Journal of Machine Learning Research ? (????) ??-?? Submitted ?/??; Published ?/??

Joint Causal Inference from Observational and Experimental Datasets

Sara Magliacane SARA.MAGLIACANE@GMAIL.COM
VU Amsterdam
Amsterdam, The Netherlands

Institute for Informatics, University of Amsterdam
Amsterdam, The Netherlands

Tom Claassen TOMC@CS.RU.NL
Radboud University Nijmegen
Nijmegen, The Netherlands

Institute for Informatics, University of Amsterdam
Amsterdam, The Netherlands

Joris M. Mooij J.M.MOOIJ@UVA.NL
Institute for Informatics, University of Amsterdam
Amsterdam, The Netherlands

Editor: ??

Abstract

We introduce Joint Causal Inference (JCI), a powerful formulation of causal discovery from multiple datasets that allows to jointly learn both the causal structure and targets of interventions from statistical independences in pooled data. Compared with existing constraint-based approaches for causal discovery from multiple data sets, JCI offers several advantages: it allows for several different types of interventions in a unified fashion, it can learn intervention targets, it systematically pools data across different datasets which improves the statistical power of independence tests, and most importantly, it improves on the accuracy and identifiability of the predicted causal relations. A technical complication that arises in JCI is the occurrence of faithfulness violations due to deterministic relations. We propose a simple but effective strategy for dealing with this type of faithfulness violations. We implement it in ACID, a determinism-tolerant extension of Ancestral Causal Inference (ACI) (Magliacane et al., 2016), a recently proposed logic-based causal discovery method that improves reliability of the output by exploiting redundant information in the data. We illustrate the benefits of JCI with ACID with an evaluation on a simulated dataset.

Keywords: Causal Inference, Structure Learning, Constraint-Based Causal Discovery, Observational and Experimental Data, Interventions

1. Introduction

Discovering causal relations from data is at the foundation of the scientific method. Traditionally, causal relations are either recovered from experimental data in which the variable of interest is perturbed, or from purely observational data, e.g., using the seminal PC and FCI algorithms (Spirtes et al., 2000; Zhang, 2008).

In recent years, several methods for combining observational and experimental data to discover causal relations have been proposed, showing that this combination can improve greatly on the

accuracy and identifiability of the predicted causal relations. Some of the proposed methods are *score-based* (e.g., Cooper and Yoo, 1999; Tian and Pearl, 2001; Eaton and Murphy, 2007; Hauser and Bühlmann, 2012; Mooij and Heskes, 2013), i.e., they evaluate models using a penalized likelihood score, while others (e.g., Tian and Pearl, 2001; Claassen and Heskes, 2010; Hyttinen et al., 2014; Triantafillou and Tsamardinos, 2015; Peters et al., 2015; Magliacane et al., 2016; Borboudakis and Tsamardinos, 2016) are *constraint-based*, i.e., they use statistical independences to express constraints over possible models.

In this work we propose Joint Causal Inference (JCI), a formulation of causal discovery over multiple datasets in which both the causal structure and targets of interventions are jointly learnt from independence test results in pooled data. A related approach was already proposed for score-based methods by Eaton and Murphy (2007), but here we extend it so that constraint-based methods can be employed. Our goal is to combine the idea of joint inference from observational and experimental data from Eaton and Murphy (2007) with the advantages that constraint-based methods have over score-based methods, namely, the ability to handle latent confounders and selection bias naturally in a nonparametric approach, and, especially in the case of logic-based methods, an easy integration of background knowledge.

Existing constraint-based methods for multiple datasets typically learn the causal structure on each dataset separately and then merge the learnt structures (e.g., Claassen and Heskes, 2010; Hyttinen et al., 2014; Triantafillou and Tsamardinos, 2015; Borboudakis and Tsamardinos, 2016). The merging process depends on the type of interventions, and most existing methods support only interventions on known targets. Instead, JCI: (1) allows for several different types of interventions; (2) can learn the intervention targets; (3) systematically pools data across different datasets, which improves the statistical power of independence tests; and (4) improves the identifiability and accuracy of the predicted causal relations.

