the same information that an 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.

5. Experiments

In this section we report on the experiments we performed with JCI, comparing various implementations of the framework 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. The source code that we used for producing the results and plots in this section is provided under a free and open source license as Online Appendix 1.

5.1. 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., 1999; Zhang, 2008a). 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 with some small extensions, also 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, σ-separation. Adapting FCI to the cyclic case seems less straightforward and is beyond the scope of this paper.[21]

Table 5 provides an overview of all implementations that we have studied here. The methods will be discussed in more detail in the next few subsections. The "CI Tests" column in the table 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 $\tilde{X}_a \perp\!\!\!\perp \tilde{X}_b \,|\, \tilde{\boldsymbol{X}}_S$ 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 $X_a \perp\!\!\!\perp X_b \,|\, \boldsymbol{X}_S$ with $\{a\} \cup \{b\} \cup S \subseteq \mathcal{I}$ and $\{a\}, \{b\}, S$ mutually disjoint.

21. The d-separation case has recently been addressed by Strobl (2018).

SS : system variables only, separately for each context; the conditional independence tests performed are of the form $X_a \perp\!\!\!\perp X_b \,|\, \boldsymbol{X}_S, \boldsymbol{C} = \boldsymbol{c}$ for $\{a\} \cup \{b\} \cup S \subseteq \mathcal{I}$ and $\{a\}, \{b\}, S$ mutually disjoint, for all values $\boldsymbol{c} \in \boldsymbol{\mathcal{C}}$ with $\mathbb{P}(\boldsymbol{c}) > 0$. Note that this method assumes that the context domain $\boldsymbol{\mathcal{C}}$ is discrete.

SF : system variables only, separately for each context, using Fisher's method;[22] the conditional independence tests performed are of the form $X_a \perp\!\!\!\perp X_b \,|\, \boldsymbol{X}_S, \boldsymbol{C} = \boldsymbol{c}$ for $\{a\} \cup \{b\} \cup S \subseteq \mathcal{I}$ and $\{a\}, \{b\}, S$ mutually disjoint, for all values $\boldsymbol{c} \in \boldsymbol{\mathcal{C}}$ with $\mathbb{P}(\boldsymbol{c}) > 0$. Note that this method assumes that the context domain $\boldsymbol{\mathcal{C}}$ is discrete.

NC : test all variables, except for conditional independences between context variables; the conditional independence tests performed are of the form $X_a \perp\!\!\!\perp \tilde{X}_b \,|\, \tilde{\boldsymbol{X}}_S$ with $a \in \mathcal{I}$, $\{b\} \cup S \subseteq \mathcal{I} \cup \mathcal{K}$ and $\{a\}, \{b\}, S$ mutually disjoint.

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

5.1.1. 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) in the acyclic case and the σ-separation adaptation of Forré and Mooij (2018) in the cyclic case, 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 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. `ASD-JCI123sc` and `ASD-JCI1sc` both use a single merged context variable $\boldsymbol{C} = (C_k)_{k \in \mathcal{K}}$, 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).

22. 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 χ^2 distribution with $2N$ degrees of freedom if all N p-values p_i are independent.

5.1.2. FCI Variants

We implemented different variants of the FCI algorithm by adapting the implementation in the `R` package `pcalg` (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 (2008a) can be ignored in the FCI algorithm, and only PAGs without (possibly) 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 4.2.4, for three different combinations of JCI Assumptions (we have not yet implemented and evaluated the speedup for `FCI-JCI123r` but leave that for future work).

The output of FCI is a PAG. Here we will not evaluate the PAG itself, since estimating the PAG is often not the ultimate task in causal discovery, but instead we will evaluate presence and absence of ancestral relations that can be identified from the estimated PAG. The procedure we use for this task is explained in Appendix B.4. For the special case of `FCI-JCI123`, we also read off the direct intervention targets (and non-targets) from the estimated PAG, as explained in Appendix B.5. We encode identified presence of a feature by $+1$, identified absence by -1, and an unidentifiable feature by 0. These predictions can then also easily be bootstrapped (or more precisely, bagged). The bagged feature scores can then be used as a score for the confidence that the feature is present.

5.1.3. 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 $\langle C, X, Y\rangle$ (see also Figure 12, where $\langle X_1, X_2, X_3\rangle$ corresponds with $\langle C, X, Y\rangle$), we use $-\log p_{C \perp\!\!\!\perp Y}$. Note that LCD predicts the presence of an ancestral relation $X \in \text{AN}(Y)$, the absence of a confounder between X and Y, and the absence of a direct causal effect of C on Y.[23]

5.1.4. ICP Variants

We also compare with the `ICP` function in the `R` package `InvariantCausalPrediction` (Peters et al., 2016).[24] Under the assumption of causal sufficiency, ICP returns direct causal relations. However, in our setting, in which causal sufficiency cannot be assumed, ICP will generally return ancestral causal relations, as shown in Corollary 17. Note that ICP assumes a single context variable, whereas one typically may have multiple context

23. The absence of a bidirected edge $X \leftrightarrow Y$ in the marginalization of the 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 $C \to Y$ in this marginalization implies the absence of the direct edge $C \to Y$ in the full graph $\mathcal{G}$.
24. We used the default arguments, except that we set `stopIfEmpty` to `TRUE`.

variables in the 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 ICP performs are of the form $\boldsymbol{C} \perp\!\!\!\perp X_i \,|\, \boldsymbol{X}_S$ 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 $\boldsymbol{X}_S$ for each context $\boldsymbol{C} = \boldsymbol{c}$ individually, and once globally. It then tests whether there exists a context $\boldsymbol{c}$ in which the mean or the variance of the regression residuals is different from the global mean or variance of the residuals. All p-values are then combined using a Bonferroni correction. Note that this test can detect more conditional dependencies than a simple partial correlation test, as it also considers the variation between the variances across contexts.

As a confidence score for the ancestral causal relation $i \in \text{AN}(j)$, we use $-\log p_{i \dashrightarrow j}$, where $p_{i \dashrightarrow j}$ is the p-value returned by ICP for system variable i being ancestor of system variable j.

5.1.5. 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 $X_i \perp\!\!\!\perp C_k \,|\, \boldsymbol{C}_{\mathcal{K} \setminus \{k\}}$ on the pooled data resulting in p-value p_{ik}. It is limited to discovery of ancestral causal relations from context to system variables. As confidence value for the ancestral causal relation $k \in \text{AN}(i)$ we report $-\log p_{ik} + \log \alpha$, where α is the threshold for the independence test.

5.1.6. 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 (or direct intervention target, for `FCI-JCI123`). For LCD, we average $-\log p_{C \perp\!\!\!\perp Y}$ 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`".

5.1.7. 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. However, the JCI framework imposes no principled restrictions on the conditional independence tests used, so one could also use non-parametric tests instead, for example.

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

Name	Data	Context variables	JCI Assumptions 1	2	3	Interventions	CI Tests
Baselines							
`ASD-obs`	observational	none	−	−	−	none	S
`ASD-pooled`	pooled	none	−	−	−	any	S
`ASD-meta`	all	none	−	−	−	any	SF
`ASD-pikt`	all	none	−	−	−	perfect, KT	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, KT	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

Table 5: Variants of implemented JCI algorithms and baselines used in our experiments. JCI Assumption 0 is always assumed. "KT" is an abbreviation of "known targets". 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 5.1.7. Bootstrapped versions of methods will be indicated with a suffix "`-bs`" (and have been omitted from this table for clarity).

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

- Otherwise, we go through all observed values $\boldsymbol{c}_{S\cap\mathcal{K}}$ of $\boldsymbol{C}_{S\cap\mathcal{K}}$, and use a standard partial correlations test to calculate a p-value for $\tilde{X}_a \perp\!\!\!\perp \tilde{X}_b \,|\, \tilde{\boldsymbol{X}}_{S\setminus\mathcal{K}}, \boldsymbol{C}_{S\cap\mathcal{K}} = \boldsymbol{c}_{S\cap\mathcal{K}}$. We then aggregate the p-values corresponding to observed values of $\boldsymbol{C}_{S\cap\mathcal{K}}$ into one overall p-value for $\tilde{X}_a \perp\!\!\!\perp \tilde{X}_b \,|\, \tilde{\boldsymbol{X}}_S$ using Fisher's method for aggregating p-values.[25]

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 implementation in the `R` package `InvariantCausalPrediction`. 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 $\tilde{X}_a \perp\!\!\!\perp \tilde{X}_b \,|\, \tilde{\boldsymbol{X}}_S$, it distinguishes the following cases:

- If $(\{a\} \cup \{b\} \cup S) \cap \mathcal{K} = \emptyset$, it reduces to a standard partial correlations test.

- If $\tilde{X}_a = \boldsymbol{C}$ is the context variable, it uses linear regression to fit $\tilde{X}_b$ as a linear function of $\boldsymbol{X}_S$, using data pooled over all contexts, and calculates the corresponding residuals. It then goes through all observed values $\boldsymbol{c}$ of $\boldsymbol{C}$, and tests whether the residuals in context $\boldsymbol{c}$ have a different distribution than the residuals in the other contexts (i.e., with $\boldsymbol{C} \neq \boldsymbol{c}$). This two-sample test is performed by comparing the means by a t-test, and the variances 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.[26]

- If $\tilde{X}_b = \boldsymbol{C}$ is the context variable, proceed similarly as in the previous case.

- If the context variable $\boldsymbol{C}$ is part of $\tilde{X}_S$, we go through all observed values $\boldsymbol{c}$ of $\boldsymbol{C}$, and use a standard partial correlations test to calculate a p-value for $X_a \perp\!\!\!\perp X_b \,|\, \boldsymbol{X}_{S\setminus\mathcal{K}}, \boldsymbol{C} = \boldsymbol{c}$. We then aggregate the p-values corresponding to observed values of $\boldsymbol{C}$ into one overall p-value for $X_a \perp\!\!\!\perp X_b \,|\, \tilde{\boldsymbol{X}}_S$ using Fisher's method for aggregating p-values.

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 subsequently combines all these p-values by adding the $\log p$-values, but taking into account how the graph structure is changed through perfect interventions.

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

25. Not to be confused with Fisher's test for causality that we described in Section 5.1.5.
26. This is the default test for continuous data in the `ICP` function of the `R` package `InvariantCausalPrediction`.

5.2. 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 interventions of two different intervention types: mechanism changes, and perfect interventions.

Random causal graphs were simulated by drawing directed edges independently between system variables with probability ϵ. For the acyclic models, we only allowed directed edges $i_1 \to i_2$ for $i_1 < i_2$ with $i_1, i_2 \in \mathcal{I}$. For cyclic models, we allowed directed edges $i_1 \to i_2$ 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 η, 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 exogenous 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 $k \in \{1, \ldots, q\}$), and the other q contexts corresponding with one of the context variables turned on (say $C_{k'} = 1$ for some $k' \in \{1, \ldots, q\}$) and the others turned off ($C_k = 0$ for the other $k \in \{1, \ldots, q\} \setminus \{k'\}$). We either took all interventions to be mechanism changes, or all interventions to be perfect. 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. This was done for each random SCM separately.

5.3. 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 between system variables. 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).

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 or 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. In the ROC curves, we use solid lines for positive scores (feature present) and negative scores (feature absent), and dotted lines for vanishing scores (feature presence/absence is unknown).

5.4. Results: Small Simulated 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 $N_{\boldsymbol{c}} = 500$ samples for each context, i.e., $N = 500(q + 1)$ samples in total.

5.4.1. 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 15 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. This is partially due to the small sample size, but also to the fact that many relationships are simply not identifiable from purely observational data alone. `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 significant improvement in precision that JCI can yield.

Figure 16 shows a similar picture in the cyclic setting, where all ASD variants make use of σ-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

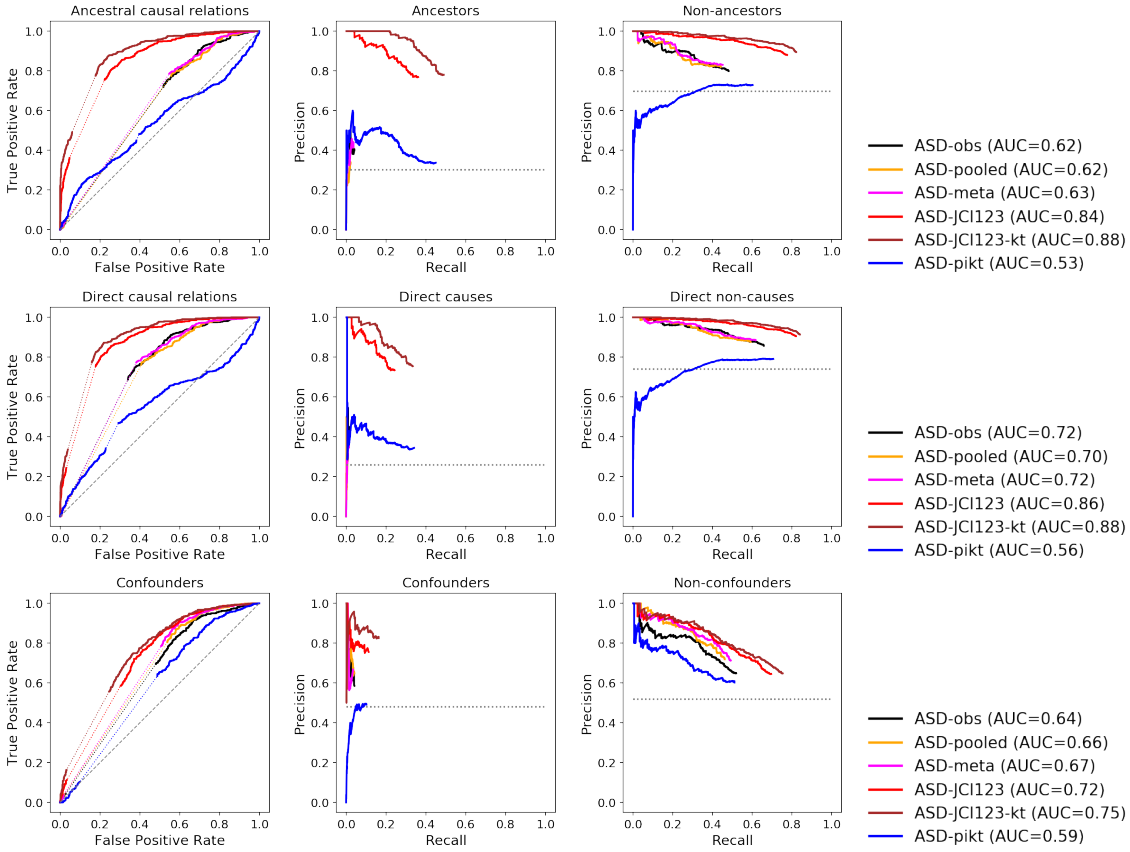


Figure 15: **Results of some ASD variants (acyclic, causal mechanism changes)** for small models. The two JCI variants (with unknown/known intervention targets) strongly outperform the baselines in this setting. From top to bottom: ancestral relations, direct causes, confounders. From left to right: ROC curves, PR curves for presence of feature, PR curves for absence of feature.

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

5.4.2. ASD-JCI vs. Baselines (Perfect Interventions)

Figures 17 and 18 show results for respectively the acyclic and cyclic settings, but now for perfect 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 perfect nature of the interventions,

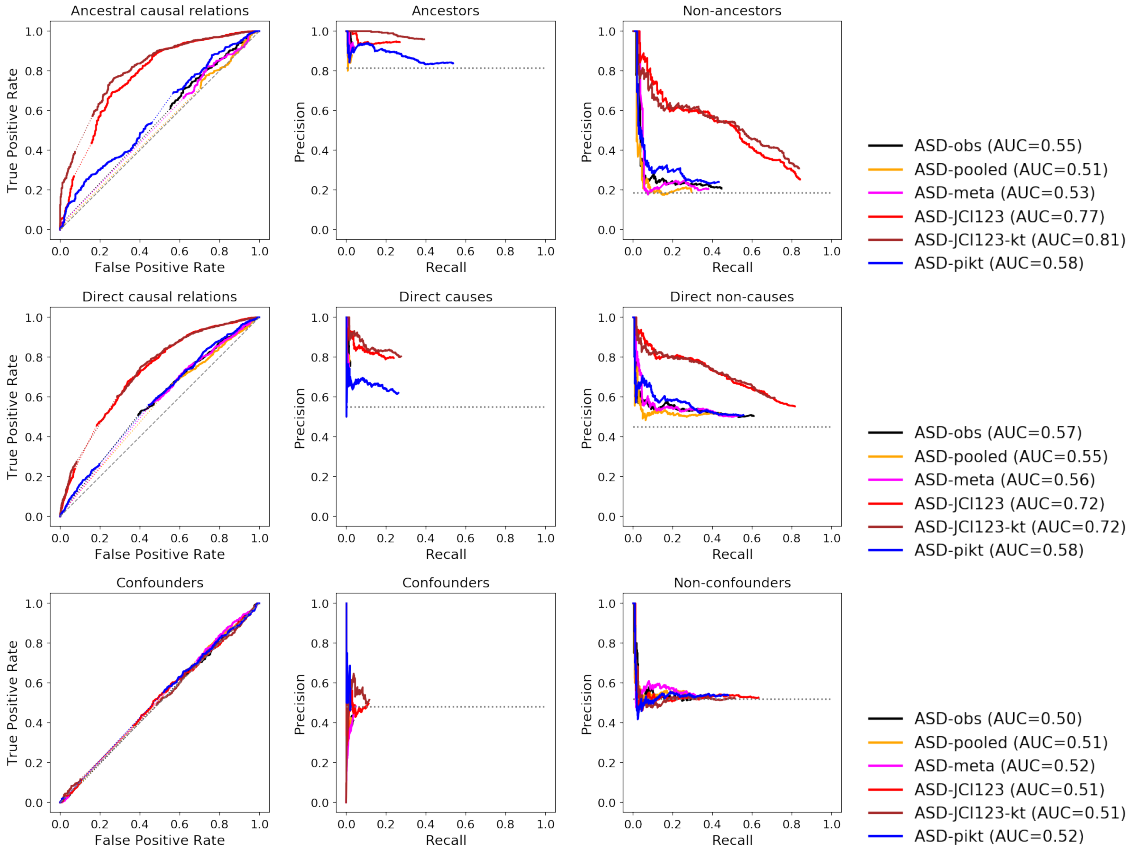


Figure 16: **Results of some ASD variants (cyclic, causal mechanism changes)** for small models. The two JCI variants substantially outperform the baselines in this setting.

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 5.4.1, `ASD-pikt` then breaks down, in contrast to the JCI variants.

For the cyclic setting with perfect interventions (Figure 18), 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). `ASD-pikt` again obtains the highest precisions, followed by the JCI variants.

5.4.3. 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 19 shows that, as expected, the more prior knowledge about the context variables is used, the better the predictions become.

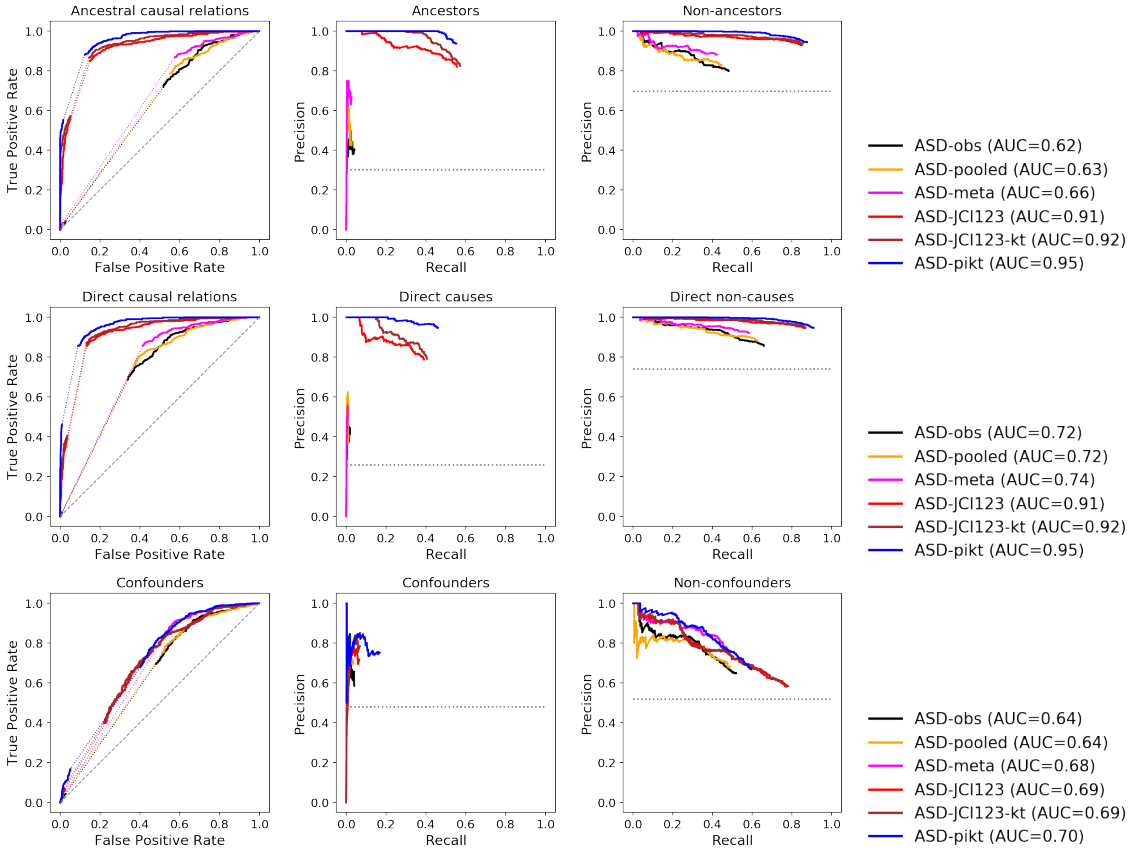


Figure 17: **Results of some ASD variants (acyclic, perfect interventions)** for small models. `ASD-pikt`, which takes into account the perfect nature of the interventions, is the best performing method in this setting. The JCI variants do not assume perfect interventions, but still yield a vast improvement of the precision of the predicted features with respect to the other baselines.

However, surprisingly, the largest 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 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 because we obtained qualitatively similar results for the other simulation settings.

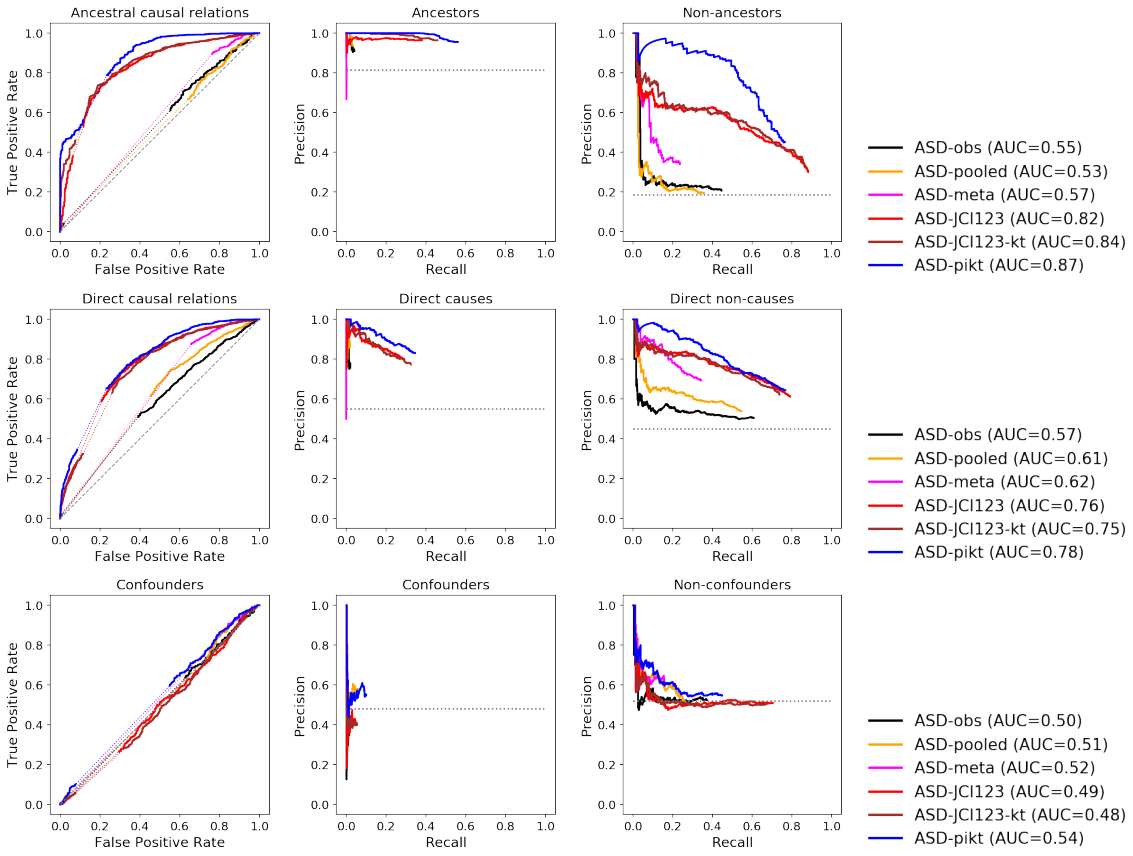


Figure 18: **Results of some ASD variants (cyclic, perfect interventions)** for small models. We see a similar picture as in the acyclic case in Figure 17.

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).

5.4.4. Multiple Context Variables vs. Single (Merged) Context Variable

Figure 20 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).

5.4.5. FCI Variants

Figure 21 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

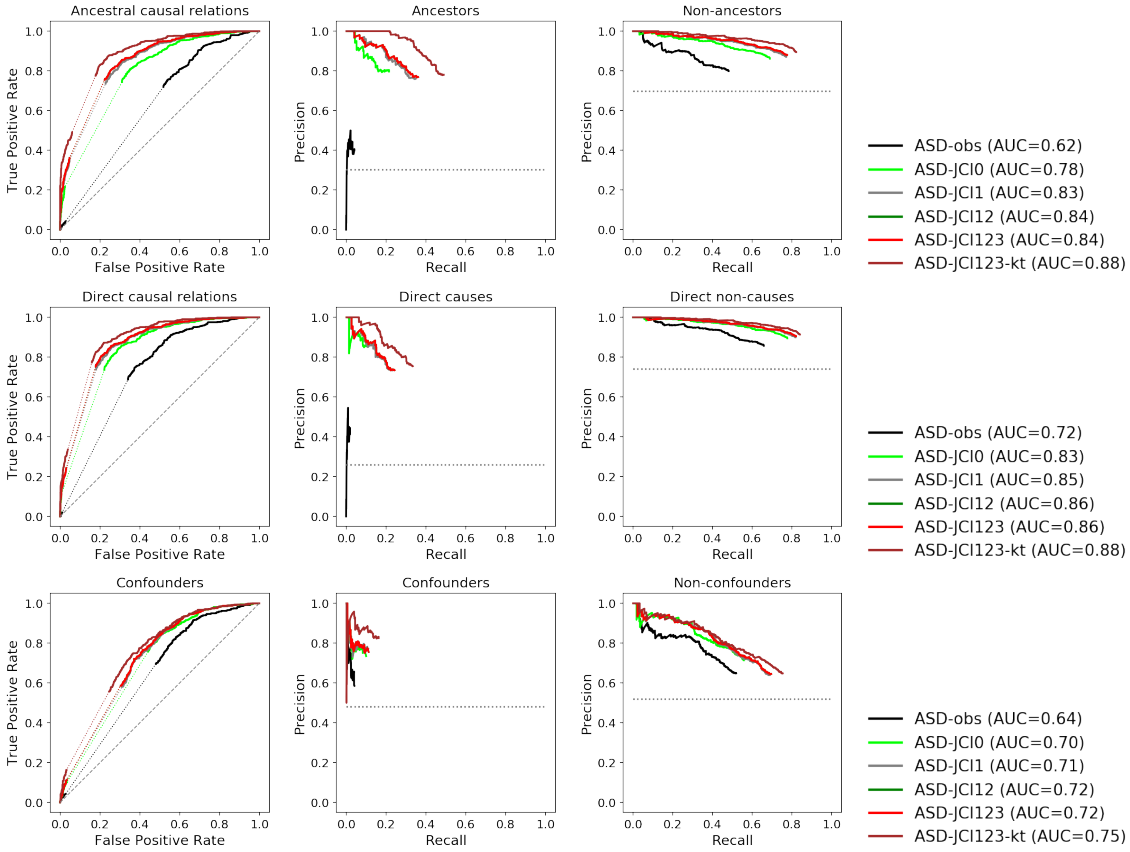


Figure 19: **Influence of JCI assumptions for ASD-JCI (acyclic, causal mechanism changes)** for small models. Exploiting more prior knowledge is seen to lead to better results. For reference, the non-JCI baseline `ASD-obs` is shown which only uses observational data.

to ASD, we conclude that the JCI variants of FCI substantially outperform the non-JCI variants. `FCI-JCI1` and `FCI-JCI123` seem to yield identical results in this setting.

Results for causal mechanism changes are very similar to those for perfect 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.

5.4.6. LCD and ICP

Figure 22 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. A possible explanation of this phenomenon could be that a combination of conditional independence tests of the form $C_k \perp\!\!\!\perp X_{i'} \,|\, X_i$ with each C_k bi-

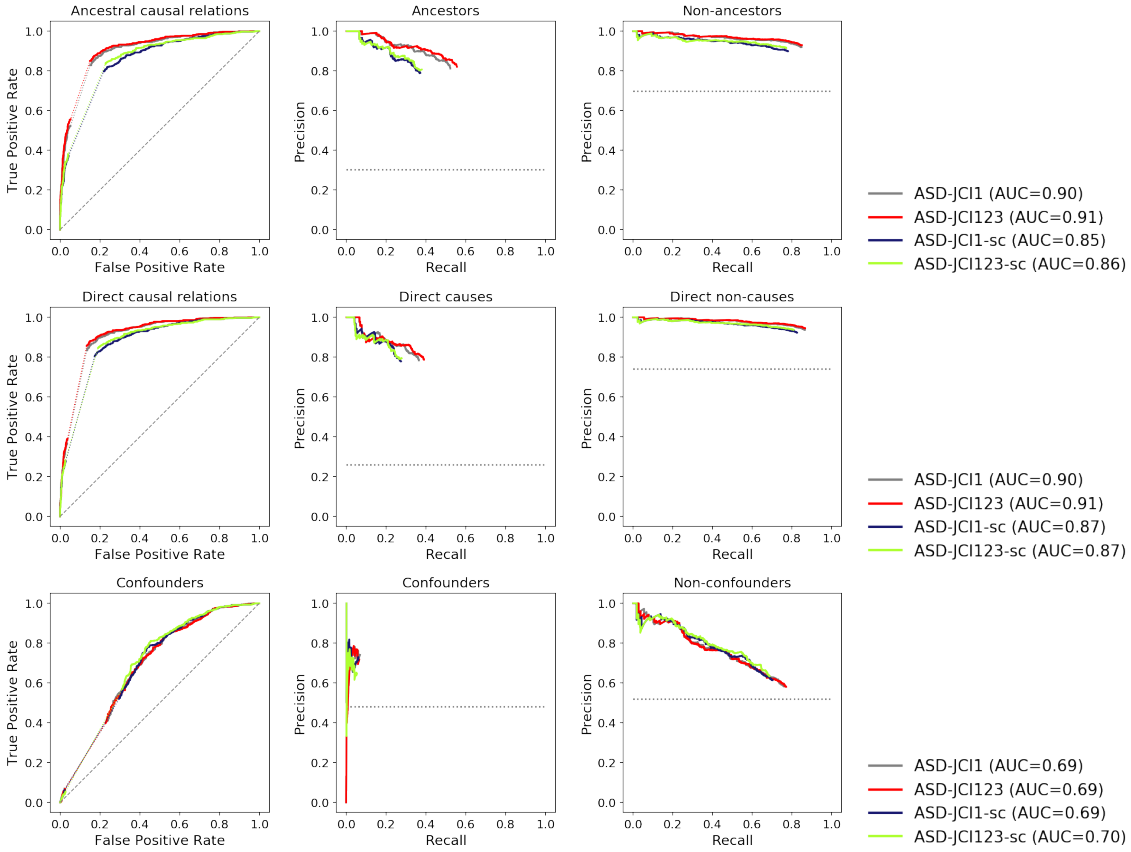


Figure 20: **Multiple context variables vs. single (merged) context variable for ASD-JCI (acyclic, perfect interventions)** for small models. Exploiting multiple context variables leads to better results than using a single (merged) context variable.

nary (and both X_i and $X_{i'}$ real-valued) might be less reliable than a single test $\boldsymbol{C} \perp\!\!\!\perp X_{i'} \,|\, X_i$ where $\boldsymbol{C}$ is categorical with $\gg 2$ states. Another observation we made is that bootstrapping does lead to only marginal improvements for these methods (not shown).

5.4.7. 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 23 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. This is in line with our expectation that jointly analyz-

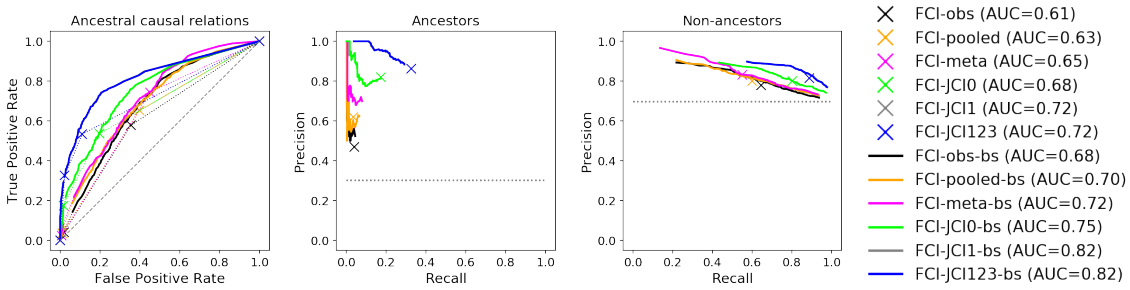


Figure 21: **Results of FCI variants (acyclic, perfect interventions)** for small models.

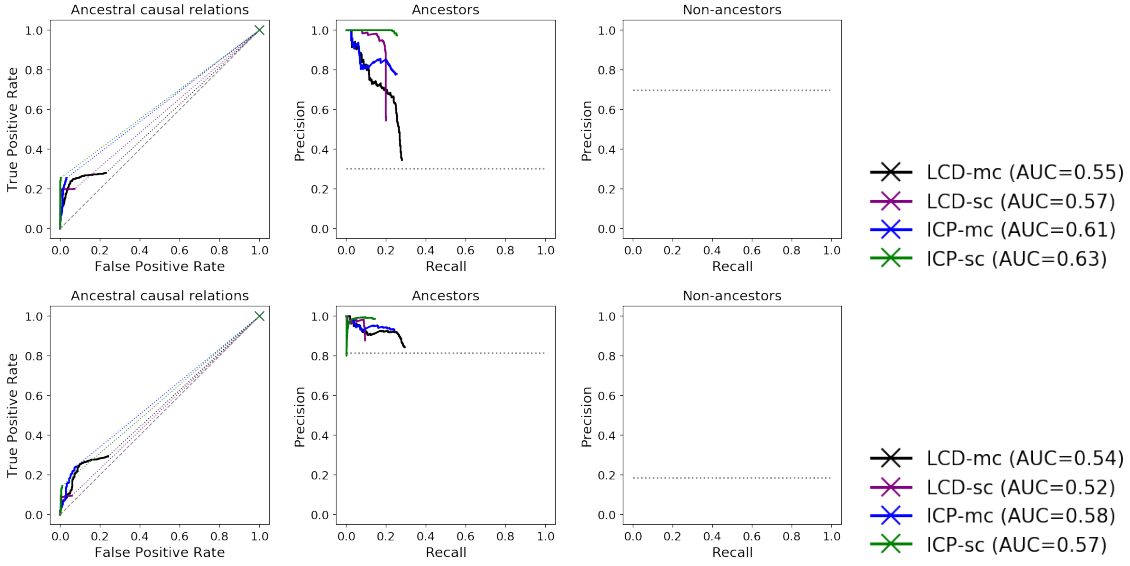


Figure 22: **Multiple context variables vs. single (merged) one for `LCD` and `ICP` (perfect interventions)** for small models. LCD and ICP both benefit from merging the context variables into a single one for the task of predicting the presence of ancestral relations. Top: acyclic; bottom: cyclic.

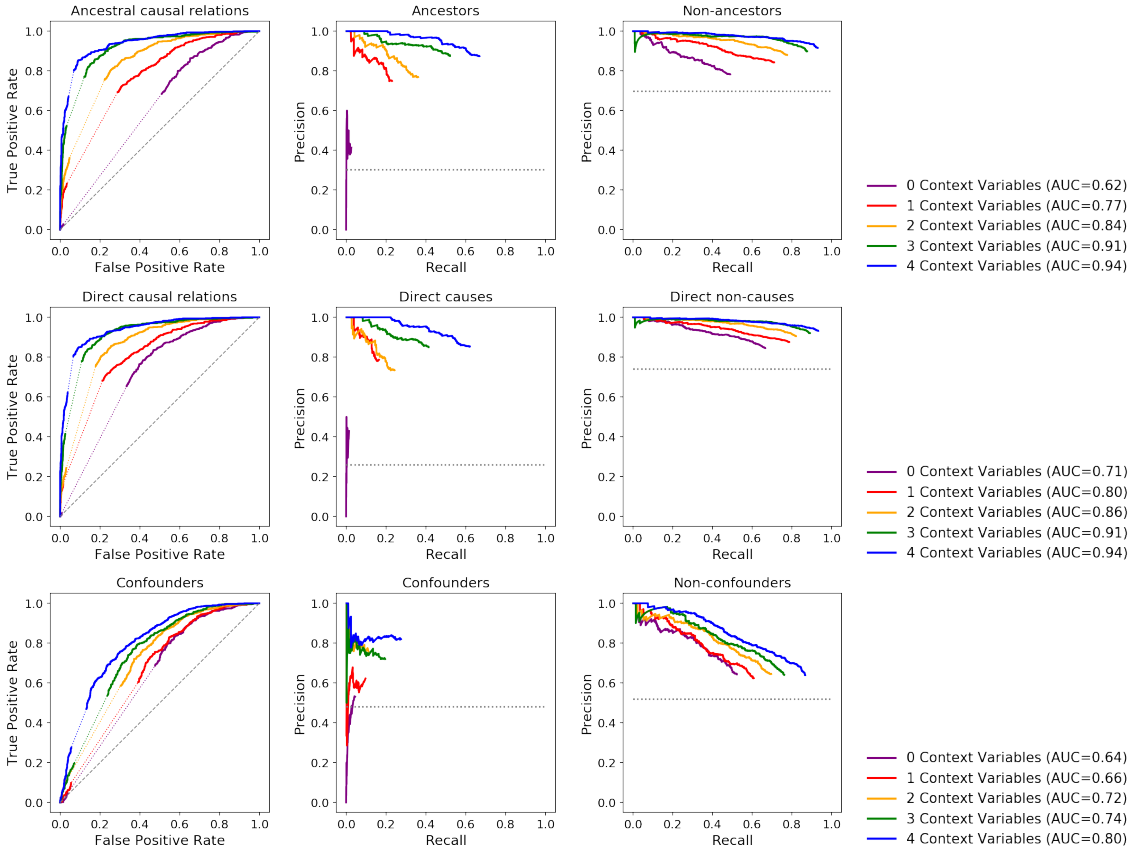


Figure 23: **Results for different numbers of context variables for `ASD-JCI123` (acyclic, causal mechanism changes)** for small models. Taking into account more context variables leads to better predictions.

ing data from multiple experiments makes it easier to estimate the causal structure of the system.

Interestingly, 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 `ICP-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.

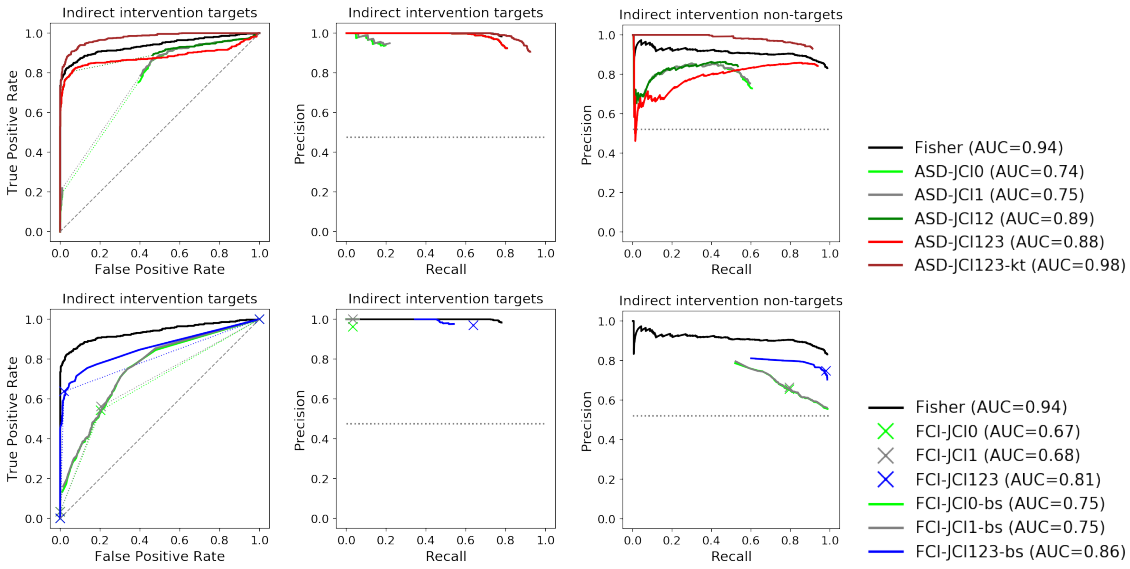


Figure 24: **Discovering indirect intervention targets (acyclic, causal mechanism changes)** on small models. Top: ASD variants; Bottom: FCI variants.