On the other hand, JCI poses challenges for current constraint-based methods because of their susceptibility to violations of the *Causal Faithfulness* assumption. Specifically, JCI induces faithfulness violations due to deterministic relations, which would typically result in erroneous inferences with standard constraint-based methods. We propose a simple but effective strategy for dealing with this type of faithfulness violations. The strategy can be applied to any constraint-based causal discovery method for observational data that can handle partial inputs, i.e. missing results for a certain independence test, thus extending it to a JCI method that can handle a combination of observational and experimental data. We implement the strategy in ACID (Ancestral Causal Inference with Determinism), a determinism-tolerant extension of ACI (Ancestral Causal Inference) (Magliacane et al., 2016), a recently proposed logic-based causal discovery method that improves reliability of the output by exploiting redundant information in its input. In our evaluation on synthetic data we show that JCI with ACID improves on the accuracy of the causal predictions with respect to simply merging separately learned causal graphs, illustrating the advantage of *joint* causal discovery.

2. Preliminaries

In this section we review a few useful concepts from the related work and introduce the notation we use in the rest of the paper. Most of the concepts described here are explained in detail in the seminal books by Pearl (2009) and Spirtes et al. (2000). In the following, we represent variables with uppercase letters, while sets of variables are denoted by boldface.

2.1 Graph terminology

Throughout the paper we assume that the data generating process can be modeled by a causal Directed Acyclic Graph (DAG) that may contain latent variables. For simplicity, we do not consider selection bias. A directed edge $X \to Y$ in the causal DAG represents a direct causal relationship of cause X on effect Y. We then say that X is a *parent* of Y, and denote the set of parents of Y as $\text{PA}(Y)$. A sequence of directed edges $X_1 \to X_2 \to \cdots \to X_n$ is a *directed path*. If there is a directed path from X to Y (or $X = Y$) then X is an *ancestor* of Y (denoted as $X \dashrightarrow Y$). We denote the set of ancestors of Y as $\text{AN}(Y)$. If there is no directed path from X to Y (and $X \neq Y$) we denote this as $X \not\dashrightarrow Y$. A sequence of unique nodes $\langle X_1, \ldots, X_n \rangle$ such that each pair $\{X_i, X_{i+1}\}$ is connected by an edge is called a *path*. A *collider* is a node on a path that has two incoming arrow heads from both its neighboring nodes on the path, i.e., of the form $X_{i-1} \to X_i \leftarrow X_{i+1}$. Any other node on a path (including the end nodes X_1, X_n) is called a *non-collider*.

For a set of variables $\boldsymbol{W}$, we extend the definition of parents $\text{PA}(\boldsymbol{W})$ to the union of all parents of any variable $W \in \boldsymbol{W}$. We similarly define $\text{AN}(\boldsymbol{W})$ as the set of ancestors of all $W \in \boldsymbol{W}$. We write $X \dashrightarrow \boldsymbol{W}$ if there exists at least one effect $Y \in \boldsymbol{W}$ that has X as an ancestor, i.e., $X \dashrightarrow Y$. We write $X \not\dashrightarrow \boldsymbol{W}$ if $X \not\dashrightarrow Y$ for all $Y \in \boldsymbol{W}$.

2.2 Independences and d-separation

For disjoint sets of random variables $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{W}$, distributed according to a joint probability distribution $\mathcal{P}$, we denote the conditional independence of $\boldsymbol{X}$ and $\boldsymbol{Y}$ given $\boldsymbol{W}$ in $\mathcal{P}$ as $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{W}\ [\mathcal{P}]$, and conditional dependence as $\boldsymbol{X} \not\perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{W}\ [\mathcal{P}]$. We often omit $[\mathcal{P}]$ when it is obvious from the context which probability distribution we are referring to. We call the cardinality $|\boldsymbol{W}|$ the *order* of the conditional (in)dependence relation.

A well-known graphical criterion for DAGs, with many implications for causal discovery, is *d-separation* (Pearl, 2009; Spirtes et al., 2000):

Definition 1 *For disjoint sets of variables $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{W}$ in a DAG $\mathcal{G}$, we say that $\boldsymbol{X}$* is d-separated *from $\boldsymbol{Y}$ by $\boldsymbol{W}$ in $\mathcal{G}$, written $X \perp_d Y \mid \boldsymbol{W}\ [\mathcal{G}]$, iff every path π in $\mathcal{G}$ that connects any $X \in \boldsymbol{X}$ with any $Y \in \boldsymbol{Y}$ is* blocked *by $\boldsymbol{W}$, i.e. at least one of the following holds:*

- *π contains a collider not in $\text{AN}(\boldsymbol{W})$, or*
- *π contains a non-collider in $\boldsymbol{W}$.*

The opposite, i.e., d-connection, is denoted as $\boldsymbol{X} \not\perp_d \boldsymbol{Y} \mid \boldsymbol{W}\ [\mathcal{G}]$. We often omit $[\mathcal{G}]$ from the notation when it is obvious which DAG we are referring to.