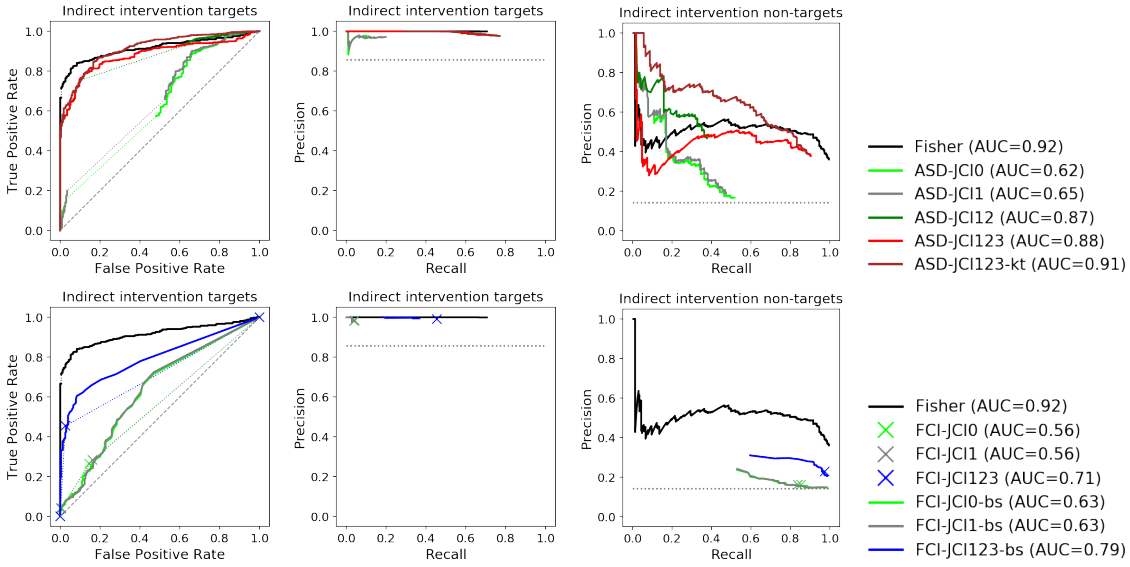


Figure 25: **Discovering indirect intervention targets (cyclic, causal mechanism changes)** on small models. Top: ASD variants; Bottom: FCI variants.

5.4.8. Discovering Indirect Intervention Targets

Figure 24 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.[27] LCD and ICP also cannot address this task.[28] 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 25 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, 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.

5.4.9. Discovering 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, as well as `FCI-JCI123`. Figure 26 shows the results for these algorithms in 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. Surprisingly, `FCI-JCI123` obtains almost perfect precision in all scenarios, outperforming `ASD-JCI123` notably in the cyclic cases. We do not understand why this is the case.

5.4.10. Computation Time

So far we have only considered the accuracy of the predictions. Another interesting aspect is the computation time that various methods need. Figure 27 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

27. 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.
28. However, when assuming also JCI Assumption 2, both LCD and ICP could be used to learn indirect intervention targets.

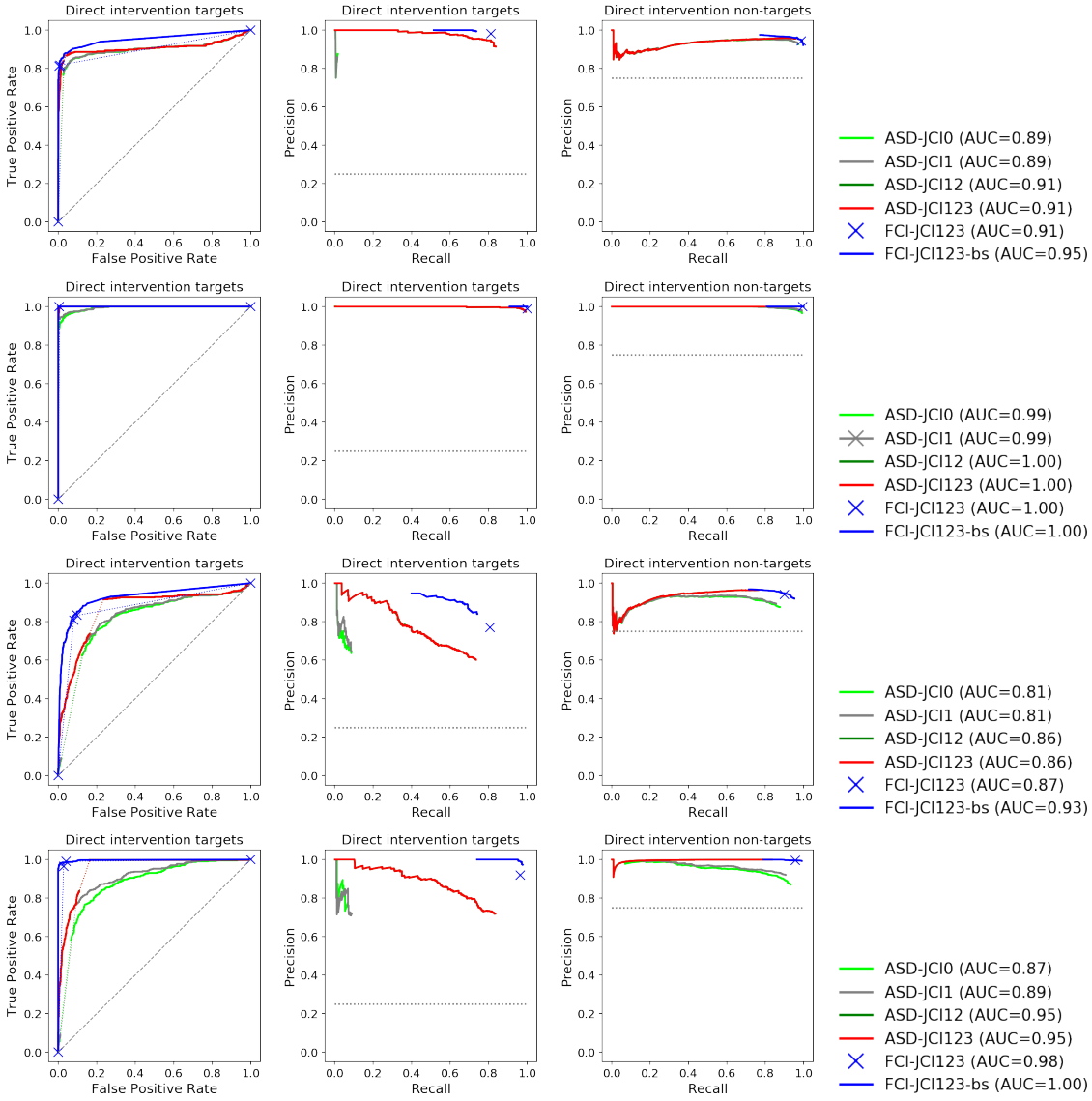


Figure 26: **Discovering direct intervention targets** on small models. From top to bottom: acyclic, causal mechanism changes; acyclic, perfect interventions; cyclic, causal mechanism changes; cyclic, perfect interventions.

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,

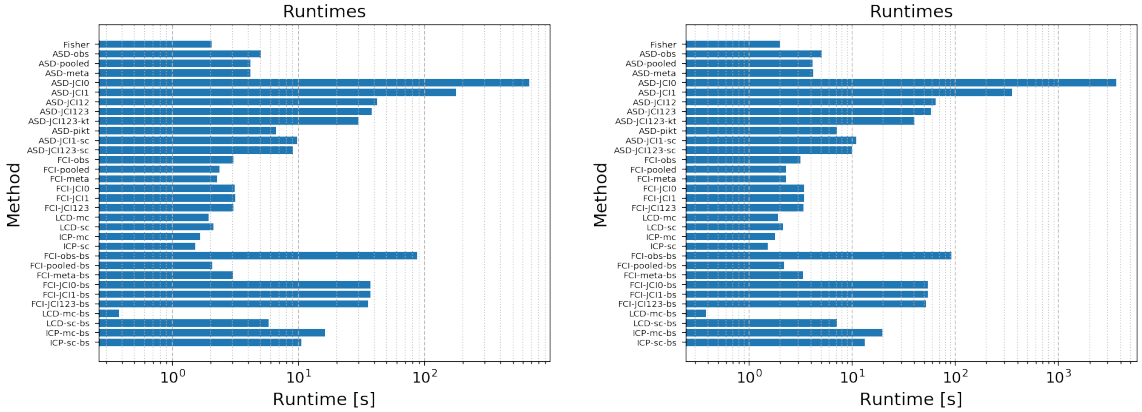


Figure 27: **Runtimes for various methods** on small models. Shown are runtimes for causal mechanism changes; for perfect interventions, runtimes are similar. Left: acyclic; Right: cyclic.

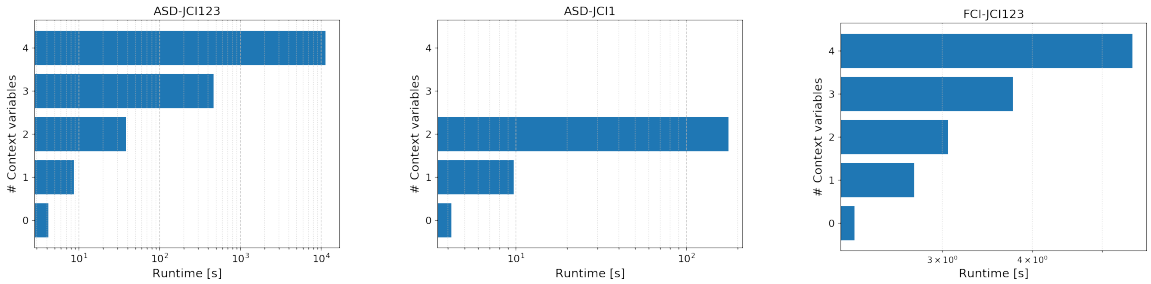


Figure 28: **Runtimes for three different algorithms as a function of the number of context variables** on small models. Note the considerably different ranges of the (logarithmic) x-axis. Results are omitted if the computation took too long to finish.

but bootstrapping also takes its toll. The fastest methods are Fisher's test, LCD and ICP. Figure 28 shows how computation time scales with the number of context variables, for three JCI implementations (`ASD-JCI123`, `ASD-JCI1` and `FCI-JCI123`).

5.5. Results: Larger Simulated Models

We now present results for larger models, with $p = 10$ system variables and $q = 10$ context variables (the meaning of the simulation parameters is explained in Section 5.2). 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

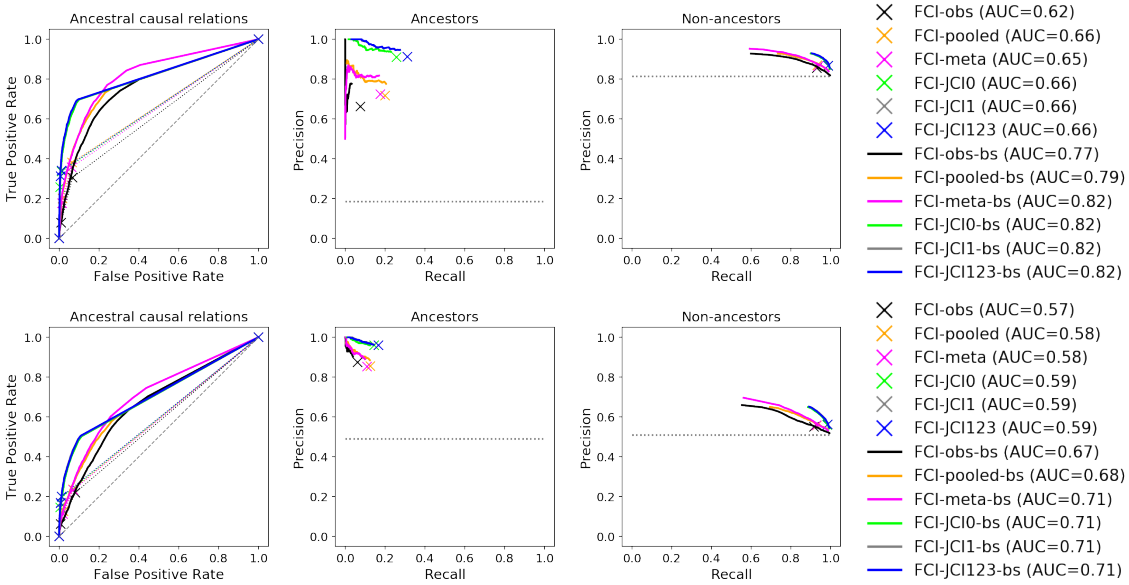


Figure 29: **FCI results for discovering ancestral causal relations between system variables** in larger models. Top: acyclic; bottom: cyclic.

in both scenarios. The motivation for these parameter choices is that they are somewhat comparable to the setting of the real-world data set that we will study in Section 5.8.

For these larger models, computation time for bootstrapped FCI methods became prohibitive for the default conditional independence test (described in Section 5.1.7). Although we can speed up the implementation of this test that we were using considerably by implementing it more efficiently, we here simply replaced it by a standard partial correlation test. This led to a speedup of about one order of magnitude at no apparent loss of accuracy.

5.5.1. FCI Variants

Figure 29 shows the accuracy for the task of predicting ancestral causal relations between system variables for various FCI-JCI variants and for various FCI baselines, in both acyclic and cyclic settings. 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 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 was designed for the acyclic case.

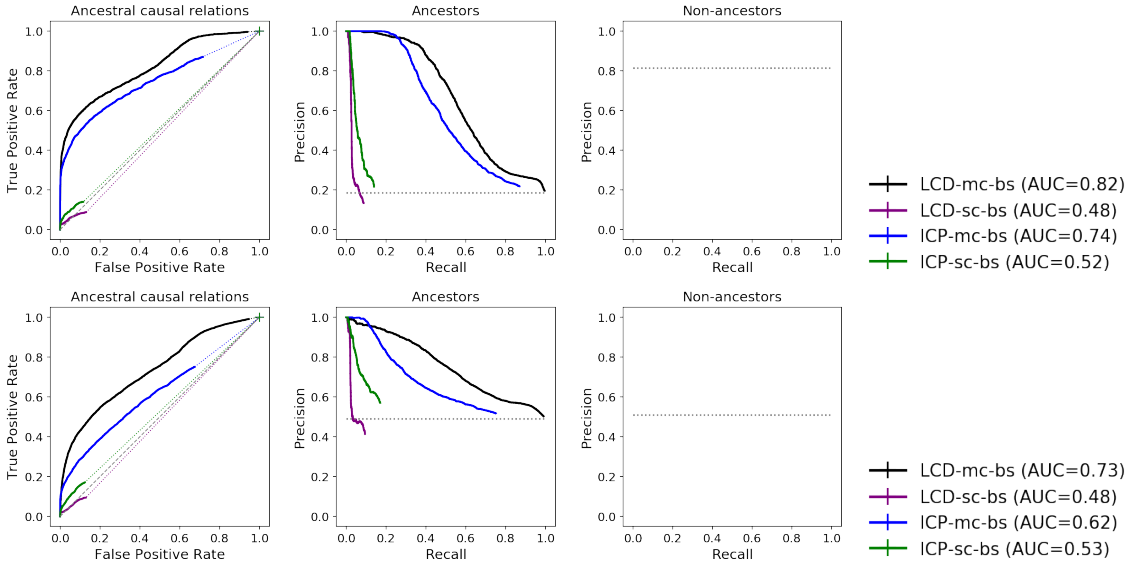


Figure 30: **Bootstrapped LCD and ICP results for discovering ancestral causal relations between system variables** for larger models. Top: acyclic; bottom: cyclic.

5.5.2. LCD and ICP

In Figure 30, 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. Contrary to what we observed for small models, the "multiple context" ("`-mc`") versions of both algorithms now clearly outperform the versions that use only a single (merged) context ("`-sc`") in these settings. Interestingly, the accuracy of LCD is quite similar to that of ICP. The additional complexity of ICP apparently does not lead to substantially better results than the LCD algorithm already offers in these settings. Also, the precision of `LCD-mc` is comparable to that of `FCI-JCI123`, the most accurate of the JCI variants of FCI (cf. Figure 29), on the task of predicting the presence of ancestral relations.

5.5.3. Discovering Intervention Targets

Figure 31 shows the performance of FCI-JCI variants (and as a baseline, Fisher's test) on the task of discovering indirect intervention targets, for both the acyclic and cyclic setting. Interestingly, JCI Assumption 2 seems necessary to obtain good results on this task. Still, `FCI-JCI123-bs` is outperformed by Fisher's test.

On the other hand, Fisher's test cannot identify direct intervention targets, whereas `FCI-JCI123` can. Figure 32 shows that `FCI-JCI123` can identify the direct intervention targets as well as the non-targets with high precision. Surprisingly, this works also in the cyclic case.

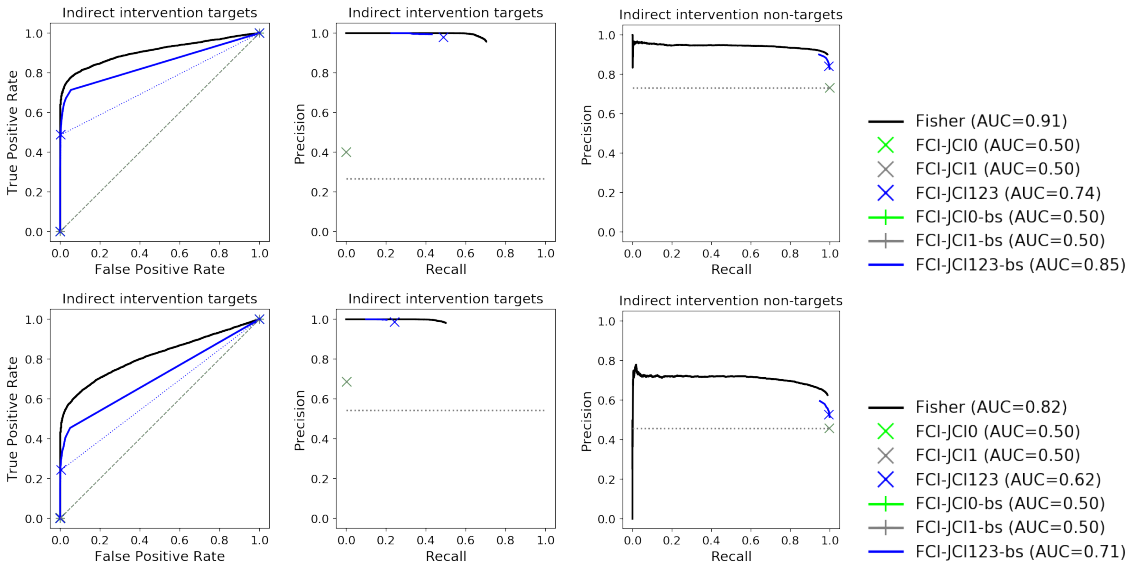


Figure 31: **FCI results for discovering indirect intervention targets** in larger models. Top: acyclic; bottom: cyclic.

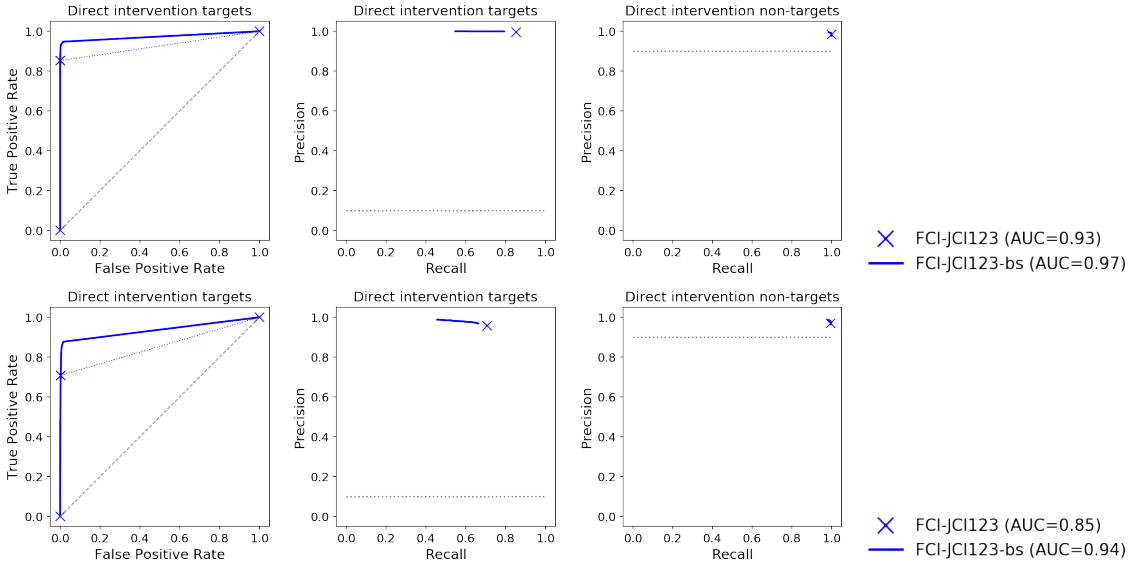


Figure 32: **FCI results for discovering direct intervention targets** in larger models. Top: acyclic; bottom: cyclic.

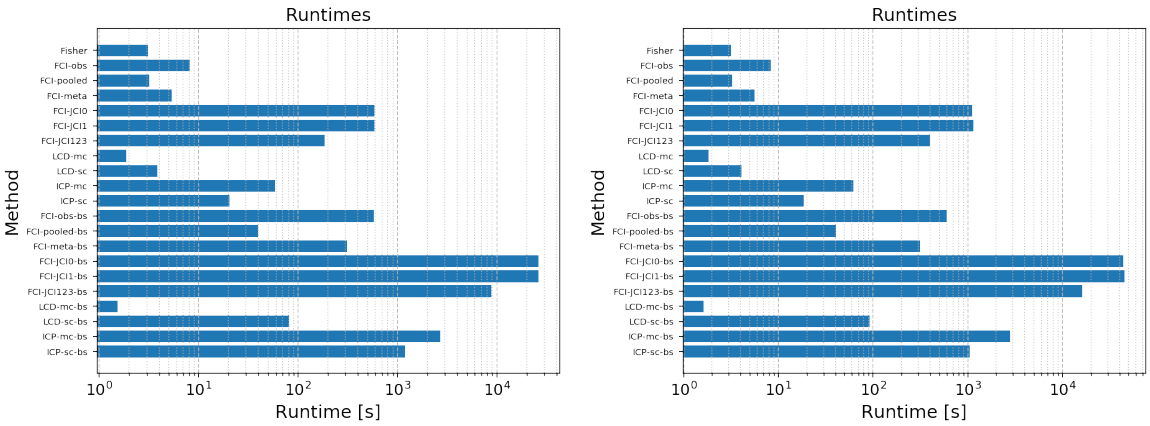


Figure 33: **Runtimes for various methods** on larger models. Left: acyclic; Right: cyclic.

5.5.4. Computation Time

Figure 33 shows total runtimes of the methods that we ran on the larger simulated models. 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^{10} subsets of all context variables in the skeleton search phase. Nevertheless, JCI variants of FCI are still computationally feasible in this setting, even with bootstrapping.

5.6. Results: Large Simulated Models

We now present results for large simulated models, with $p = 100$ system variables and $q = 10$ context variables. We only consider causal mechanism changes with unknown targets and only the acyclic setting. We used $\epsilon = \eta = 0.02$, which yields rather sparse graphs, in order to avoid that the computations would take too long (the meaning of the simulation parameters is explained in Section 5.2). We used only 100 samples per context, because the tasks would become too easy otherwise due to the sparsity of the graphs. We again used the standard partial correlation test for FCI variants in this setting instead of the default conditional independence test described in Section 5.1.7 for computational efficiency reasons.[29]

29. It is still possible to use the standard test, and the results are slightly better, but the small gain in precision does not seem to justify the large increase in computation time. Implementing `FCI-JCI123r` as proposed in Section 4.2.5 would probably yield a significant reduction in computation time. Also, alternatives for the skeleton search phase such as FCI+ (Claassen et al., 2013) could be employed to gain further speedups. Last but not least, a more efficient implementation of the standard conditional independence test would help considerably.

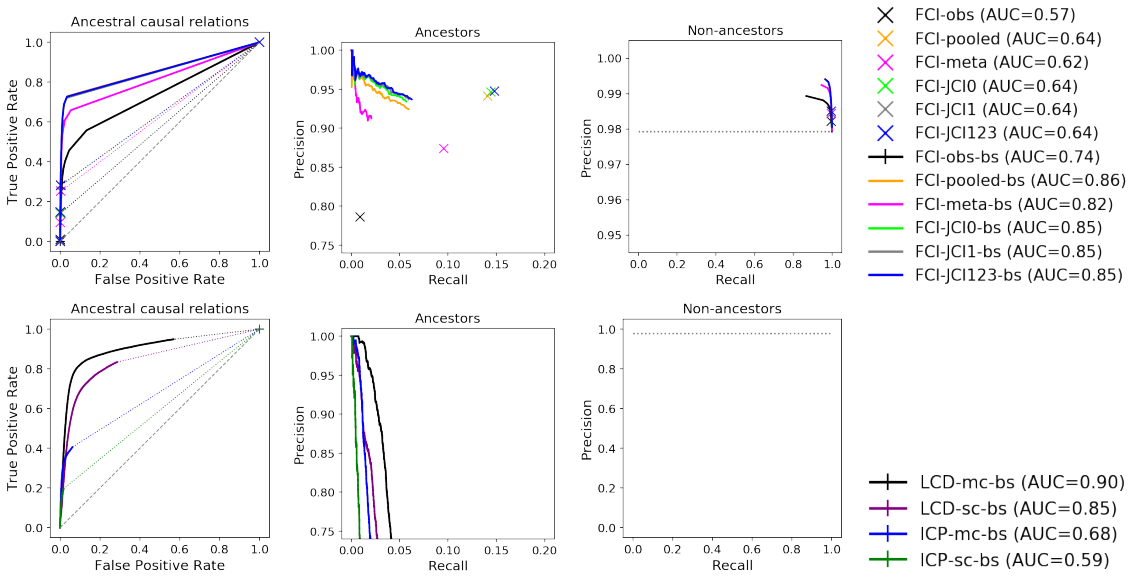


Figure 34: **Discovering ancestral causal relations between system variables** in large models. Top: FCI variants; Bottom: Bootstrapped LCD and ICP variants. Note that we zoomed in on the PR curves.

5.6.1. Ancestral Causal Relations between System Variables

Figure 34 shows the accuracy for the task of discovering ancestral causal relations between system variables for various methods and baselines. Overall, we see that JCI variants outperform non-JCI baselines. On a detailed level, the conclusions are somewhat different than what we saw for smaller models.

We start by discussing the results for the task of predicting the presence of causal relations. Like before, we see that bootstrapping FCI helps considerably to increase the precision of the predictions for the lower recall range. On the other hand, it reduces recall, possibly because only half of the available data is used and the independence threshold was not adjusted. As before, `FCI-obs` (which uses only observational data) performs worst. However, `FCI-obs` outperforms random guessing by a large margin in this setting. This is especially noteworthy given that it is only using 100 observational samples. Interestingly, the main improvement in this setting is obtained by pooling the data; whether one includes the context variables (`FCI-JCI0`) or not (`FCI-pooled`) does not seem to make much of a difference. Also, using more JCI background knowledge yields only small improvements: the differences between `FCI-JCI0`, `FCI-JCI1` and `FCI-JCI123` are small.

We observe that `LCD-mc-bs` obtains a higher precision at low recall than the FCI-JCI variants. On the other hand, the FCI-JCI variants maintain a decent precision over a larger recall range, contrary to the LCD and ICP variants that only look for very specific patterns, which may explain that their precision drops off at lower recall than it does for FCI-JCI. The "multiple context" ("`-mc`") versions of LCD and ICP outperform the versions that use

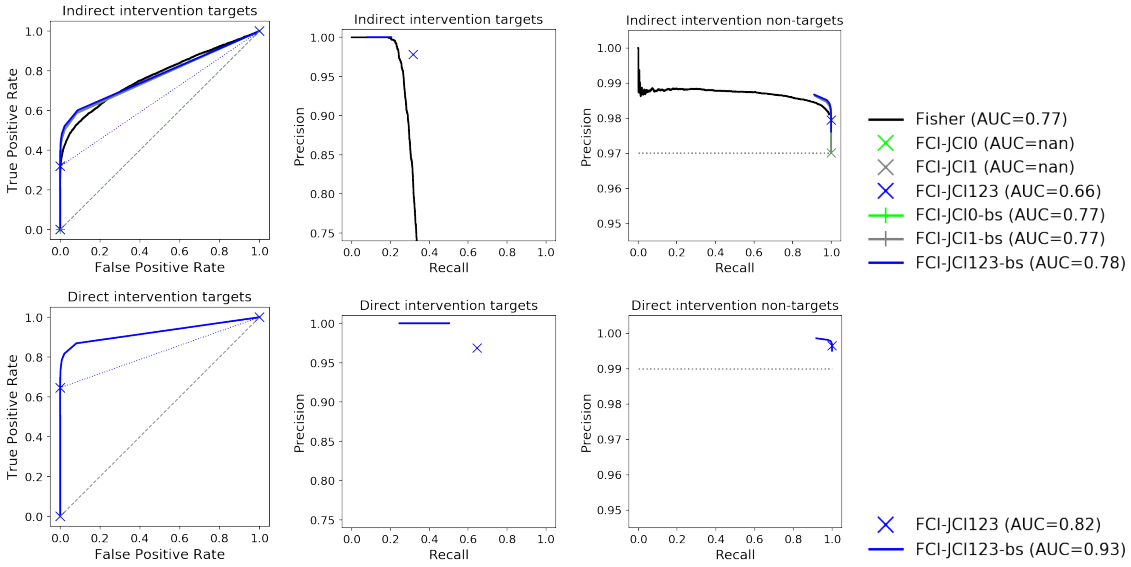


Figure 35: **Results for discovering intervention targets** in large models. Top: indirect intervention targets; Bottom: direct intervention targets. Note that we zoomed in on the PR curves.

only a single (merged) context ("`-sc`") in these settings. Interestingly, both variants of LCD outperform the corresponding variant of the more complicated ICP algorithm.

For predicting the absence of causal relations, the random guessing baseline already obtains a high precision, because of the sparsity of the graphs. FCI variants improve on this, roughly halving the error for the most confident predictions when using the bootstrapped versions. FCI-JCI variants again obtain the highest precision on this task, but don't significantly outperform `FCI-pooled`.

5.6.2. Discovering Intervention Targets

Figure 35 shows the performance of FCI-JCI variants (and as a baseline, Fisher's test) on the task of discovering intervention targets. For discovering indirect intervention targets, Fisher's test is now slightly outperformed by `FCI-JCI123-bs`. Interestingly, JCI Assumption 2 seems necessary to obtain any results on that particular task. Investigating the PAGs shows that the edges between context and system variables are mostly bidirected for `FCI-JCI1` and `FCI-JCI0`, which explains why these two algorithms yield no predictions at all. We do not have a good explanation for this behavior, but speculate that the sparse setting with many nodes makes certain empirical violations of faithfulness quite likely. One of the features of `FCI-JCI123` is that it can discover *direct* intervention targets, something that Fisher's test cannot. We find that `FCI-JCI123` identifies with high precision direct intervention targets as well as non-targets in this simulation setting.

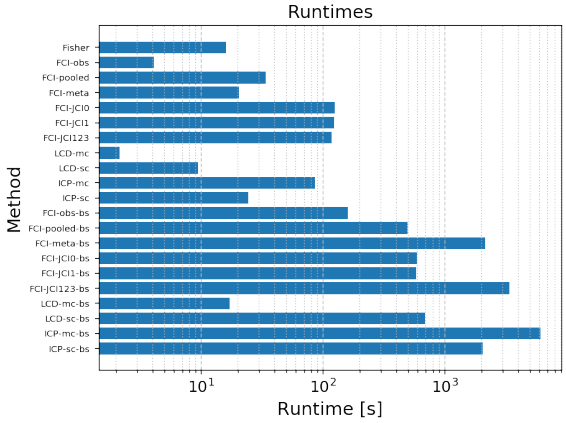


Figure 36: **Runtimes for various methods** on large models.

5.6.3. Computation Time

Figure 36 shows total runtimes of the methods that we ran on the large simulated models. For most methods, the largest part of the total running time is spent on performing independence tests. In particular, the tests that subdivide data according to context are relatively slow since we have not seriously optimized their implementation.

5.7. Summary of Results on Simulated Data

We have seen in our experiments with simulated data that JCI methods typically outperform non-JCI methods, in some settings by a large margin. For certain tasks, our newly proposed FCI-JCI algorithms provide the new state-of-the-art. Interestingly, Fisher's baseline turned out to be hard to beat on the task of discovering indirect intervention targets. However, there are other tasks for which Fisher's baseline cannot be applied but for which our newly proposed methods do apply, such as the task of discovering direct intervention targets.

As expected, LCD, ICP and ASD variants work in both the acyclic and cyclic setting. While FCI variants were expected to work only in the acyclic setting, we were surprised by how well they perform in the cyclic setting.[30] Often, but not always, adding more context variables leads to better results. Having multiple contexts was seen to be mostly beneficial if each context variable targets only a small subset of system variables, and then only for methods that can explicitly take into account multiple context variables. An interesting exception to this are LCD and ICP, which due to their sensitivity to very specific patterns actually degrade in performance when too many system variables become directly targeted by context variables. Exploiting more JCI background knowledge typically led to better results, but it depends on the task and simulation setting how large the benefits are. Interestingly, the largest boost in accuracy for discovery of causal relations between system variables comes already from JCI Assumption 0 (i.e., from pooling the data and adding the context variables).

30. These empirical observations led us to conjecture that FCI does not need to be adapted for the σ-separation setting. It was shown very recently that this is indeed the case: FCI is also sound and complete in that setting (Mooij and Claassen, 2020).

As to the relative merits of the various JCI methods that we compared, it is difficult to state this concisely. Different methods behave differently on different tasks in different simulation settings, both in terms of precision and recall as well as in terms of computation time, and for many methods there is a combination of a task and a simulation setting in which they do relatively well. Generally speaking, LCD and ICP behave quite similarly, are relatively fast and obtain high precision but have low recall; ASD-JCI variants are among the most accurate and have highest recall, but computation time explodes for more than a handful of variables; performance of FCI-JCI methods turns out to be somewhere in between, both in terms of accuracy and in terms of computation time. When implemented properly, their scalability is comparable to that of standard FCI, where the total number of variables $|\mathcal{I}| + |\mathcal{K}|$ and the sparsity of the underlying MAGs are important factors in the computation time.

Some aspects that should be kept in mind when interpreting these results is that all simulations have been done under JCI Assumptions 1, 2, 3, and there was no model misspecification. From that perspective, it was to be expected that JCI methods would work well. Also, we make no claims as to how our conclusions would generalize to different settings, for example, with non-Gaussian noise distributions, discrete variables, or continuous variables with non-linear interactions. For those settings, the choice of the conditional independence tests could have a large influence on the results, for example. A detailed study of that is beyond the scope of this paper.

5.8. Results: Real-world (Flow Cytometry) Data

In this subsection, we present an application of the Joint Causal Inference framework on real-world data: the flow cytometry data of Sachs et al. (2005). The data consists of a collection of data sets, where each data set corresponds with a different experimental condition in which a system was perturbed and subsequently measured. The system consists of an individual cell, drawn randomly from a collection of primary human immune system cells. The system variables measure the abundances of several phosphorylated protein and phospholipid components in an individual cell using a measurement technology known as flow cytometry. Performing the measurement destroys the cell, and hence, it is not possible to obtain multiple measurements over time from the same cell. Instead, snap-shot measurements of thousands of individual cells are available, obtained in different experimental conditions in which the cells were perturbed with molecular interventions, performed by administering certain reagents to the cells. Most of these interventions are not perfect, but rather change the activity of some component by an unknown amount. There is prior knowledge about the targets of these interventions (see Table 3), but it is not clear whether interventions are as specific as claimed. Many existing causal discovery approaches assume that the true causal graph is acyclic and that the system variables are causally sufficient. However, it is known that these cellular signaling networks contain strong feedback loops, and it is quite likely that some of the variables may be subject to latent confounding. Thus, this type of experimental data constitutes a compelling motivation for the Joint Causal Inference framework.

Over the years, this particular flow cytometry data set has become a "benchmark" in causal discovery (see e.g. Ramsey and Andrews (2018) for some references). Many causal

C_γ	C_δ	C_ϵ	C_ζ	C_η	$(C_\alpha, C_\theta, C_\iota)$	$N_{\boldsymbol{C}}$	Reagents added
0	0	0	0	0	(1,0,0)	853	α-CD3, α-CD28
1	0	0	0	0	(1,0,0)	911	α-CD3, α-CD28, AKT inhibitor
0	1	0	0	0	(1,0,0)	723	α-CD3, α-CD28, G0076
0	0	1	0	0	(1,0,0)	810	α-CD3, α-CD28, Psitectorigenin
0	0	0	1	0	(1,0,0)	799	α-CD3, α-CD28, U0126
0	0	0	0	1	(1,0,0)	848	α-CD3, α-CD28, LY294002
0	0	0	0	0	(0,1,0)	913	PMA
0	0	0	0	0	(0,0,1)	707	β2CAMP

Table 6: Experimental design for part of the Sachs et al. (2005) flow cytometry data used in our experiments.

discovery methods have been applied to this data, and in many cases, the "consensus network" in Sachs et al. (2005), visualized in Figure 38(a), was used as a ground truth to evaluate the results of the causal discovery procedure. However, we would like to point out that there are good reasons to be skeptical about the assumption that the "consensus network" represents the true causal graph of the system (as acknowledged by several domain experts we spoke). Indeed, by inspecting the data one can find many examples where the data is incompatible with the hypothesis that the "consensus network" is a realistic and complete description of the underlying system. For example, according to the "consensus network", an intervention that inhibits the activity of Mek should have no effect on Raf (because Raf is a direct cause of Mek and there is no feedback loop from Mek to Raf). However, in the data we see an increase of more than one order of magnitude in the abundance of Raf when U0126 (a reagent assumed to inhibit Mek activity) is added to the cells (see Figure 37). So either U0126 also directly targets Raf, or Mek must be a cause of Raf, in both cases contradicting the "consensus network". In the literature regarding this signaling pathway (which has been studied in great detail since it plays an important role in many human cancers) it is often suggested that there should be a feedback loop back from Erk to Raf (whose molecular mechanism is still unknown). This would be in line with our observations from the data in Figure 37, and also imply that the "consensus network" is incomplete. This is just one example illustrating that the data is not entirely compatible with the "consensus network", and it is easy to find more of these examples via visual inspection of the data. Therefore, we will *not* make use of the "consensus network" as a ground truth to compare with when evaluating the output of the causal discovery algorithms.