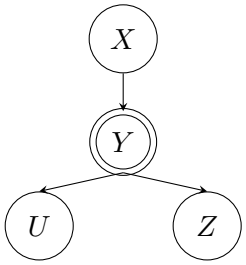
The following two key assumptions for constraint-based causal discovery have been thoroughly discussed in the literature (see for example Spirtes et al., 2000). They connect conditional independences in the observational distribution $\mathcal{P}$ with d-separations in the underlying causal DAG $\mathcal{G}$.

- *Causal Markov Assumption*: d-separation in the causal DAG $\mathcal{G}$ implies conditional independence in the observational distribution $\mathcal{P}$. For all disjoint sets of variables $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{W}$:

$$\boldsymbol{X} \perp_d \boldsymbol{Y} \mid \boldsymbol{W}\ [\mathcal{G}] \implies \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{W}\ [\mathcal{P}],$$

  which can also be expressed contrapositively as:

$$\boldsymbol{X} \not\perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{W}\ [\mathcal{P}] \implies \boldsymbol{X} \not\perp_d \boldsymbol{Y} \mid \boldsymbol{W}\ [\mathcal{G}].$$



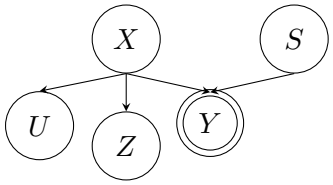
$Y = f(X) \implies$

$Y \not\perp\!\!\!\perp_d Z \,|\, X$ $\quad$ $Y \perp\!\!\!\perp Z \,|\, X$

$Y \not\perp\!\!\!\perp_d U \,|\, X$ $\quad$ $Y \perp\!\!\!\perp U \,|\, X$

$Z \not\perp\!\!\!\perp_d U \,|\, X$ $\quad$ $Z \perp\!\!\!\perp U \,|\, X$

Functionally determined example



$Y = X \oplus S \implies$

$X \not\perp\!\!\!\perp_d U \,|\, \{S, Y\}$ $\quad$ $X \perp\!\!\!\perp U \,|\, \{S, Y\}$

$X \not\perp\!\!\!\perp_d Z \,|\, \{S, Y\}$ $\quad$ $X \perp\!\!\!\perp Z \,|\, \{S, Y\}$

$Z \not\perp\!\!\!\perp_d U \,|\, \{S, Y\}$ $\quad$ $Z \perp\!\!\!\perp U \,|\, \{S, Y\}$

Non-functionally determined example

Figure 1: **Examples of faithfulness violations**: In the left example, we show a set of faithfulness violations due to a functionally determined relation, in which the parent X fully determines the child Y. In the right example, X and S are binary variables and $\oplus$ is the XOR function. Conditioning on $\{S, Y\}$ fully determines X, even though $\{S, Y\}$ are not ancestors of X. This creates a series of non-trivial faithfulness violations, listed below the graph. In both graphs, we represent the variables resulting from structural equations without noise terms with a double circled node.

- *Causal Faithfulness Assumption*: the inverse, i.e., for all disjoint sets of variables $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{W}$:

$$\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \,|\, \boldsymbol{W} \; [\mathcal{P}] \implies \boldsymbol{X} \perp_d \boldsymbol{Y} \,|\, \boldsymbol{W} \; [\mathcal{G}].$$

If we assume both the Causal Markov and Causal Faithfulness assumptions to hold, the conditional independences of the observational distribution correspond one-to-one with the d-separations in the causal DAG. This setting is very favourable for causal discovery, thus both assumptions are usually made in constraint-based approaches.