Sachs et al. (2005) use an MCMC method to estimate the structure of a causal Bayesian network from the combined observational and interventional data, making use of the modified BDe score proposed by Cooper and Yoo (1999). Eaton and Murphy (2007) later used a dynamic programming algorithm to solve the estimation problem exactly. Like the original analysis by Sachs et al. (2005), many causal discovery methods that have been applied on this data rely on the background knowledge about the intervention types and targets, for which we provide (our interpretation) in Table 3. A notable exception is Eaton and Murphy (2007), who were the first to estimate the intervention targets directly from the data. Exploiting the background knowledge on intervention targets and types simplifies the causal discovery problem considerably. However, the accuracy of this background knowl-

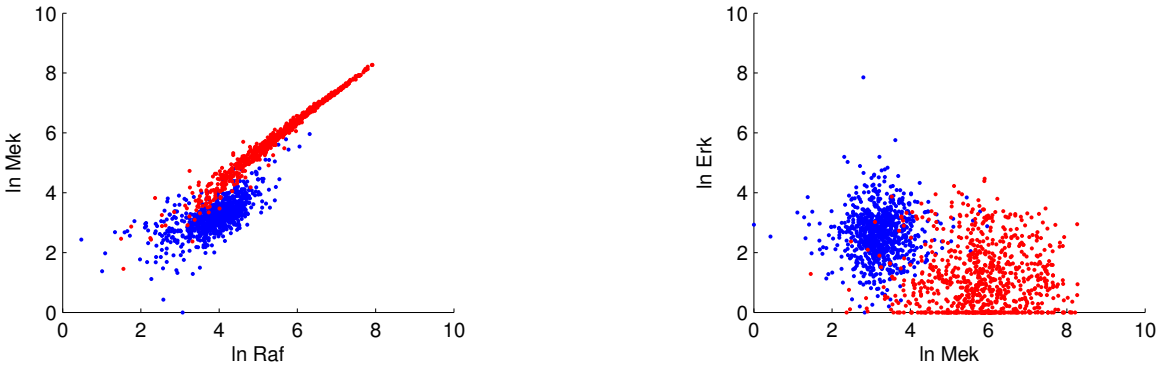


Figure 37: Log-abundances of Mek vs. Raf (left) and of Erk vs. Mek (right). Blue: observational baseline ($C_\alpha = 1, \boldsymbol{C}_{\setminus\alpha} = \mathbf{0}$); Red: reagent U0126 added ($C_\zeta = 1$). We observe: (i) the measurement noise is quite small; (ii) Raf and Mek are highly correlated ("consensus network": Raf is a direct cause of Mek); (iii) strong evidence for feedback (intervening on Mek changes Raf abundance) if we assume that U0126 directly targets Mek but not Raf; (iv) the Mek *inhibitor* U0126 *increases* Mek abundance (so modeling this as a perfect intervention would not be realistic); (v) Mek and Erk are independent in both contexts (even though Mek is a direct cause of Erk according to the "consensus network"), an apparent violation of faithfulness.

edge is not universally accepted. In particular, many biologists that we spoke with were skeptical about the assumed specificity of the interventions (i.e., the interventions may have additional direct effects that are not listed in the table).

We ran various FCI-JCI variants on a subset of the flow cytometry data.[31] The experimental design of the original data is described in Table 2 (left), p. 34. In order to avoid deterministic relations between context variables, we merged context variables C_α (α-CD3/CD28), C_θ (PMA) and C_ι (β2CAMP), as discussed in Section 4.1, leading to the experimental design in Table 2 (right). This means that we must interpret the merged context variable as referring to the addition of PMA or β2CAMP, combined with the *omission* of α-CD3/CD28. However, note that there are still approximate conditional independences in this experimental design of the form $C_\beta \perp\!\!\!\perp C_k \,|\, (C_\alpha, C_\theta, C_\iota)$ for $k \in \{\gamma, \delta, \epsilon, \zeta, \eta\}$. This could lead to problems with JCI Assumption 3. For that reason, but also in order 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.

Similarly to Eaton and Murphy (2007), we do not use the background knowledge regarding intervention types or targets. 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

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

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 applies. JCI Assumption 3 then also applies since there are no conditional independences in the context distribution (after merging C_α, C_θ and C_ι and leaving out all contexts with $C_\beta = 1$). When using `FCI-JCI123`, we can then learn the intervention targets from the data itself, without making use of the background knowledge on intervention types and targets.

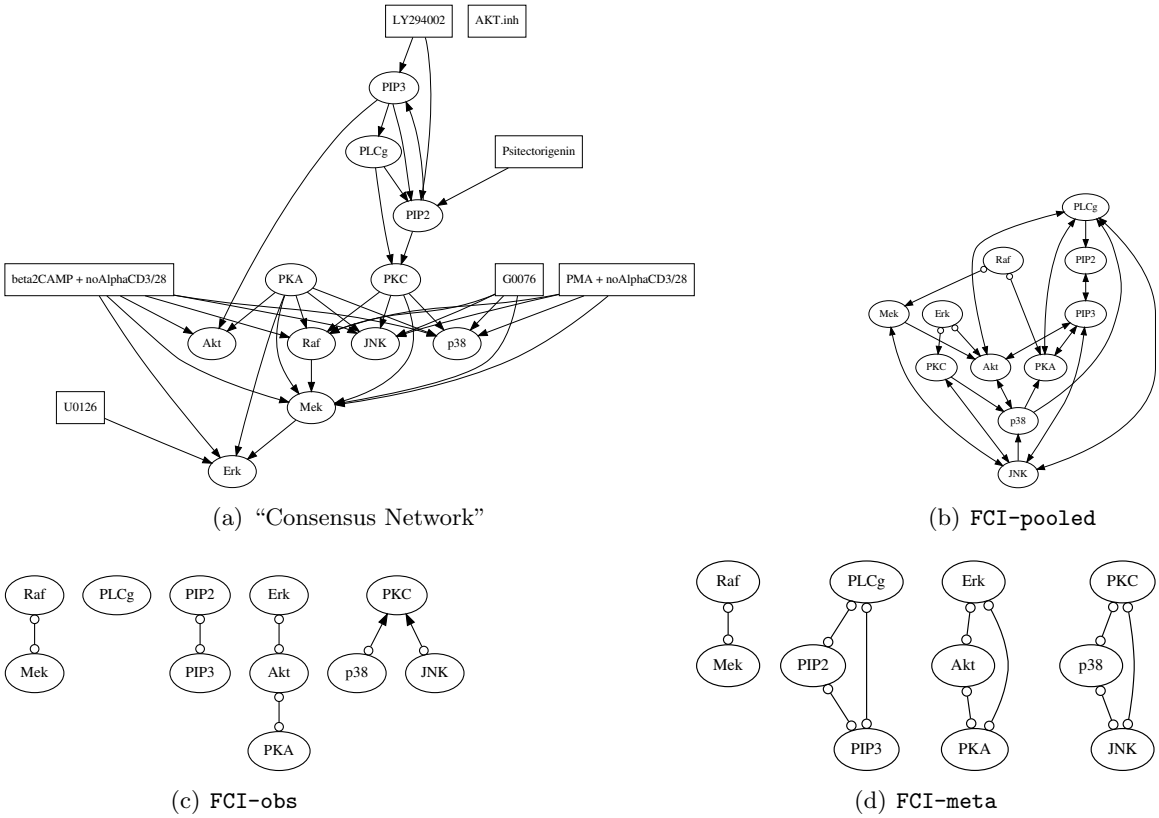
For comparison, we also ran `FCI-obs`, i.e., standard FCI using only the observational data set (i.e., the one in which only global activators α-CD3 and α-CD28 have been administered). We also ran `FCI-meta`, which uses Fisher's method to combine p-values of conditional independence tests in the 8 separate 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. Computation times of various implementations are reported in Figure 42.

The "consensus network" and the PAGs obtained by the FCI baselines are shown in Figure 38. Figure 39 shows PAGs obtained by the FCI-JCI variants. Note that we show the PAGs obtained without bootstrapping, although these are not necessarily stable. Therefore, we show in Figure 40 also the bootstrapped results for the learned ancestral causal relations between system variables, and the learned intervention targets. One can see here that most, but not all, of these features are stably predicted by the FCI-JCI variants. In particular, the causal relations[32] Mek$\dashrightarrow$Raf, PLCg$\dashrightarrow$PIP2, Akt$\dashrightarrow$Erk, P38$\dashrightarrow$PKC are predicted by both `FCI-JCI123` and `FCI-JCI1` with high confidence.

Regarding the learned indirect intervention targets, `FCI-JCI123` has an advantage over `FCI-JCI1` because it can exclude bidirected edges between context and system variables. Nonetheless, `FCI-JCI1` predicts (in accordance with `FCI-JCI123`) that G0076 affects PLCg, PIP2, PKC and P38. For discovering indirect intervention targets, `Fisher`'s test for causality is simple and powerful. However, it is not able to learn the *direct* intervention targets, like `FCI-JCI123` can. Although the direct intervention targets learned by `FCI-JCI123` do not all correspond with the "consensus network", they do agree for example on Psitectorigenin directly targeting PIP2. Most direct intervention targets learned by `FCI-JCI123` were also found by Eaton and Murphy (2007). It is interesting to see that `FCI-JCI123` considers both Mek and Erk as direct intervention targets of U0126, while Raf is identified as an indirect target.

In the absence of a reliable ground truth, 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. In that case, 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 appears

32. We write $i \dashrightarrow j$ if i is a cause of j.

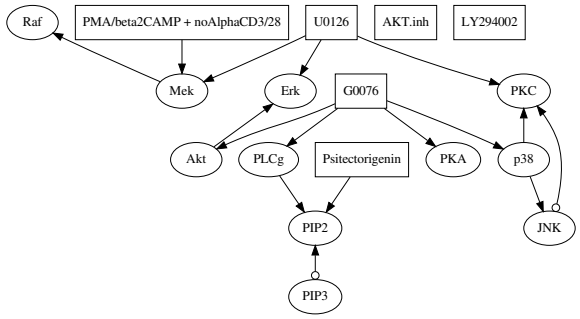


(a) "Consensus Network" (b) `FCI-pooled`

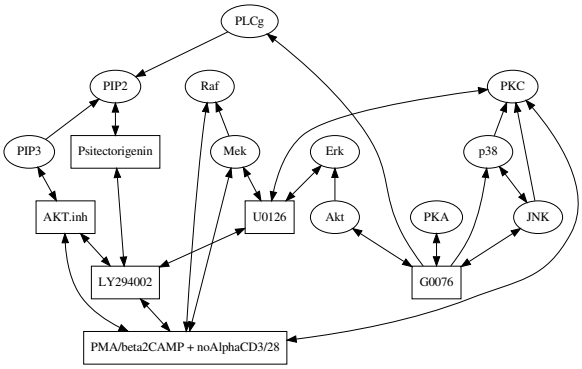
(c) `FCI-obs` (d) `FCI-meta`

Figure 38: **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 ellipses. 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") data set in which only global activators α-CD3 and α-CD28 have been administered; (d) FCI with as input the result of Fisher's method for combining conditional independence test results from all data sets.

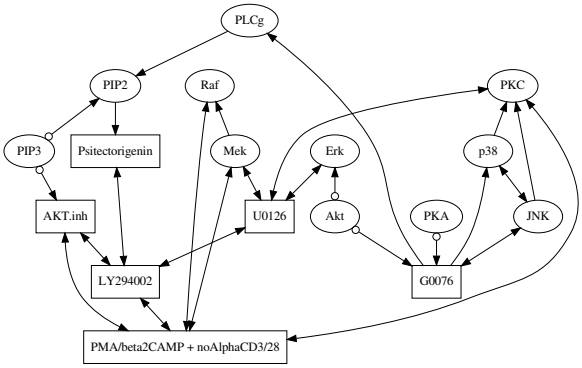
to be a considerable improvement over the simple FCI baselines `FCI-obs`, `FCI-meta` and `FCI-pooled`. Remarkably, `FCI-JCI123` manages to orient most of the edges. The output also resembles the consensus network, although some of the edges seem to be reversed. Considering that we have not taken into account the available background knowledge on the intervention types and targets (Table 3), and that we have not used any tuning of the parameters we consider this still an impressive and encouraging result that illustrates the potential that JCI has for analyzing complex scientific data sets.



(a) `FCI-JCI123`



(b) `FCI-JCI1`



(c) `FCI-JCI0`

Figure 39: **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 ellipses. 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 40 shows also the corresponding bootstrap estimates for certain features of the PAGs.

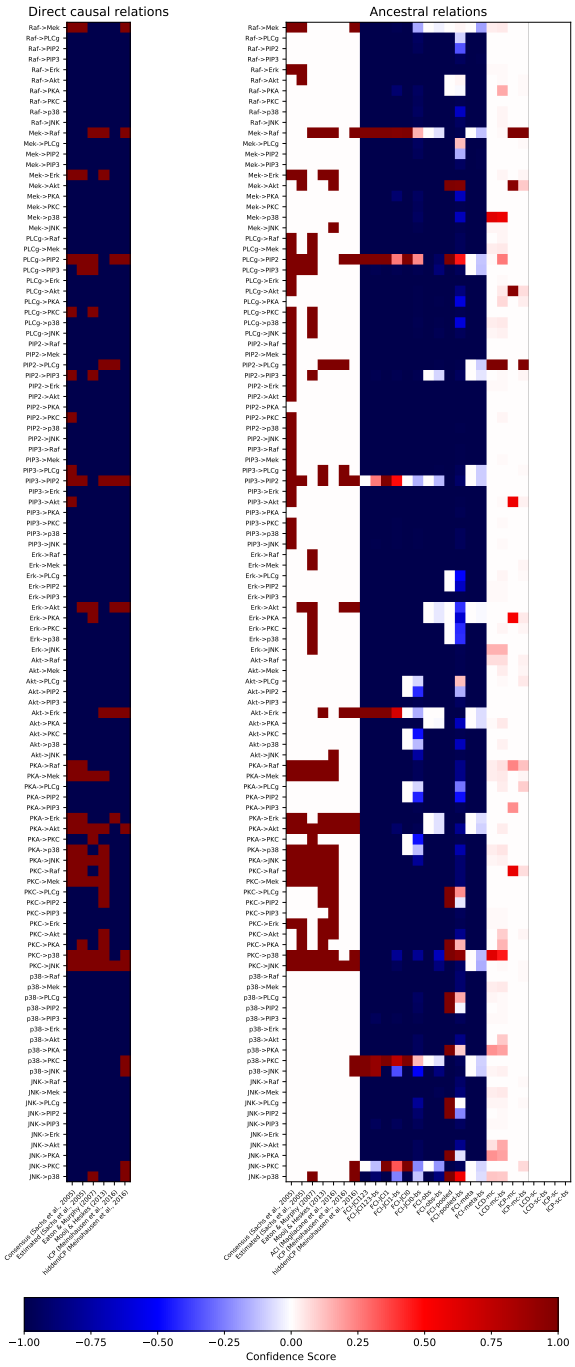


Figure 40: **Causal relationships between the biochemical agents in the flow cytometry data of Sachs et al. (2005)**, according to different causal discovery methods and the "consensus network" according to Sachs et al. (2005) (which we do not consider as a reliable complete ground truth). We also included results of causal discovery methods reported in other works.

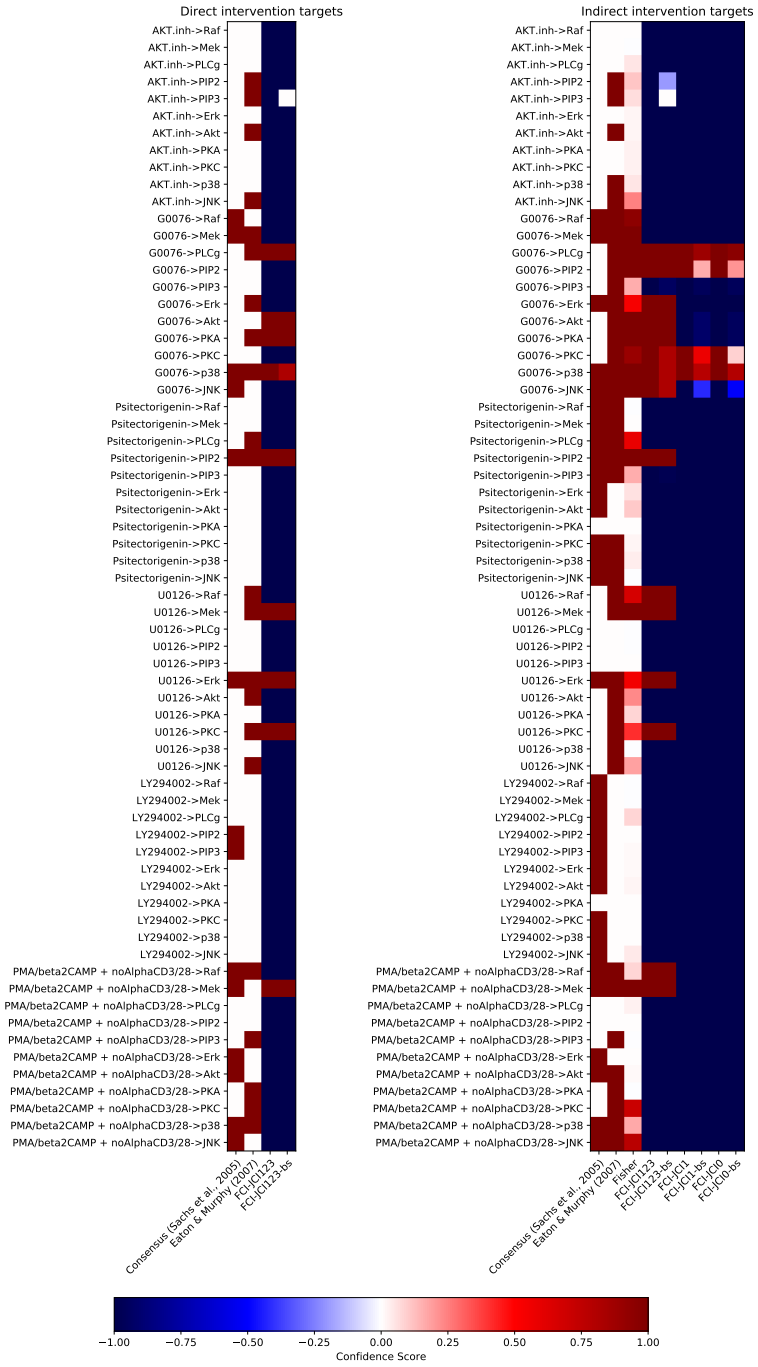


Figure 41: **Intervention effects on biochemical agents in the flow cytometry data of Sachs et al. (2005)**, according to different causal discovery methods and the "consensus network" according to Sachs et al. (2005) (which we do not consider as a reliable complete ground truth). We also included results of causal discovery methods reported in other works.

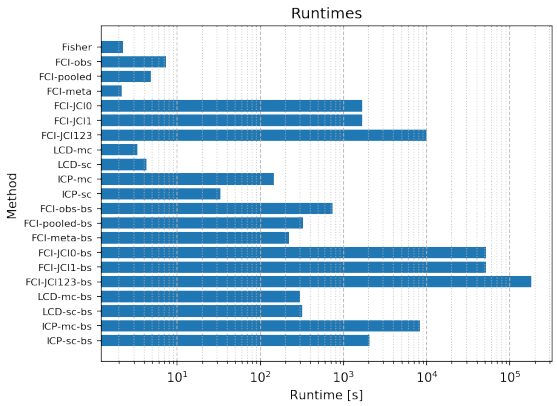


Figure 42: **Runtimes for various methods on the flow cytometry data of Sachs et al. (2005).**

6. Conclusions and Discussion

In this work, we proposed Joint Causal Inference (JCI), a powerful and elegant framework for causal discovery from data sets from multiple contexts. JCI generalizes the ideas of 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 data sets 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 (with context variables included), 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 by 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.

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 method of Hyttinen et al. (2014) combined with ideas from Magliacane et al. (2016b), which is very flexible and accurate. By replacing d-separation with σ-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 about 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 on small models, 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 interventions with known targets. The reason is that none of the JCI implementations exploited the perfect 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 perfect 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 discovering indirect (ancestral) causal relations, the classic (and very simple and fast) LCD algorithm can be competitive with 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 locally. 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 data sets 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 and editor 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 supported by EU-FP7 grant agreement n.603016 (MATRICS).

Appendix A. Proofs

In this appendix we provide the proofs that were omitted from the main text.

A.1. JCI Foundations

Theorem 13 *Assume that JCI Assumptions 0, 1 and 2 hold for SCM $\mathcal{M}$:*

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

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

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

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

(i) the conditional system graphs coincide: $\mathcal{G}(\mathcal{M})_{\mathrm{do}(\mathcal{K})} = \mathcal{G}(\tilde{\mathcal{M}})_{\mathrm{do}(\mathcal{K})}$;

(ii) if $\tilde{\mathcal{M}}$ and $\mathcal{M}$ induce the same context distribution, i.e., $\mathbb{P}_{\mathcal{M}}(\boldsymbol{C}) = \mathbb{P}_{\tilde{\mathcal{M}}}(\boldsymbol{C})$, then for any perfect intervention on the system variables $\mathrm{do}(I, \boldsymbol{\xi}_I)$ with $I \subseteq \mathcal{I}$ (including the non-intervention $I = \emptyset$), $\tilde{\mathcal{M}}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}$ is observationally equivalent to $\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}$.

(iii) if the context graphs $\mathcal{G}(\tilde{\mathcal{M}})_{\mathcal{K}}$ 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 σ-separations).

Proof Let $\mathcal{M}$ be an SCM of the form (5). Under JCI Assumption 1, the structural equations for the context variables do not depend on the system variables:

$$C_k = f_k(\boldsymbol{C}_{\mathrm{PA}_{\mathcal{H}}(k)\cap\mathcal{K}}, \boldsymbol{E}_{\mathrm{PA}_{\mathcal{H}}(k)\cap\mathcal{J}}), \qquad k \in \mathcal{K}.$$

Because of JCI Assumption 2, $\mathrm{PA}_{\mathcal{H}}(\mathcal{K}) \cap \mathrm{PA}_{\mathcal{H}}(\mathcal{I}) \cap \mathcal{J} = \emptyset$, 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 $k \to i$ 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 $\boldsymbol{C}$ is independent of $\boldsymbol{E}_{\mathrm{PA}_{\mathcal{H}}(\mathcal{I})}$ (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 $\mathbb{P}_{\mathcal{M}}(\boldsymbol{C}) = \mathbb{P}_{\tilde{\mathcal{M}}}(\boldsymbol{C})$.