2.3 Deterministic relations and faithfulness violations

Although often reasonable, the Causal Faithfulness assumption is violated in some cases, notably in the common case of deterministic relations among variables, e.g., for Structural Causal Model equations (Pearl, 2009) in which there is no noise term. Some of the faithfulness violations related to determinism are captured by an extension to the d-separation criterion, the *D-separation* criterion, first introduced by Geiger et al. (1990) and later extended by Spirtes et al. (2000). Under the Causal Markov assumption, the formulation of D-separation presented by Geiger et al. (1990) is proven to be complete for the restricted setting where determinism arises only due to *functionally determined* relations, defined recursively as variables that are fully determined by their parents, or more precisely:

Definition 2 *A variable X is* functionally determined *by a set $\boldsymbol{W}$ for a given DAG $\mathcal{G}$ if $X \in \boldsymbol{W}$, or all parents of X are functionally determined by $\boldsymbol{W}$.*

We show an example of faithfulness violations due to a functionally determined relation in Figure 1 (left). Following Geiger et al. (1990), we represent the variables resulting from structural equations without noise terms with a double circled node.

In Section 3.8 of their book, Spirtes et al. (2000) extend D-separation to model also some deterministic relations that are due to variables being determined when conditioning on their non-ancestors. In Figure 1 (right) we show an example in which the definition of D-separation from Geiger et al. (1990) fails to capture the faithfulness violation that is due to the deterministic relation between non-ancestors of X, but the extended notion of D-separation by Spirtes et al. (2000) correctly captures it.

Although the version of D-separation by Spirtes et al. (2000) retains completeness for the restricted case of functionally determined relations, it is not proven to be complete in general. Nevertheless, Spirtes et al. (2000) introduce several useful concepts for handling general deterministic relations, so we summarize their findings here, adapting them to our notation. We start with the assumption that we have complete knowledge of all deterministic relations in the system.

Assumption 1 *$\mathcal{D}$ is a complete set of all the deterministic relations among variables, where each entry $\langle\{V_1, \ldots, V_{n-1}\}, V_n\rangle$ in the set indicates that variable V_n is a deterministic function of variables $\{V_1, \ldots, V_{n-1}\}$, but it is not of any strict subset of $\{V_1, \ldots, V_{n-1}\}$.*

For example, in Figure 1 left, $\mathcal{D} = \{\langle\{X\}, Y\rangle\}$, and in Figure 1 right, $\mathcal{D} = \{\langle\{X, S\}, Y\rangle$, $\langle\{Y, X\}, S\rangle$, $\langle\{Y, S\}, X\rangle\}$. This assumption is not as restrictive as it may seem at first, because in practice one can easily reconstruct deterministic relations in the data by using several standard methods. We use $\mathcal{D}$ to define a function that maps a given set of variables $\boldsymbol{W}$ to the set of all variables that are determined by $\boldsymbol{W}$:

Definition 3 *Given a set of variables $\boldsymbol{W}$ and a complete set of deterministic relations $\mathcal{D}$, we define $\text{DET}_{\mathcal{D}}(\boldsymbol{W})$ as the set of variables determined according to $\mathcal{D}$ by (a subset of) $\boldsymbol{W}$. We omit $\mathcal{D}$, using only $\text{DET}(\boldsymbol{W})$, if it is obvious from the context which set we are referring to.*

Note that $\text{DET}_{\mathcal{D}}(\boldsymbol{W})$ trivially includes $\boldsymbol{W}$ itself. Also, any variable with constant value is by definition in $\text{DET}_{\mathcal{D}}(\boldsymbol{W})$ for all $\boldsymbol{W}$ as it is determined by $\emptyset$. Following Spirtes et al. (2000), we can use $\text{DET}()$ to extend d-separation for deterministic relations.

Definition 4 *Given a DAG $\mathcal{G}$, three disjoint sets of variables $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{W}$, and the complete set of deterministic relations $\mathcal{D}$, we define $\boldsymbol{X}$ and $\boldsymbol{Y}$ to be* D-separated *by $\boldsymbol{W}$ w.r.t. $\mathcal{D}$ and $\mathcal{G}$, denoted as $\boldsymbol{X} \perp_D \boldsymbol{Y} \mid \boldsymbol{W}$ $[\mathcal{D}, \mathcal{G}]$, iff for every path π in $\mathcal{G}$ between any $X \in \boldsymbol{X}$ and any $Y \in \boldsymbol{Y}$, at least one of the following holds:*

- *π contains a collider not in $\text{AN}(\boldsymbol{W})$, or*
- *π contains a non-collider[1] in $\text{DET}_{\mathcal{D}}(\boldsymbol{W})$.*

If $\text{DET}_{\mathcal{D}}(\boldsymbol{W}) = \boldsymbol{W}$, D-separation reduces to standard d-separation. We omit $\mathcal{D}$ and $\mathcal{G}$ if it is obvious from the context which set and graph we are referring to.