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 "σ-open" or as "d-open". Note that by assumption, $\mathcal{G}(\mathcal{M})_{\mathrm{do}(\mathcal{K})} = \mathcal{G}(\tilde{\mathcal{M}})_{\mathrm{do}(\mathcal{K})}$, and that the edges in $\mathcal{G}(\mathcal{M})_{\mathrm{do}(\mathcal{K})}$ are a subset of those in $\mathcal{G}(\mathcal{M})$, and of those in $\mathcal{G}(\tilde{\mathcal{M}})$. 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 π in $\mathcal{G}(\mathcal{M})_{\mathcal{K}}$ such that π is A-open in $\mathcal{G}(\mathcal{M})$ for some $A \subseteq \mathcal{I} \cup \mathcal{K}$, we can find a path π' 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 $B \subseteq \mathcal{K} \cap \mathrm{AN}_{\mathcal{G}(\mathcal{M})_{\mathrm{do}(\mathcal{K})}}(A \setminus \mathcal{K})$. For π to be A-open in $\mathcal{G}(\mathcal{M})$, any collider on π 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 $\mathcal{K} \cap \mathrm{AN}_{\mathcal{G}(\mathcal{M})_{\mathrm{do}(\mathcal{K})}}(A \setminus \mathcal{K})$ (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 π, but that we can construct another path π' for which this can be done.

Consider a path π 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 $B \subseteq \mathcal{K} \cap \mathrm{AN}_{\mathcal{G}(\mathcal{M})_{\mathrm{do}(\mathcal{K})}}(A \setminus \mathcal{K})$. We will proceed by induction. Suppose a walk π_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 $B_m = (\mathcal{K} \cap \mathrm{AN}_{\mathcal{G}(\mathcal{M})_{\mathrm{do}(\mathcal{K})}}(A \setminus \mathcal{K})) \setminus \mathrm{NCOL}(\pi_m)$. Consider the "problematic" colliders $\mathrm{COL}(\pi_m) \setminus \mathrm{AN}_{\mathcal{G}(\mathcal{M})}(A \cap \mathcal{K} \cup B_m)$ on π_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 π_m. Since c is not $\mathcal{G}(\mathcal{M})$-ancestor of $A \cap \mathcal{K} \cup B_m$, but π_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 π_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 π_m between c and k_c

by a directed path $c \to \cdots \to k_c$ 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 π_{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 π_{m+1} is at least one less than on π_m: c is no longer a collider, and if k_c became a collider on π_{m+1}, it won't be problematic (as it is itself in $(\mathcal{K} \cap \text{AN}_{\mathcal{G}(\mathcal{M})_{\text{do}(\mathcal{K})}}(A \setminus \mathcal{K}))$ and cannot also occur as non-collider on π_{m+1}), We repeat this procedure until no problematic colliders are present anymore. This yields a walk π_M that is both A-open and A'-open, with $A' = (A \cap \mathcal{K}) \cup B$ where $B = B_M = (\mathcal{K} \cap \text{AN}_{\mathcal{G}(\mathcal{M})_{\text{do}(\mathcal{K})}}(A \setminus \mathcal{K})) \setminus \text{NCOL}(\pi_M)$. We now shorten this A'-open walk π_M in $\mathcal{G}(\mathcal{M})_{\mathcal{K}}$ into an A'-open path π' 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 $k \to i$, 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 σ-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 σ-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. ■

Corollary 14 *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 conditional system graphs coincide: $\mathcal{G}(\mathcal{M})_{\text{do}(\mathcal{K})} = \mathcal{G}(\tilde{\mathcal{M}})_{\text{do}(\mathcal{K})}$;

(ii) for any perfect intervention on the system variables $\text{do}(I, \boldsymbol{\xi}_I)$ with $I \subseteq \mathcal{I}$ (including the non-intervention $I = \emptyset$), $\tilde{\mathcal{M}}_{\text{do}(I,\boldsymbol{\xi}_I)}$ is observationally equivalent to $\mathcal{M}_{\text{do}(I,\boldsymbol{\xi}_I)}$;

(iii) if the context distribution $\mathbb{P}_{\mathcal{M}}(\boldsymbol{C})$ contains no conditional or marginal independences, then the same σ-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 (5). Under JCI Assumption 1, the structural equations for the context variables do not depend on the system variables:

$$C_k = f_k(\boldsymbol{C}_{\mathrm{PA}_{\mathcal{H}}(k)\cap\mathcal{K}}, \boldsymbol{E}_{\mathrm{PA}_{\mathcal{H}}(k)\cap\mathcal{J}}), \qquad k \in \mathcal{K}.$$

Because of JCI Assumption 2, $\mathrm{PA}_{\mathcal{H}}(\mathcal{K}) \cap \mathrm{PA}_{\mathcal{H}}(\mathcal{I}) \cap \mathcal{J} = \emptyset$, 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:

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

where $\tilde{\mathcal{J}} = \mathcal{J} \cup \{C\}$ contains an additional exogenous variable $\boldsymbol{E}_C \in \prod_{k\in\mathcal{K}} \mathcal{E}_k$ with components $(\boldsymbol{E}_C)_k \in \mathcal{C}_k$ with distribution $\mathbb{P}(\boldsymbol{E}_C) = \mathbb{P}_{\mathcal{M}}(\boldsymbol{C})$ and g_k the projection on the k^{th} component $g_k : \boldsymbol{E}_C \mapsto (\boldsymbol{E}_C)_k$. By construction, this SCM $\tilde{\mathcal{M}}$ satisfies JCI Assumptions 1 and 2. The only aspect that requires some work is to prove that $\tilde{\mathcal{M}}$ as constructed above is simple (Definition 4).

Take $\mathcal{O} \subseteq \mathcal{I}$ and consider the solution function for $\mathcal{O}$ according to $\mathcal{M}$:

$$\boldsymbol{g}_{\mathcal{O}} : \boldsymbol{\mathcal{X}}_{(\mathrm{PA}_{\mathcal{H}}(\mathcal{O})\setminus\mathcal{O})\cap\mathcal{I}} \times \boldsymbol{\mathcal{C}}_{(\mathrm{PA}_{\mathcal{H}}(\mathcal{O})\setminus\mathcal{O})\cap\mathcal{K}} \times \boldsymbol{\mathcal{E}}_{\mathrm{PA}_{\mathcal{H}}(\mathcal{O})\cap\mathcal{J}} \to \boldsymbol{\mathcal{X}}_{\mathcal{I}\setminus\mathcal{O}}.$$

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 $\boldsymbol{g}_{\mathcal{Q}} : \mathcal{E}_C \to \boldsymbol{\mathcal{C}}_Q$ with components $\mathcal{E}_C \to \mathcal{C}_k : \boldsymbol{e}_C \mapsto g_k(\boldsymbol{e}_C)$, $k \in \mathcal{Q}$. 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 $\mathbb{P}_{\tilde{\mathcal{M}}}(\boldsymbol{E}) = \mathbb{P}_{\mathcal{M}}(\boldsymbol{E})$ and satisfies JCI Assumption 3 by construction. The other statements now follow by applying Theorem 13, 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 σ-separations if the context distribution $\mathbb{P}_{\mathcal{M}}(\boldsymbol{C})$ 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 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 $\mathbb{P}_{\mathcal{M}}(\boldsymbol{C})$ has no conditional independences, there must be a K-σ-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 $\mathcal{G}(\mathcal{M})_{\mathcal{K}}$ 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 $\{k, k'\} \cap K = \emptyset$. ■

The following Lemma and Corollary extend these fundamental results further, which enables one to state a precise relationship between our JCI approach of jointly modeling system and context with alternative approaches based on modeling the system conditional on its context (e.g., Yang et al. (2018)).

Lemma 23 *Let $\mathcal{M}$ be an SCM that satisfies JCI Assumptions 0, 1, 2. Then the same* restricted *separations hold in $\mathcal{G}(\mathcal{M})$ as in the conditional system graph $\mathcal{G}(\mathcal{M})_{\mathrm{do}(\mathcal{K})}$, i.e.,*

$$X \underset{\mathcal{G}(\mathcal{M})_{\mathrm{do}(\mathcal{K})}}{\perp} Y \,|\, Z \iff X \underset{\mathcal{G}(\mathcal{M})}{\perp} Y \,|\, Z$$

whenever $X, Y, Z \subseteq \mathcal{I} \cup \mathcal{K}$ with $X \cap \mathcal{K} = \emptyset$ and $\mathcal{K} \subseteq Y \cup Z$ (where "separations" can refer to either d-separations or σ-separations).

Proof Let $X, Y, Z \subseteq \mathcal{I} \cup \mathcal{K}$ be such that $X \cap \mathcal{K} = \emptyset$ and $\mathcal{K} \subseteq Y \cup Z$. Let $\mathcal{G}_1, \mathcal{G}_2$ be two graphs in $\{\mathcal{G}(\mathcal{M}),\, \mathcal{G}(\mathcal{M})_{\mathrm{do}(\mathcal{K})}\}$. Let π be a path in $\mathcal{G}_1$ between a node in X and a node in Y that is open in $\mathcal{G}_1$ and that does not contain any non-endpoint nodes in $X \cup Y$. It cannot have non-endpoint nodes in $\mathcal{K}$, because those would be either in Y (a contradiction), or in Z (and since they would be non-colliders with an outgoing directed edge pointing to another strongly-connected component, they would block the path, another contradiction). But then the same path π must be present in $\mathcal{G}_2$ as well. It is easy to see that it must also be open in $\mathcal{G}_2$, since for each $i \in \mathcal{I}$, $\mathrm{DE}_{\mathcal{G}_1}(i) = \mathrm{DE}_{\mathcal{G}_2}(i)$ and $\mathrm{SC}_{\mathcal{G}_1}(i) = \mathrm{SC}_{\mathcal{G}_2}(i)$. ■

We can now formulate the following slightly adapted version of Corollary 14:

Corollary 24 *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 conditional system graphs coincide: $\mathcal{G}(\mathcal{M})_{\mathrm{do}(\mathcal{K})} = \mathcal{G}(\tilde{\mathcal{M}})_{\mathrm{do}(\mathcal{K})}$;

(ii) as a consequence, the same restricted *separations hold in $\mathcal{G}(\tilde{\mathcal{M}})$ as in $\mathcal{G}(\mathcal{M})$ and in their corresponding conditional system graphs, i.e.,*

$$X \underset{\mathcal{G}(\tilde{\mathcal{M}})}{\perp} Y \,|\, Z \iff X \underset{\mathcal{G}(\tilde{\mathcal{M}})_{\mathrm{do}(\mathcal{K})}}{\perp} Y \,|\, Z \iff X \underset{\mathcal{G}(\mathcal{M})_{\mathrm{do}(\mathcal{K})}}{\perp} Y \,|\, Z \iff X \underset{\mathcal{G}(\mathcal{M})}{\perp} Y \,|\, Z$$

whenever $X, Y, Z \subseteq \mathcal{I} \cup \mathcal{K}$ with $X \cap \mathcal{K} = \emptyset$ and $\mathcal{K} \subseteq Y \cup Z$ (where "separations" can refer to either d-separations or σ-separations);

(iii) for any perfect intervention on the system variables $\mathrm{do}(I, \boldsymbol{\xi}_I)$ with $I \subseteq \mathcal{I}$ (including the non-intervention $I = \emptyset$), and any perfect intervention on all context variables $\mathrm{do}(\mathcal{K}, \boldsymbol{c})$:

$$\mathbb{P}_{\tilde{\mathcal{M}}}(\boldsymbol{X} \,|\, \mathrm{do}(\mathcal{K}, \boldsymbol{c}), \mathrm{do}(I, \boldsymbol{\xi}_I)) = \mathbb{P}_{\mathcal{M}}(\boldsymbol{X} \,|\, \mathrm{do}(\mathcal{K}, \boldsymbol{c}), \mathrm{do}(I, \boldsymbol{\xi}_I));$$

(iv) as a consequence, $\mathbb{P}_{\mathcal{M}}(\boldsymbol{X} \,|\, \boldsymbol{C}) = \mathbb{P}_{\tilde{\mathcal{M}}}(\boldsymbol{X} \,|\, \boldsymbol{C})$, and in particular, the same restricted conditional independences hold, i.e.,

$$X \underset{\mathcal{G}(\tilde{\mathcal{M}})}{\perp\!\!\!\perp} Y \,|\, Z \iff X \underset{\mathcal{G}(\tilde{\mathcal{M}})_{\mathrm{do}(\mathcal{K})}}{\perp\!\!\!\perp} Y \,|\, Z \iff X \underset{\mathcal{G}(\mathcal{M})_{\mathrm{do}(\mathcal{K})}}{\perp\!\!\!\perp} Y \,|\, Z \iff X \underset{\mathcal{G}(\mathcal{M})}{\perp\!\!\!\perp} Y \,|\, Z$$

whenever $X, Y, Z \subseteq \mathcal{I} \cup \mathcal{K}$ with $X \cap \mathcal{K} = \emptyset$ and $\mathcal{K} \subseteq Y \cup Z$;

(v) the context distribution $\mathbb{P}_{\tilde{\mathcal{M}}}(\boldsymbol{C})$ contains no conditional or marginal independences.

Proof The same SCM $\tilde{\mathcal{M}}$ as constructed in the proof of Corollary 14, but with a generic distribution of $\mathbb{P}(\boldsymbol{E}_C)$ that contains no conditional or marginal independences, is easily seen to fulfill all requirements. ■

A.2. Minimal Conditional (In)Dependencies

In this section we generalize two useful Lemmas from Claassen and Heskes (2011) to the cyclic setting.

Definition 25 *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 σ-separation ($\perp^\sigma$). We say that the* minimal separation

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

holds if and only if

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

In words: all nodes in Z are required (in the context of the nodes in S) to separate X from Y. Similarly: we say that the minimal connection

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

holds if and only if

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

In words: all nodes in Z are required (in the context of the nodes in S) to connect X with Y.

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 26 *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 σ-separation ($\perp^\sigma$), we have that:*

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

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 σ-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 27 *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 σ-separation ($\perp^\sigma$), we have that:*

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

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 σ-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 end node. 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 ω represents $\{X, Y\} \cup S$) and add an edge $Z_i \to \omega$ 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 $Z_i \to Z_j$ 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 implication of this is that the intersection of all sets that separate a node X from a node Y can only consist of ancestors of X or Y:

Proposition 28 *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 σ-separation ($\perp^\sigma$), consider $Z^* := \bigcap \{Z \subseteq \mathcal{V} : X \notin Z, Y \notin Z, X \perp_{\mathcal{G}} Y \mid Z\}$. Then $Z^* \subseteq \text{AN}_{\mathcal{G}}(\{X, Y\})$.*

Proof First, note that $Z^* = \bigcap \{Z \subseteq \mathcal{V} : X \notin Z, Y \notin Z, X \perp_{\mathcal{G}} Y \mid [Z]\}$. From Lemma 27, $X \perp_{\mathcal{G}} Y \mid [Z]$ implies $Z \subseteq \text{AN}_{\mathcal{G}}(\{X, Y\})$. Hence $Z^* \subseteq \text{AN}_{\mathcal{G}}(\{X, Y\})$. ■

Appendix B. Soundness, Consistency and Completeness Properties of FCI-JCI

In this appendix we will formulate and prove various results concerning soundness and completeness of FCI-JCI variants.

B.1. Preliminaries on MAGs and PAGs

We start by summarizing the basic definitions and results from the theory of maximal ancestral graphs and partial ancestral graphs (Spirtes et al., 2000; Richardson and Spirtes, 2002; Zhang, 2006, 2008a,b) that we will need.

B.1.1. Directed Maximal Ancestral Graphs

Richardson and Spirtes (2002) introduced a class of graphs known as *maximal ancestral graphs (MAGs)*. The general formulation of MAGs allows for undirected edges which are useful when modeling selection bias, but here we will only use *directed* maximal ancestral graphs without undirected edges (sometimes abbreviated as DMAGs in the literature) as we assume for simplicity that there is no selection bias. In order to define a directed maximal ancestral graph, we need the notion of inducing path.

Definition 29 *Let $\mathcal{G} = \langle \mathcal{V}, \mathcal{E}, \mathcal{F} \rangle$ be an acyclic directed mixed graph (ADMG). An* inducing path between two nodes $u, v \in \mathcal{V}$ *is a path in $\mathcal{G}$ between u and v on which every node (except for the end nodes) is a collider on the path and an ancestor in $\mathcal{G}$ of an end node of the path.*

We can now state:

Definition 30 *A directed mixed graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E}, \mathcal{F} \rangle$ is called a* directed maximal ancestral graph (DMAG) *if all of the following conditions hold:*

1. *Between any two different nodes there is at most one edge, and there are no self-cycles;*
2. *The graph contains no directed or almost directed cycles ("ancestral");*
3. *There is no inducing path between any two non-adjacent nodes ("maximal").*

Given an ADMG, we can define a corresponding DMAG (Richardson and Spirtes, 2002):

Definition 31 *Let $\mathcal{G} = \langle \mathcal{V}, \mathcal{E}, \mathcal{F} \rangle$ be an ADMG. The* directed maximal ancestral graph induced by $\mathcal{G}$ *is denoted* $\text{DMAG}(\mathcal{G})$ *and is defined as* $\text{DMAG}(\mathcal{G}) = \langle \tilde{\mathcal{V}}, \tilde{\mathcal{E}}, \tilde{\mathcal{F}} \rangle$ *such that $\tilde{\mathcal{V}} = \mathcal{V}$ and for each pair $u, v \in \mathcal{V}$ with $u \neq v$, there is an edge in* $\text{DMAG}(\mathcal{G})$ *between u and v if and only if there is an inducing path between u and v in $\mathcal{G}$, and in that case the edge*

in $\mathrm{DMAG}(\mathcal{G})$ *connecting* u *and* v *is:*

$$\begin{cases} u \to v & \text{if } u \in \mathrm{AN}_{\mathcal{G}}(v), \\ u \leftarrow v & \text{if } v \in \mathrm{AN}_{\mathcal{G}}(u), \\ u \leftrightarrow v & \text{if } u \notin \mathrm{AN}_{\mathcal{G}}(v) \text{ and } v \notin \mathrm{AN}_{\mathcal{G}}(u). \end{cases}$$

An important property of the induced DMAG is that it preserves all ancestral and non-ancestral relations. More precisely, for two nodes u, v in ADMG $\mathcal{G}$: $u \in \mathrm{AN}_{\mathcal{G}}(v)$ if and only if $u \in \mathrm{AN}_{\mathrm{DMAG}(\mathcal{G})}(v)$. Another important property of the induced DMAG is that it preserves all d-separations. Indeed, $A \perp^d_{\mathrm{DMAG}(\mathcal{G})} B \,|\, C \iff A \perp^d_{\mathcal{G}} B \,|\, C$ for all $A, B, C \subseteq \mathcal{V}$. We sometimes identify a DMAG $\mathcal{H}$ with the set of ADMGs $\mathcal{G}$ such that $\mathrm{DMAG}(\mathcal{G}) = \mathcal{H}$. For an acyclic SCM $\mathcal{M}$, we will define its induced DMAG as $\mathrm{DMAG}(\mathcal{M}) := \mathrm{DMAG}(\mathcal{G}(\mathcal{M}))$.

For a directed maximal ancestral graph $\mathcal{H}$, define its independence model to be

$$\mathrm{IM}(\mathcal{H}) := \{\langle A, B, C\rangle : A, B, C \subseteq \mathcal{V}, A \overset{d}{\underset{\mathcal{H}}{\perp}} B \,|\, C\},$$

i.e., the set of all d-separations entailed by the DMAG. For a simple SCM $\mathcal{M}$ with endogenous index set $\mathcal{I}$ and distribution $\mathbb{P}_{\mathcal{M}}(\boldsymbol{X})$, we define its independence model to be

$$\mathrm{IM}(\mathcal{M}) := \{\langle A, B, C\rangle : A, B, C \subseteq \mathcal{I}, \boldsymbol{X}_A \underset{\mathbb{P}_{\mathcal{M}}}{\perp\!\!\!\perp} \boldsymbol{X}_B \,|\, \boldsymbol{X}_C\},$$

i.e., the set of all (conditional) independences that hold in its (observational) distribution. If $\mathcal{M}$ is acyclic, then by the Markov property, $\mathrm{IM}(\mathcal{M}) \supseteq \mathrm{IM}(\mathrm{DMAG}(\mathcal{M}))$; the faithfulness assumption then means that $\mathrm{IM}(\mathcal{M}) \subseteq \mathrm{IM}(\mathrm{DMAG}(\mathcal{M}))$.