Under the Causal Markov Assumption, this formulation of D-separation is proven to imply independence (Spirtes et al., 2000). More precisely: $\boldsymbol{X} \perp_D \boldsymbol{Y} \mid \boldsymbol{W}$ $[\mathcal{D}, \mathcal{G}] \implies \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{W}$ $[\mathcal{P}]$ if $\mathcal{P}$ is Markov with respect to $\mathcal{G}$ and $\mathcal{D}$ is the complete set of deterministic relations that hold in $\mathcal{P}$. For the case of functionally determined relations, this version of D-separation is complete, i.e., it completely identifies all additional independences due to functionally determined relations, because in that setting it reduces to the version of D-separation by Geiger et al. (1990), which was shown to be complete by Geiger et al. (1990).

1. Note that we also refer to the end nodes of a path as non-colliders.

3. Related work

Given a set of observational and interventional datasets, most constraint-based methods that handle multiple datasets learn the causal structure from each dataset separately and then merge the learned structures (e.g., Claassen and Heskes, 2010; Hyttinen et al., 2014; Triantafillou and Tsamardinos, 2015; Borboudakis and Tsamardinos, 2016). Some of these methods, e.g., COMBINE (Triantafillou and Tsamardinos, 2015), perform these two steps sequentially, applying a greedy procedure to resolve any potential conflict from the causal structures learnt in the first step. Others, e.g., HEJ (Hyttinen et al., 2014), combine learning and merging in a single procedure, solving potential conflicts by formulating an optimization problem. A recent approach, ETIO (Borboudakis and Tsamardinos, 2016), combines the aspects of the previous approaches by learning and merging in a single procedure, but using a greedy algorithm for resolving conflicts.

Merging causal structures learnt on each dataset separately has several drawbacks with respect to a method that can jointly use all datasets, as for example the score-based method from Eaton and Murphy (2007). First, merging approaches require known targets and cannot learn the targets of the interventions, since this type of information is only available when considering multiple datasets jointly. Moreover, they cannot take advantage of certain interventional datasets, e.g., in the case of a single data point per interventional setting, as for example happens in a popular genomics dataset (Kemmeren et al., 2014). Two other important drawbacks are a loss of statistical power because of the separation into smaller isolated datasets, and, as we will show with some examples in Section 4.1, less identifiable relations with respect to a joint causal inference method.

There is some related work on special cases in which to apply constraint-based methods with mixtures of observational and experimental datasets (e.g., Tian and Pearl, 2001; Eberhardt, 2008; Lagani et al., 2012; Peters et al., 2015; Borboudakis and Tsamardinos, 2016), but the problem has not been systematically discussed and formalized in a general framework yet. In particular, to the best of our knowledge, no existing work addresses learning the intervention targets, possibly jointly with the causal graph, from independence tests.

The approach described by Tian and Pearl (2001) learns the causal structure by combining a standard constraint-based method on observational data with information extracted from changes in the marginal probability of each variable. This information can be extracted from a sequence of (interventional) datasets by comparing each pair of datasets in the sequence, under the assumption that the only difference between the two datasets is a mechanism change on a single known variable. This is a quite restrictive assumption in practice.

Other approaches describe sufficient, although restrictive, conditions under which pooling data does not change the conditional distribution of the variables under consideration. In particular, Eberhardt (2008) describes how naively pooling data from different experimental settings, while discarding the information of which experimental setting a sample was taken, may give wrong results. Thus Eberhardt (2008) proposes a sufficient condition that allows one to pool data for a given independence test when the conditional distribution of the tested variables is the same in all experimental conditions. Lagani et al. (2012) present two other approaches: (i) perform the conditional independence tests separately in each dataset, then define the pooled dependence as a disjunction of the single dependences; (ii) pool experimental conditions that differ only in the value of at most one intervened variable. Both of these approaches describe restrictive conditions in which one can pool datasets, while in this paper we argue that, when done systematically, e.g., as we will show in the next section, one can *always* pool all available datasets.

Other approaches like Invariant Causal Prediction (ICP) (Peters et al., 2015) focus on certain specific combinations of independence tests that are performed jointly on all datasets. ICP is a causal discovery method that looks for invariance across different experimental settings, returning a *conservative subset* of ancestors (or parents, if one assumes there are no latent confounders) for a given target variable Y. The main assumption is that the conditional distribution of Y given its parents does not change in the different interventional settings (in particular, that Y is not directly intervened upon). This assumption is also referred to as invariance or modularity (Pearl, 2009; Spirtes et al., 2000). Since the method searches for patterns that are invariant across different settings, it can safely pool together a subset of settings in a new virtual "experimental" setting to increase the statistical power for settings with few data. On the other hand, as we show with some examples in Section 4.2, the conservativeness of the ICP estimates sometimes significantly reduces the causal information that can be inferred. Like ICP, JCI makes an invariance assumption that allows it to combine different datasets. The invariance assumption made by JCI is that the causal structure is invariant across experimental settings (but model parameters are allowed to change).

4. Joint Causal Inference (JCI)

We propose to model jointly with a single causal graph n observational or experimental datasets $\{D_r\}_{r\in\{1,\dots,n\}}$. We assume that there is a unique underlying causal DAG $G_{\mathcal{X}}$ in all of these datasets, defined over the same set of variables that we call the *system variables*, $\{X_j\}_{j\in\mathcal{X}}$, some of which are possibly hidden.

Each dataset D_r has an associated joint probability distribution $\mathcal{P}_r((X_j)_{j\in\mathcal{X}})$ and represents the d_r data points collected after a set of interventions on possibly unknown intervention targets. In the context of this paper, observational data are simply datasets with an empty set of interventions. We assume each distribution $\mathcal{P}_r$ ($r = 1, \dots, n$) to be Markov and faithful with respect to the causal DAG $G_{\mathcal{X}}$. This assumption precludes certain types of interventions, notably, perfect interventions (Pearl, 2009). On the other hand, it allows for many other types of interventions, e.g., soft interventions (Markowetz et al., 2005), mechanism changes (Tian and Pearl, 2001), fat-hand interventions (Eaton and Murphy, 2007), activity interventions (Mooij and Heskes, 2013), etc., as long as they do not induce new (in)dependences, which can be seen as modifications to the underlying DAG $G_{\mathcal{X}}$.

Using the terminology from Dawid (2002), we call the different distributions in the datasets *regimes*. In related work different names have been used, e.g., *experimental conditions* or *environments* (Mooij and Heskes, 2013; Peters et al., 2015). We introduce two types of dummy variables in the data:

- a **regime variable** R, representing which dataset D_r a data point is from, i.e., $\forall r = 1, \dots, n$, $R = r$ for data from D_r.
- **intervention variables** $\{I_i\}_{i\in\mathcal{I}}$, which are deterministic functions of the regime R. Intervention variables represent the interventions performed in each dataset. In absence of any information on the interventions performed in the datasets, we can use as intervention variables the indicator variables for each of the datasets.

We can now state the main assumption of JCI.

Assumption 2 *We assume that the causal relations between system variables $\{X_j\}_{j\in\mathcal{X}}$ and the introduced dummy variables R and $\{I_i\}_{i\in\mathcal{I}}$ can be represented as an acyclic Structural Causal*

D_1 (20°, no drug):

X_1	X_2	X_4
0.1	0.2	0.5
...	...	...

...

D_4 (30°, drug):

X_1	X_2	X_4
2.2	3.4	1.5
...	...	...

R	I_1	I_2	X_1	X_2	X_4
1	20°	0	0.1	0.2	0.5
1	20°	0	...	...	...
2	20°	1	0.1	3.1	0.6
2	20°	1	...	...	...
3	30°	0	2.3	1.3	1.4
3	30°	0	...	...	...
4	30°	1	2.2	3.4	1.5
4	30°	1	...	...	...

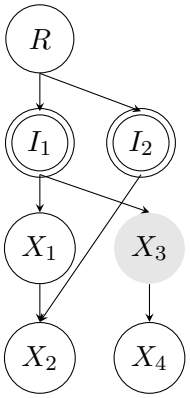


Figure 2: **Prototypical example of JCI setting:** A set of four experimental datasets in raw form (left), in a pooled tabular form with the addition of dummy variables (center) and as a causal DAG $\mathcal{G}$ (right) representing the causal structure of the system variables $X_1, \ldots, X_4$, regime variable R and intervention variables I_1, I_2. The intervention variable I_1 represents the temperature at which each experiment was performed, while I_2 represents the addition of a drug in some of the experiments.