B.1.2. Directed Partial Ancestral Graphs

Since in many cases, the true DMAG is unknown, it is often convenient when performing causal reasoning to be able to represent a set of hypothetical DMAGs in a compact way. For this purpose, *partial ancestral graphs (PAGs)* have been introduced (Zhang, 2006). Again, since we are assuming no selection bias for simplicity, we will only discuss *directed* PAGs (that is, PAGs without undirected or circle-tail edges, i.e., edges of the form $\{—, —\circ, \circ—\}$).

Definition 32 *We call a mixed graph* $\mathcal{G} = \langle \mathcal{V}, \mathcal{E}\rangle$ *with nodes* $\mathcal{V}$ *and edges* $\mathcal{E}$ *of the types* $\{\to, \leftarrow, \leftarrow\!\circ, \leftrightarrow, \circ\!\circ, \circ\!\to\}$ *a* directed partial ancestral graph (DPAG) *if:*

1. *Between any two different nodes there is at most one edge, and there are no self-cycles;*
2. *The graph contains no directed or almost directed cycles ("ancestral");*
3. *There is no inducing path between any two non-adjacent nodes ("maximal").*

Given a DMAG or DPAG, its induced *skeleton* is an undirected graph with the same nodes and with an edge between any pair of nodes if and only if the two nodes are adjacent in the DMAG or DPAG. We often identify a DPAG with the set of all DMAGs that have the same skeleton as the DPAG, have an arrowhead (tail) on each edge mark for which the DPAG has an arrowhead (tail) at that corresponding edge mark, and for each circle in the DPAG, have either an arrowhead or a tail at the corresponding edge mark. Hence, the circles in a DPAG can be thought of as to represent either an arrowhead or a tail.

We extend the definitions of (directed) walks, (directed) paths and colliders for directed mixed graphs to apply also to DPAGs. Edges of the form $i \leftarrow j, i \leftarrow\!\circ\, j, i \leftrightarrow j$ are called *into* i, and similarly, edges of the form $i \rightarrow j, i \circ\!\!\rightarrow j, i \leftrightarrow j$ are called *into* j. Edges of the form $i \rightarrow j$ and $j \leftarrow i$ are called *out of* i. In addition, we define:

Definition 33 *A path $v_0, e_1, v_1, \ldots, v_n$ between nodes v_0 and v_n in a DPAG $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ is called a* possibly directed path from v_0 to v_n *if for each $i = 1, \ldots, n$, the edge e_i between v_{i-1} and v_i is not into v_{i-1} (i.e., is of the form $v_{i-1} \circ\!\!-\!\!\circ v_i$, $v_{i-1} \circ\!\!\rightarrow v_i$, or $v_{i-1} \rightarrow v_i$). The path is called* uncovered *if every subsequent triple is unshielded, i.e., v_i and v_{i-2} are not adjacent in $\mathcal{G}$ for $i = 2, \ldots, n$.*

If $\mathcal{H}_1$ and $\mathcal{H}_2$ are DMAGs, then we call them *Markov equivalent* if $\mathrm{IM}(\mathcal{H}_1) = \mathrm{IM}(\mathcal{H}_2)$. One can show that this implies that $\mathcal{H}_1$ and $\mathcal{H}_2$ must have the same skeleton and the same unshielded colliders. The FCI algorithm maps the independence model $\mathrm{IM}(\mathcal{H})$ of a DMAG $\mathcal{H}$ to a DPAG $\mathcal{P}$. Zhang (2008a) showed that FCI is *sound* and *complete*, which means that

- $\mathcal{P}$ has the same skeleton as $\mathcal{H}$;
- As a set of DMAGs, $\mathcal{P}$ contains $\mathcal{H}$ and all Markov equivalent DMAGs;
- For every circle edge mark in $\mathcal{P}$, there exists a DMAG in $\mathcal{P}$ Markov equivalent to $\mathcal{H}$ that has a tail at the corresponding edge mark, and there exists a DMAG in $\mathcal{P}$ Markov equivalent to $\mathcal{H}$ that has an arrowhead at the corresponding edge mark.

We will denote the completely oriented directed partial ancestral graph that contains the Markov equivalence class of a DMAG $\mathcal{H}$ by $\mathrm{CDPAG}(\mathcal{H})$. For an acyclic SCM $\mathcal{M}$ we will denote its corresponding CDPAG representation as $\mathrm{CDPAG}(\mathcal{M}) := \mathrm{CDPAG}(\mathrm{DMAG}(\mathcal{G}(\mathcal{M})))$.

We will make use of the notion of *(in)visible* edges in a DMAG (Zhang, 2008b):

Definition 34 *A directed edge $i \rightarrow j$ in a DMAG is said to be* visible *if there is a node k not adjacent to j, such that either there is an edge between k and i that is into i, or there is a collider path between k and i that is into i and every collider on the path is a parent of j. Otherwise $i \rightarrow j$ is said to be* invisible.

We will use the same notion in a DPAG, but call it *definitely visible* (and its negation *possibly invisible*). If a directed edge in a DPAG is definitely visible, it must be visible in all DMAGs in the DPAG.

B.2. Soundness and Consistency of FCI-JCI

We are now equipped to prove the soundness (and for some cases, completeness) of FCI-JCI, the adaptation of FCI that incorporates the JCI background knowledge that we introduced in Section 4.2.4.

First, the soundness of FCI-JCI is easy to prove by checking that the soundness of the FCI orientation rules is not invalidated by the JCI background knowledge.

Theorem 35 *Let $\mathcal{M}$ be an acyclic SCM that satisfies JCI Assumption 0 and a subset of JCI Assumptions 1, 2, 3. Suppose that its distribution $\mathbb{P}_{\mathcal{M}}(\boldsymbol{X}, \boldsymbol{C})$ is faithful w.r.t. the graph $\mathcal{G}(\mathcal{M})$. With input $\mathrm{IM}(\mathcal{M})$, and with the right JCI Assumptions, FCI-JCI outputs a DPAG that contains $\mathrm{DMAG}(\mathcal{M})$.*

Proof First note that the skeleton obtained by FCI-JCI must coincide with that of DMAG($\mathcal{M}$) (as it would for standard FCI). Indeed, if JCI Assumption 3 is made, the context nodes are all adjacent in DMAG($\mathcal{M}$) by assumption. For all other edges, and also for edges between context nodes if JCI Assumption 3 is not made: the edge is in the skeleton found by FCI-JCI if and only if it is in the skeleton of DMAG($\mathcal{M}$), for the same reason as for standard FCI.

Furthermore, one can easily see that FCI rule $\mathcal{R}0$ is still sound and will not conflict with the application of the background knowledge stemming from the JCI Assumptions.

The extra orientation rules to incorporate the JCI background knowledge are easily seen to be sound themselves, as they just impose additional features on the DPAG that are satisfied by DMAG($\mathcal{M}$) by assumption.

By checking the soundness proofs of each of the standard FCI orientation rules $\mathcal{R}1$-$\mathcal{R}4$ and $\mathcal{R}8$-$\mathcal{R}10$ in Zhang (2006), it is obvious that all these rules are sound when applied on any DPAG as long as (i) it contains the true DMAG and (ii) FCI rule $\mathcal{R}0$ has been completely applied, i.e., all unshielded colliders have been oriented as such. Rules $\mathcal{R}5$-$\mathcal{R}7$ are not needed since we assumed no selection bias.

Hence all the rules applied by FCI-JCI are sound (as long as they are applied in the prescribed ordering), and hence the final DPAG must contain the true DMAG($\mathcal{M}$). ■

In general, soundness of a constraint-based causal discovery algorithm implies consistency of the algorithm when using appropriate conditional independence tests.

Lemma 36 *If a conditional independence test, including the choice of the sample-size dependent threshold (to decide between the null and alternative hypothesis), is consistent, then any sound constraint-based causal discovery algorithm based on the test is asymptotically consistent.*

Proof Consistency of the conditional independence test means that for any distribution, the probability of a Type I or Type II error converges to 0 as sample size $N \to \infty$. Since the number of tests is finite for a fixed number of variables, and the number of possible predictions made by the algorithm is finite, the probability of *any* error then converges to 0. Any constraint-based algorithm that is sound (i.e., that would return correct answers when using an independence oracle, including the possible answer "unknown") is therefore asymptotically consistent if it makes use of that conditional independence test. ■

As an example of a consistent test, Kalisch and Bühlmann (2007) provide a choice of the threshold for the standard partial correlation test that ensures asymptotic consistency of the test under the assumption that the distribution is multivariate Gaussian. Another example of a distribution-free and even strongly-consistent conditional independence test is proposed by Györfi and Walk (2012). In general, when using the p-value as a test statistic, one should choose the sample-size dependent threshold α_N (where the p-value of the test result is used to decide "dependence" if $p \leq \alpha_N$ and "independence" otherwise) in such a way that $\alpha_N \to 0$ as $N \to \infty$ at a suitable rate.

B.3. Completeness of FCI-JCI

Regarding completeness, we currently only know how to prove the completeness of the variants `FCI-JCI0` and `FCI-JCI123`. In particular, we do not know whether `FCI-JCI1` is complete. The completeness of `FCI-JCI0` is obvious, because `FCI-JCI0` reduces to the standard FCI algorithm without additional background knowledge.

Theorem 37 *Let $\mathcal{M}$ be an acyclic SCM that satisfies JCI Assumption 0. Suppose that its distribution $\mathbb{P}_{\mathcal{M}}(\boldsymbol{X}, \boldsymbol{C})$ is faithful w.r.t. the graph $\mathcal{G}(\mathcal{M})$. With input $\mathrm{IM}(\mathcal{M})$, the output of `FCI-JCI0` is a CDPAG in which all edge marks that can possibly be identified from $\mathrm{IM}(\mathcal{M})$ have been oriented.*

Proof Follows immediately from the completeness of FCI (Zhang, 2008a) under the additional assumption of no selection bias. ■

Proving the completeness of `FCI-JCI123` is more work. The proof strategy is to introduce additional variables that mimic the JCI background knowledge. We can then apply the completeness results for the standard FCI algorithm (Zhang, 2008a).

Theorem 38 *Let $\mathcal{M}$ be an acyclic SCM that satisfies JCI Assumptions 0, 1, 2, 3. Suppose that its distribution $\mathbb{P}_{\mathcal{M}}(\boldsymbol{X}, \boldsymbol{C})$ is faithful w.r.t. the graph $\mathcal{G}(\mathcal{M})$. With input $\mathrm{IM}(\mathcal{M})$, the output of `FCI-JCI123` is a DPAG in which all edge marks that can possibly be identified from $\mathrm{IM}(\mathcal{M})$ and the JCI background knowledge have been oriented.*

Proof The adjacency phase (skeleton search) of `FCI-JCI123` and the orientation of unshielded triples by applying FCI rule $\mathcal{R}0$ are both sound, as we have seen in Theorem 35. Furthermore, the skeleton and unshielded colliders found by `FCI-JCI123` will be the same as found by standard FCI (in particular, note that FCI would not orient any unshielded colliders on a context node, since the true $\mathrm{DMAG}(\mathcal{M})$ does not have these).[33]

Before continuing with the FCI orientation rules $\mathcal{R}1$–$\mathcal{R}4$ and $\mathcal{R}8$–$\mathcal{R}10$, `FCI-JCI123` now uses the JCI background knowledge to orient the following edges:

- $k \leftrightarrow k'$ for all $k \neq k' \in \mathcal{K}$;
- $k \to i$ for $k \in \mathcal{K}$, $i \in \mathcal{I}$ if k and i are adjacent.

After this background orientation step, the only edges in the skeleton that remain to be (further) oriented are the ones connecting two system variables. Denote the DPAG identified by `FCI-JCI123` so far by $\mathcal{P}$.

Each DMAG $\mathcal{H}$ with $\mathrm{IM}(\mathcal{H}) = \mathrm{IM}(\mathcal{M})$ and that satisfies the JCI Assumptions 0, 1, 2, 3 must be contained in $\mathcal{P}$. Consider any such DMAG $\mathcal{H}$. We can extend it to a DMAG $\mathcal{H}^*$, defined over an extended set of variables $\mathcal{I} \dot\cup \mathcal{K} \dot\cup \{r\} \dot\cup \bar{\mathcal{K}}$ where $\bar{\mathcal{K}} := \{\bar{k} : k \in \mathcal{K}\}$ is a copy of $\mathcal{K}$, by adding edges $r \to k$ for all $k \in \mathcal{K}$, adding edges $\bar{k} \to k$ for all $k \in \mathcal{K}$, and removing all bidirected edges $k \leftrightarrow k'$ for all $k \neq k' \in \mathcal{K}$ (see also Figure 43). By construction, the marginal DMAG of $\mathcal{H}^*$ on $\mathcal{I} \cup \mathcal{K}$ is $\mathcal{H}$.[34]

If we run FCI on $\mathrm{IM}(\mathcal{H}^*)$ then the first stages of the algorithm yield a DPAG $\mathcal{P}^*$ with:

33. If we would use the results of statistical conditional independence tests on a finite data sample, then there could be differences between the DPAGs constructed by FCI and `FCI-JCI123` after these stages.
34. For the notion of marginal DMAG, see Richardson and Spirtes (2002).

- the skeleton of $\mathcal{H}^*$, which equals the skeleton of $\mathcal{P}$ together with the additionally constructed edges ($r \ast\!\!-\!\!\ast k$ for all $k \in \mathcal{K}$, and $\bar{k} \ast\!\!-\!\!\ast k$ for all $k \in \mathcal{K}$) but without the edges between the context variables ($k \ast\!\!-\!\!\ast k'$ for $k \neq k' \in \mathcal{K}$);

- the unshielded colliders in $\mathcal{P}$ plus the additionally constructed unshielded colliders ($r \ast\!\!\rightarrow k \leftarrow\!\!\ast \bar{k}$ for all $k \in \mathcal{K}$), which are all identified by rule $\mathcal{R}0$;

- the arrowheads identified by rule $\mathcal{R}1$ that could also be found in $\mathcal{P}$, plus the edge orientations $k \rightarrow i$ for $k \in \mathcal{K}$, $i \in \mathcal{I} \cap \mathrm{CH}_{\mathcal{G}(\mathcal{M})}(k)$ that are obtained from rule $\mathcal{R}1$;

This means that the subgraph of DPAG $\mathcal{P}$ (obtained by `FCI-JCI123` so far) induced on the system nodes $\mathcal{I}$ is identical to the subgraph of DPAG $\mathcal{P}^*$ (obtained by FCI so far) induced on the system nodes $\mathcal{I}$. In addition, all pairs of a system node $i \in \mathcal{I}$ and a context node $k \in \mathcal{K}$ are identically connected in the two DPAGs.

By examining rules $\mathcal{R}1$-$\mathcal{R}4$ and $\mathcal{R}8$-$\mathcal{R}10$ of FCI (which are used without modifications in `FCI-JCI123`) in detail, one can check that they will perform exactly the same edge mark orientations on $\mathcal{P}$ as on $\mathcal{P}^*$. For rules $\mathcal{R}2$, $\mathcal{R}3$ and $\mathcal{R}8$-$\mathcal{R}10$ this is obvious because the only subsets of nodes that play a role in those rules necessarily must be in $\mathcal{I}$. For rules $\mathcal{R}1$ and $\mathcal{R}4$ the situation is only slightly more complicated: a single node appearing in those rules can be in $\mathcal{I} \cup \mathcal{K}$, while all others must be in $\mathcal{I}$. Hence each of these rules is applicable to some tuple of nodes in $\mathcal{P}$ if and only if it is applicable to the same tuple of nodes in $\mathcal{P}^*$. Hence, the final DPAG obtained by `FCI-JCI123` from IM($\mathcal{M}$) and the final DPAG obtained by FCI from IM($\mathcal{H}^*$) induce identical subgraphs on the system nodes $\mathcal{I}$.

Thus, if all DMAGs $\mathcal{H}$ in IM($\mathcal{M}$) that satisfy JCI Assumptions 0, 1, 2, 3 have a certain invariant edge mark on an edge between two system variables, then all extended DMAGs $\mathcal{H}^*$ must have the same invariant edge mark. All these extended DMAGs $\mathcal{H}^*$ must be Markov equivalent, since FCI arrives at the same CDPAG for all IM($\mathcal{H}^*$). Now suppose `FCI-JCI123` left an edge mark on an edge between two system variables unoriented. Then FCI must also leave the corresponding edge mark unoriented when it is run on IM(DMAG($\mathcal{M}$)*). This means that there must exist DMAGs that are Markov equivalent to DMAG($\mathcal{M}$)* that have an arrowhead at that spot, but also DMAGs that are Markov equivalent to DMAG($\mathcal{M}$)* that have a tail at that spot. Marginalizing those DMAGs down to $\mathcal{I} \cup \mathcal{K}$ gives DMAGs that are Markov equivalent to DMAG($\mathcal{M}$), satisfy JCI Assumptions 0, 1, 2, 3 and have an arrowhead respectively a tail at that spot. This means that all edge marks between system variables that could possibly be oriented, have been oriented by `FCI-JCI123`. This completes the proof of arrowhead and tail completeness of `FCI-JCI123`. ■

B.4. Reading off Definite (Non-)Ancestors From a DPAG

Zhang (2006) conjectured the soundness and completeness of a criterion to read off definite ancestral relations from a CDPAG. Roumpelaki et al. (2016) proved soundness of this criterion.[35] We will need a slightly stronger result (with a similar proof) for DPAGs:

35. Roumpelaki et al. (2016) also claim to have proved completeness, but their proof is flawed: the last part of the proof that aims to prove that u, v are non-adjacent appears to be incomplete.

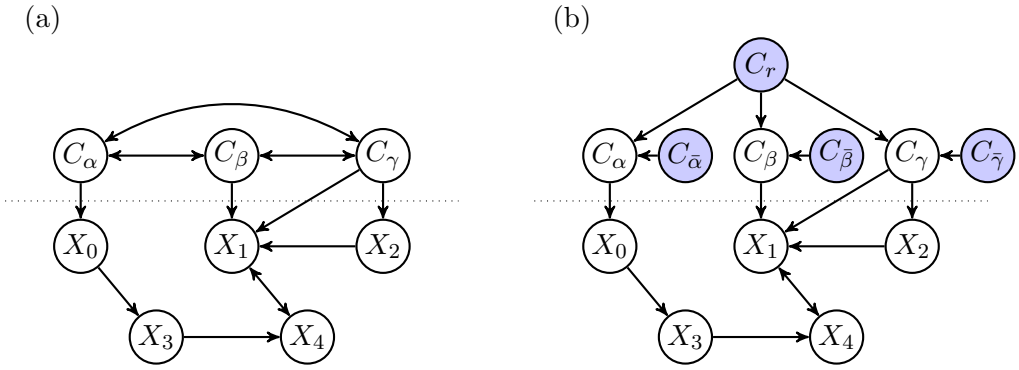


Figure 43: (a) Example DMAG satisfying JCI Assumptions 1, 2, 3 and (b) corresponding extended DMAG with additional variables as used in the proof of Theorem 38.

Proposition 39 *Let $\mathcal{M}$ be an acyclic SCM. Let $\mathcal{P}$ be a DPAG that contains* $\mathrm{DMAG}(\mathcal{M})$*, and in which all unshielded colliders in* $\mathrm{DMAG}(\mathcal{M})$ *have been oriented. For two nodes $i, j \in \mathcal{P}$: If*

- *there is a directed path from i to j in $\mathcal{P}$, or*
- *there exist uncovered possibly directed paths from i to j in $\mathcal{P}$ of the form $i, u, \ldots, j$ and $i, v, \ldots, j$ such that u, v are non-adjacent nodes in $\mathcal{P}$,*

then i causes j according to $\mathcal{M}$, i.e., $i \in \mathrm{AN}_{\mathcal{G}(\mathcal{M})}(j)$.

Proof First, if there is a directed path from i to j in $\mathcal{P}$, it must be in any DMAG in $\mathcal{P}$, hence there must be a directed path from i to j in $\mathrm{DMAG}(\mathcal{M})$ as well. Therefore $i \in \mathrm{AN}_{\mathcal{G}(\mathcal{M})}(j)$.

Second, assume that there exist uncovered possibly directed paths from i to j in $\mathcal{P}$ of the form $i, u, \ldots, j$ and $i, v, \ldots, j$ such that u, v are non-adjacent in $\mathcal{P}$. If $\mathrm{DMAG}(\mathcal{M})$ has $i \to u$, the path $i, u, \ldots, j$ must actually correspond to a directed path in $\mathrm{DMAG}(\mathcal{M})$ because otherwise it would contain unshielded colliders that were not oriented, contradicting the assumptions. If $\mathrm{DMAG}(\mathcal{M})$ has $i \leftarrow\!\!* u$ instead, it must have $i \to v$ to avoid an unshielded collider $u \mathrel{*\!\!\rightarrow} i \leftarrow\!\!* v$ that was not oriented, and hence must have a directed path $i, v, \ldots, j$. In both cases, $\mathrm{DMAG}(\mathcal{M})$ must have a directed path from i to j, and hence $i \in \mathrm{AN}_{\mathcal{G}(\mathcal{M})}(j)$. ■

Zhang (2006, p. 137) provides a sound and complete criterion to read off definite non-ancestors from a CDPAG. It is easy to prove the soundness of the criterion also for (arbitrary) DPAGs:

Proposition 40 *Let $\mathcal{M}$ be an acyclic SCM. Let $\mathcal{P}$ be a DPAG that contains* $\mathrm{DMAG}(\mathcal{M})$*. For two nodes $i, j \in \mathcal{P}$: if there is no possibly-directed path from i to j in $\mathcal{P}$ then $i \notin \mathrm{AN}_{\mathcal{G}(\mathcal{M})}(j)$.*

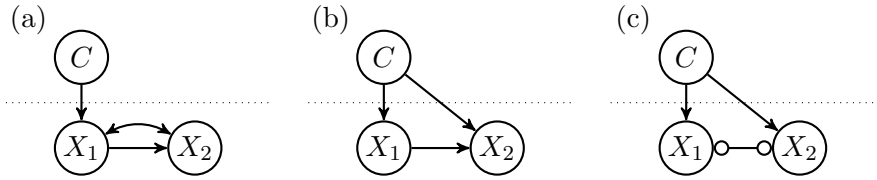


Figure 44: Example to illustrate that directed edges in the DPAG obtained by `FCI-JCI123` do not necessarily correspond with a direct cause. (a) Graph $\mathcal{G}(\mathcal{M})$, satisfying JCI Assumptions 1, 2, 3; (b) corresponding DMAG($\mathcal{M}$); (c) corresponding DPAG $\mathcal{P}$ output by `FCI-JCI123`. The directed edge $C \to X_2$ in $\mathcal{P}$ identified by `FCI-JCI123` does not correspond with a direct causal effect of C on X_2 (note that there is no directed edge $C \to X_2$ in $\mathcal{G}(\mathcal{M})$).

Proof If $i \in \text{AN}_{\mathcal{G}(\mathcal{M})}(j)$ then there is a directed path from i to j in DMAG($\mathcal{M}$). Since $\mathcal{P}$ contains DMAG($\mathcal{M}$), this must correspond with a possibly-directed path from i to j in $\mathcal{P}$. ■

These two propositions allow us to read off (a subset of the) ancestral and non-ancestral relations that are identifiable from the conditional independences in the joint distribution and the JCI background knowledge (if applicable) from the DPAGs output by the various FCI variants (`FCI-JCI123`, `FCI-JCI1`, `FCI-JCI0`, FCI).

B.5. Discovering Direct Intervention Targets with `FCI-JCI123`

One of the features of `FCI-JCI123` is that it allows one to read off direct intervention targets from the DPAG that it outputs. Naïvely interpreting a directed edge $k \to i$ from context node $k \in \mathcal{K}$ to system node $i \in \mathcal{I}$ in the DPAG output by `FCI-JCI123` as meaning that k directly targets i is incorrect, as can be seen from the example in Figure 44. Here we propose a provably correct (but possibly incomplete) procedure.

We will first consider how to read direct intervention targets from DMAGs before applying this to DPAGs.

Lemma 41 *Let $\mathcal{M}$ be an acyclic SCM that satisfies JCI Assumptions 0, 1, 2, 3. For $k \in \mathcal{K}$, $i \in \mathcal{I}$: if*

- *$k \to i$ in* DMAG($\mathcal{M}$)*, and*
- *for all nodes j in* DMAG($\mathcal{M}$) *s.t. $k \to j \to i$ in* DMAG($\mathcal{M}$)*, $j \to i$ is visible,*

then k is a direct cause of i according to $\mathcal{M}$, i.e., $k \to i \in \mathcal{G}(\mathcal{M})$.

Proof Because the edge $k \to i$ is present in DMAG($\mathcal{M}$), there must be an inducing path between k and i in $\mathcal{G}(\mathcal{M})$. This must be a collider path into i where each collider is ancestor of i. First suppose that the path consists of more than a single edge. Denote its first collider (the one adjacent to k) by j. If the first edge on the path would be into k, then it must be $k \leftrightarrow j$ and j must be a context node (because of JCI Assumptions 1, 2). Similarly, all subsequent nodes on the inducing path (except for the final node i) must be collider nodes

and hence in $\mathcal{K}$. But then the final edge is between a context node and system node i and into the context node, contradicting JCI Assumption 1 or 2. Hence the first edge on the inducing path must be $k \to j$. The same edge $k \to j$ must then occur in DMAG($\mathcal{M}$). The remainder of the inducing path is actually an inducing path between j and i that is into j. By Zhang (2008b, Lemma 9), $j \to i$ is in DMAG($\mathcal{M}$) and it is invisible. This contradicts the assumption. Therefore the inducing path in $\mathcal{G}(\mathcal{M})$ between k and i must consist of a single edge. This must be out of k because of JCI Assumptions 1, 2, and is thus necessarily of the form $k \to i$. Hence $k \to i$ is in $\mathcal{G}(\mathcal{M})$. ■

The following result enables us to read off direct intervention targets from the DPAG output by `FCI-JCI123`.

Proposition 42 *Let $\mathcal{M}$ be an acyclic SCM that satisfies JCI Assumptions 0, 1, 2, 3. Suppose that its distribution $\mathbb{P}_{\mathcal{M}}(\boldsymbol{X}, \boldsymbol{C})$ is faithful w.r.t. the graph $\mathcal{G}(\mathcal{M})$. Let $\mathcal{P}$ be the DPAG output by `FCI-JCI123` with input* IM($\mathcal{M}$). *Let $k \in \mathcal{K}$, $i \in \mathcal{I}$.*

- *If k is not adjacent to i in $\mathcal{P}$, k is not a direct cause of i according to $\mathcal{M}$, i.e., $k \to i \notin \mathcal{G}(\mathcal{M})$.*
- *If:*
  1. *$k \to i$ in $\mathcal{P}$, and*
  2. *for all system nodes $j \in \mathcal{I}$ s.t. $k \to j$ in $\mathcal{P}$ and $j \circ\!\!-\!\!\circ i$ or $j \circ\!\!\to i$ or $j \to i$ in $\mathcal{P}$, the edge $j \to i$ is definitely visible in the DPAG obtained from $\mathcal{P}$ by replacing the edge between j and i by $j \to i$,*

  then k is a direct cause of i according to $\mathcal{M}$, i.e., $k \to i \in \mathcal{G}(\mathcal{M})$.

Proof Because `FCI-JCI123` is sound (Theorem 38), $\mathcal{P}$ contains DMAG($\mathcal{M}$).

For the first statement: if k is not adjacent to i in $\mathcal{P}$, then the two nodes are not adjacent in any DMAG in $\mathcal{P}$, and in particular, in DMAG($\mathcal{M}$). This means that $k \to i \notin \mathcal{G}(\mathcal{M})$, because otherwise, $k \to i$ would be in DMAG($\mathcal{M}$), a contradiction.

The second statement follows from Lemma 41 and from the JCI Assumptions 1, 2, 3. ■

If the context variables represent interventions, then this allows us to learn (a subset of) the direct targets and non-targets of each intervention. While it is easy to see that this criterion is sound, we do not know whether it is complete.

Appendix C. ASD: Accounting for Strong Dependencies

The causal discovery and reasoning algorithm that we refer to as ASD (Accounting for Strong Dependencies) used in this work is based on an algorithm proposed by Hyttinen et al. (2014) and extensions proposed by Magliacane et al. (2016b) and Forré and Mooij (2018). To make this paper more self-contained, we give here a short description of this algorithm, referring the reader for more details to the original publications.

Hyttinen et al. (2014) formulate causal discovery as an optimization problem where a loss function is minimized over possible causal graphs. Intuitively, the loss function can be thought of as measuring the amount of evidence *against* the hypothesis that the data was generated by an SCM with a particular graph. The loss function depends on the hypothetical causal graph and on a list of input statements. For the purely observational case,

the input consists of a list $S = \big((a_j, b_j, Z_j, \lambda_j)\big)_{j=1}^n$ of weighted conditional independence statements. Here, the weighted statement $(a_j, b_j, Z_j, \lambda_j)$ with $\{a_j\}, \{b_j\}, Z_j$ disjoint sets of endogenous variable indices and $\lambda_j \in \bar{\mathbb{R}} := \mathbb{R} \cup \{-\infty, +\infty\}$ encodes that the conditional independence $X_{a_j} \perp\!\!\!\perp X_{b_j} \,|\, \boldsymbol{X}_{Z_j}$ holds with "confidence" λ_j, where a finite value of λ_j gives a "soft constraint" and a value of $\lambda_j = \pm\infty$ imposes a "hard constraint". Positive weights encode that we have empirical support *in favor* of the independence, whereas negative weights encode empirical support *against* the independence (in other words, in favor of *dependence*). The loss function simply sums the absolute weights of all the input statements that would be violated if the true causal graph would consist of the hypothetical one:

$$\mathcal{L}(\mathcal{G}, S) := \sum_{(a_j, b_j, Z_j, \lambda_j) \in S} \lambda_j (\mathbb{1}_{\lambda_j > 0} - \mathbb{1}_{a_j \perp_{\mathcal{G}} b_j \,|\, Z_j}),$$

where $\mathbb{1}$ is the indicator function. While the original implementation by Hyttinen et al. (2014) makes use of d-separation, Forré and Mooij (2018) show how this can be modified for σ-separation. Causal discovery can now be formulated as the optimization problem:

$$\mathcal{G}^* = \operatorname*{arg\,min}_{\mathcal{G} \in \mathbb{G}(\mathcal{I})} \mathcal{L}(\mathcal{G}, S), \tag{7}$$

where $\mathbb{G}(\mathcal{I})$ denotes the set of all possible causal graphs with nodes $\mathcal{I}$ (ADMGs in the acyclic case, and DMGs in the cyclic case).

The optimization problem (7) may have multiple minima, for example because the underlying causal graph is not 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. Let $f : \mathbb{G}(\mathcal{I}) \to \{0, 1\}$ be a feature, i.e., a Boolean function of the causal graph $\mathcal{G}$. We employ the method proposed by Magliacane et al. (2016b) for scoring the confidence that feature f is present in the causal graph by calculating the difference between the optimal losses under the additional hard constraints that the feature f is present vs. that the feature f is absent in the causal graph:

$$C(f, S) := \min_{\mathcal{G} \in \mathbb{G}(\mathcal{I}) : \neg f(\mathcal{G})} \mathcal{L}(\mathcal{G}, S) - \min_{\mathcal{G} \in \mathbb{G}(\mathcal{I}) : f(\mathcal{G})} \mathcal{L}(\mathcal{G}, S). \tag{8}$$

This confidence is positive if there is less evidence against its presence than against its absence, negative if there is less evidence against its absence than its presence, and vanishes if there is as much evidence against its presence as there is against its absence. As features, we can consider for example the presence of a direct causal relation, the presence of an (ancestral) causal relation, and the presence of a latent confounder.

Magliacane et al. (2016b) showed that this scoring method is sound for oracle inputs.

Theorem 43 *For any feature $f : \mathbb{G}(\mathcal{I}) \to \{0, 1\}$, the ASD confidence score $C(f, S)$ of (8) is sound and complete for oracle inputs with infinite weights. In other words, $C(f, S) = \infty$ if f is identifiable from the inputs, $C(f, S) = -\infty$ if $\neg f$ is identifiable from the inputs, and $C(f, S) = 0$ if f is unidentifiable from the inputs.*

Additionally, Magliacane et al. (2016b) showed that the scoring method is asymptotically consistent under a consistency condition on the weights that encode the confidence of conditional (in)dependence.

Theorem 44 *Assume that the weights are asymptotically consistent, meaning that*

$$\lambda_j \xrightarrow{P} \begin{cases} -\infty & X_{a_j} \not\perp\!\!\!\perp X_{b_j} \,|\, \boldsymbol{X}_{Z_j} \\ +\infty & X_{a_j} \perp\!\!\!\perp X_{b_j} \,|\, \boldsymbol{X}_{Z_j}, \end{cases} \tag{9}$$

(where $\xrightarrow{P}$ means convergence in probability) as the number of samples $N \to \infty$. Then for any feature $f : \mathbb{G}(\mathcal{I}) \to \{0, 1\}$, the ASD confidence score $C(f, S)$ of (8) is asymptotically consistent, i.e., $C(f, S) \xrightarrow{P} \infty$ if f is identifiably true, $C(f, S) \xrightarrow{P} -\infty$ if f is identifiably false, and $C(f, S) \xrightarrow{P} 0$ otherwise.

In our experiments, we used 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 with independence as null hypothesis, and α is a significance level (e.g., 1%). These weights have the desirable property that independences typically get a smaller absolute weight than strong dependencies. This leads to the strong dependencies dominating the loss function, which explains the acronym ASD (Accounting for Strong Dependencies) that we use here to describe this method. By choosing a sample-size dependent threshold α_N such that $\alpha_N \to 0$ as $N \to \infty$ at a suitable rate, these weights may become asymptotically consistent. Kalisch and Bühlmann (2007) provide a choice of α_N for partial correlation tests that ensures asymptotic consistency under the assumption that the distribution is multivariate Gaussian. Another possibility for obtaining consistent weights would be to base them on the distribution-free and strongly-consistent conditional independence test proposed by Györfi and Walk (2012).

References

R. Ayesha Ali, Thomas S. Richardson, and Peter Spirtes. Markov equivalence for ancestral graphs. *The Annals of Statistics*, 37(5B):2808–2837, 2009.

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 2018)*, 2018.

Stephan Bongers, Patrick Forré, Jonas Peters, Bernhard Schölkopf, and Joris M. Mooij. Foundations of structural causal models with cycles and latent variables. *arXiv.org preprint*, arXiv:1611.06221v3 [stat.ME], May 2020. URL `https://arxiv.org/abs/1611.06221v3`.

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 *Advances in Neural Information Processing Systems 23 (NIPS 2010)*, pages 415–423, Vancouver, British Columbia, Canada, 2010.

Tom Claassen and Tom Heskes. A logical characterization of constraint-based causal discovery. In *Proceedings of the 27th Annual Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, pages 135–144, 2011.

Tom Claassen, Joris M. Mooij, and Tom Heskes. Learning sparse causal models is not NP-hard. In Ann Nicholson and Padhraic Smyth, editors, *Proceedings of the 29th Annual Conference on Uncertainty in Artificial Intelligence (UAI 2013)*, pages 172–181. AUAI Press, 2013.

Diego Colombo and Marloes H. Maathuis. Order-independent constraint-based causal structure learning. *The Journal of Machine Learning Research*, 15(1):3741–3782, 2014.

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 Edward Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.

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 22nd 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.

Doris Entner and Patrik 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.

Ronald A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, 1925.

Ronald 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 2018)*, 2018.

Patrick Forré and Joris M. Mooij. Causal calculus in the presence of cycles, latent confounders and selection bias. In *Proceedings of the 35th Annual Conference on Uncertainty in Artificial Intelligence (UAI 2019)*, 2019.

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.

László Györfi and Harro Walk. Strongly consistent nonparametric tests of conditional independence. *Statistics & Probability Letters*, 82:1145–1150, June 2012.

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. *Journal of Causal Inference*, 6(2):20170016, 2018.

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 13th Conference on Uncertainty in Artificial Intelligence (UAI 2014)*, pages 340–349, Quebec City, Quebec, Canada, 2014.

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

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.

Patrick Kemmeren, Katrin Sameith, Loes A.L. van de Pasch, Joris J. Benschop, Tineke L. Lenstra, Thanasis Margaritis, Eoghan ODuibhir, Eva Apweiler, Sake van Wageningen, Cheuk W. Ko, Sebastiaan van Heesch, Mehdi M. Kashani, Giannis Ampatziadis-Michailidis, Mariel O. Brok, Nathalie A.C.H. Brabers, Anthony J. Miles, Diane Bouwmeester, Sander R. van Hooff, Harm van Bakel, Erik Sluiters, Linda V. Bakker, Berend Snel, Philip Lijnzaad, Dik van Leenen, Marian J.A. Groot Koerkamp, and Frank C.P. Holstege. Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell*, 157(3):740–752, 2014.

Mikko Koivisto and Kismat Sood. Exact Bayesian structure discovery in Bayesian networks. *Journal of Machine Learning Research*, 5:549–573, 2004.

Jan T. A. Koster. On the validity of the Markov interpretation of path diagrams of Gaussian structural equations systems with correlated errors. *Scandinavian Journal of Statistics*, 26:413–431, 1999.

Jan T.A. Koster. Markov properties of nonrecursive causal models. *Annals of Statistics*, 24 (5):2148–2177, 1996.

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 *Advances in Neural Information Processing Systems 29 (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 (NeurIPS 2018)*, pages 10869–10879. Curran Associates, Inc., 2018.

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 11th Conference on Uncertainty in Artificial Intelligence (UAI 1995)*, pages 411–419, 1995.

Nicolai Meinshausen, Alain Hauser, Joris M. Mooij, Jonas Peters, Philip Versteeg, and Peter Bühlmann. Methods for causal inference from gene perturbation experiments and validation. *Proceedings of the National Academy of Sciences of the United States of America*, 113(27):7361–7368, 2016.

Joris M. Mooij and Tom Claassen. Constraint-based causal discovery in the presence of cycles. *arXiv.org preprint*, arXiv:2005.00610 [math.ST], May 2020. URL `https://arxiv.org/abs/2005.00610`.

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`.

Radford M. Neal. On deducing conditional independence from d-separation in causal graphs with feedback. *Journal of Artificial Intelligence Research*, 12:87–91, 2000.

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.

Judea 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.

Judea Pearl and Rina Dechter. Identifying independence in causal graphs with feedback. In *Proceedings of the 12th Annual Conference on Uncertainty in Artificial Intelligence (UAI 1996)*, pages 420–426, 1996.

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.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, USA, 2017.

Joseph Ramsey and Bryan Andrews. FASK with interventional knowledge recovers edges from the Sachs model. *arXiv.org preprint*, arXiv:1805.03108 [q-bio.MN], 2018. URL `https://arxiv.org/abs/1805.03108`.

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. BACKSHIFT: 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. Conditional independence in directed cyclic graphical models for feedback. Technical Report CMU-PHIL-54, Carnegie Mellon University, 1994.

Peter Spirtes. Directed cyclic graphical representations of feedback models. In Philippe Besnard and Steve Hanks, editors, *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (UAI 1995)*, pages 499–506, San Francisco, CA, USA, 1995. Morgan Kaufmann.

Peter Spirtes, Thomas Richardson, Cristopher Meek, Richard Scheines, and Clark Glymour. Using path diagrams as a structural equation modelling tool. *Sociological Methods & Research*, 27:182–225, 1998.

Peter Spirtes, Christopher Meek, and Thomas S. Richardson. An algorithm for causal inference in the presence of latent variables and selection bias. In Clark Glymour and Gregory F. Cooper, editors, *Computation, Causation and Discovery*, chapter 6, pages 211–252. The MIT Press, 1999.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2000.

Eric V. Strobl. A constraint-based algorithm for causal discovery with cycles, latent variables and selection bias. *International Journal of Data Science and Analytics*, 2018.

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 2017)*, 2017.

Philip J.J.P. Versteeg and Joris M. Mooij. Boosting local causal discovery in high-dimensional expression data. *arXiv.org preprint*, arXiv:1910.02505v2 [stat.ML], November 2019. URL `https://arxiv.org/abs/1910.02505v2`. Accepted for publication in BIBM 2019.

Larry Wasserman. *All of Statistics*. Springer Texts in Statistics. Springer, 2004.

Sewall Wright. Correlation and causation. *Journal of Agricultural Research*, 20:557–585, 1921.

Karren D. Yang, Abigail Katcoff, and Caroline Uhler. Characterizing and learning equivalence classes of causal DAGs under interventions. In *Proceedings of Machine Learning Research 80 (ICML 2018)*, pages 5537–5546, 2018.

Jiji Zhang. *Causal Inference and Reasoning in Causally Insufficient Systems*. PhD thesis, Carnegie Mellon University, July 2006. URL `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.466.7206&rep=rep1&type=pdf`.

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, 2008a.

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

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 2017)*, pages 1347–1353, 2017.