Model (SCM) $\mathcal{M}$ with jointly independent exogenous variables $\{E_k\}_{k\in\mathcal{X}\cup\{R\}}$:

$$\begin{cases} R &= E_R, \\ I_i &= g_i(R), \quad i \in \mathcal{I}, \\ X_j &= f_j(X_{\text{PA}_{\mathcal{X}}(X_j)}, I_{\text{PA}_{\mathcal{I}}(X_j)}, E_j), \quad j \in \mathcal{X}, \end{cases}$$
$$p\big((E_k)_{k\in\mathcal{X}\cup\{R\}}\big) = \prod_{k\in\mathcal{X}\cup\{R\}} p(E_k).$$

Here, $\text{PA}_{\mathcal{X}}(X_j)$ are the system variable parents of X_j, while $\text{PA}_{\mathcal{I}}(X_j)$ denote its intervention parents and E_j is the exogenous parent of X_j. The distribution $\mathcal{P}_r$ corresponding to dataset D_r is given by $p((X_j)_{j\in\mathcal{X}} \mid R = r)$.

We denote the corresponding causal DAG as $\mathcal{G}$. The Causal Markov assumption then holds by construction. We show an example in Figure 2, where we model four datasets with the same underlying causal structure.

The JCI assumptions are applicable when the value of the regime/intervention variables are determined by the experimenter *before* the system variables are measured. More generally, the assumption is that *the system variables cannot cause the regime/intervention variables*. In addition, we assume that the values set by the experimenter are chosen independently of any other possible cause of the system variables. In other words, we assume there to be no confounders between the regime variable and the system variables, or between the intervention variables and the system variables. For the purposes of causal discovery as intended in this paper, there is nothing else that really distinguishes the regime/intervention variables from the system variables: they can both be considered to be random variables, where the distribution of the regime/intervention variables just reflects the empirical distribution of the experimental design chosen by the experimenter. Moreover, there is no distinction between observational and experimental datasets, allowing for several observational datasets, possibly from different contexts.

R	$I_{\text{Akt-Inh}}$	I_{U0126}	I_{ICAM}	$p(R)$
1	0	1	0	0.375
2	1	0	0	0.125
3	0	1	1	0.2
4	1	0	1	0.3

R	I_{drug}	I_{ICAM}	$p(R)$
1	0	0	0.375
2	1	0	0.125
3	0	1	0.2
4	1	1	0.3

Table 1: Example of experimental design matrix with additional deterministic relations beyond the ones allowed in JCI (left) and a reduced version with only allowed deterministic relations (right). In the right version we joined $I_{\text{Akt-Inh}}$ and I_{U0126} in a single intervention variable I_{drug} representing the addition of a single drug (U0126 when the value is 0, Akt-Inh when it is 1).

Intervention variables are functions of the regime variable, and do not have any associated noise. This means that they are *determined* by the regime. We represent these functions as a matrix:

Definition 5 *We define the **experimental design matrix** as the matrix representing the functional relations between R and each intervention variable I_i, and the corresponding probabilities of the regime variable $p(R = r) = \frac{d_r}{\Sigma_{s\in\{1,\ldots,n\}} d_s}$, where d_s is the number of data points in dataset D_s.*

We assume that the intervention variables are complete in the sense that every effect of the regime variable is mediated through an intervention variable. In other words, we assume that there are no direct effects of R on any of the system variables.

In general, other deterministic relations between dummy variables may arise. For example, consider the example in Table 1 left in which $I_{\text{Akt-Inh}}$ represents a drug that was added when the regime is an odd number, while I_{U0126} indicates another drug that was added when the regime is an even number. These two variables determine each other. Even though this is clear from the experimental design matrix, it is not visible in the causal influence diagram.

In this paper, we focus on a special case by allowing only certain types of deterministic relations. We assume that the regime R determines each of the intervention variables $\{I_i\}_{i\in\mathcal{I}}$. Optionally, we allow one additional deterministic relation, namely that all intervention variables $\{I_i\}_{i\in\mathcal{I}}$ together determine the regime R. We assume that there are no other deterministic relations.

Assumption 3 *The deterministic relations that hold in the joint distribution $\mathcal{P}(R, \{I_i\}_{i\in\mathcal{I}}, \{X_j\}_{j\in\mathcal{X}})$ are $\langle\{R\}, I_i\rangle$ for all $i \in \mathcal{I}$, and optionally, $\langle\{I_i\}_{i\in\mathcal{I}}, R\rangle$. No other deterministic relations hold in the joint distribution.*

In practice, one can often "normalize" a system that does not satisfy this assumption. For example, the experimental design matrix in Table 1 left contains also deterministic relations that are not allowed in JCI, but that arise from "redundant" intervention variables. Table 1 right shows how joining two intervention variables can yield a "normalized" system that satisfies the JCI assumptions.

D-separation has been shown to be sound (Spirtes et al., 2000), but was only conjectured to be complete. We prove that in our restricted setting, D-separation is actually complete.

Theorem 6 *Under Assumptions 1–3, D-separation is complete, i.e., it gives* all *conditional independences that are entailed by the assumptions.*

Proof First we consider the case that the regime variable R is *not* determined by the intervention variables. Then, all deterministic relations are functionally determined relations (see Definition 2),

as they correspond to each intervention variable being a function of the regime variable only. The notion of D-separation introduced by Geiger et al. (1990) is proved to be sounds and complete under the Causal Markov assumption for functionally determined relations. If all deterministic relations are functionally determined relations, then their notion of D-separation is equivalent to the one by Spirtes et al. (2000) that we use here. Thus the statement follows.

For the other case, we will show that the D-separations do not change by removing the deterministic relation that the regime variable R is determined by all intervention variables $\{I_i\}_{i\in\mathcal{I}}$. Let $\mathcal{D}$ denote the complete set of deterministic relations according to $\mathcal{M}$ and assume $\langle\{I_i\}_{i\in\mathcal{I}}, R\rangle \in \mathcal{D}$. Let $\mathcal{D}^* = \mathcal{D} \setminus \langle\{I_i\}_{i\in\mathcal{I}}, R\rangle$. Let $\mathcal{G}$ be the DAG associated with $\mathcal{M}$. We claim that for disjoint sets of variables $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{W}$:

$$\boldsymbol{X} \perp_D \boldsymbol{Y} \mid \boldsymbol{W}\ [\mathcal{D}, \mathcal{G}] \iff \boldsymbol{X} \perp_D \boldsymbol{Y} \mid \boldsymbol{W}\ [\mathcal{D}^*, \mathcal{G}].$$

Assume $\boldsymbol{X} \perp_D \boldsymbol{Y} \mid \boldsymbol{W}\ [\mathcal{D}, \mathcal{G}]$ and $\boldsymbol{X} \not\perp_D \boldsymbol{Y} \mid \boldsymbol{W}\ [\mathcal{D}^*, \mathcal{G}]$. This can only happen when $\{I_i\}_{i\in\mathcal{I}} \subseteq \boldsymbol{W}$, because otherwise $\text{DET}_{\mathcal{D}}(\boldsymbol{W}) = \text{DET}_{\mathcal{D}^*}(\boldsymbol{W})$ and then the two D-separations are identical by definition. So there must exist a path π in $\mathcal{G}$ that is D-open w.r.t. $\mathcal{D}^*$ but D-closed w.r.t. $\mathcal{D}$. This means it must contain R as a non-collider. Also, $R \notin \boldsymbol{W}$ otherwise the path would be closed w.r.t. $\mathcal{D}^*$. Since $\boldsymbol{X}$ and $\boldsymbol{Y}$ are disjoint, π must contain at least one other node that is adjacent to R, which must be one of the intervention variables $\{I_i\}_{i\in\mathcal{I}}$. Since the intervention variables can only be non-colliders on π by the JCI assumptions, and $\{I_i\}_{i\in\mathcal{I}} \subseteq \boldsymbol{W}$, they must d-block π. Hence we have arrived at a contradiction: π cannot be D-open w.r.t. $\mathcal{D}^*$.

Moreover, $\boldsymbol{X} \not\perp_D \boldsymbol{Y} \mid \boldsymbol{W}\ [\mathcal{D}, \mathcal{G}]$ and $\boldsymbol{X} \perp_D \boldsymbol{Y} \mid \boldsymbol{W}\ [\mathcal{D}^*, \mathcal{G}]$ cannot happen, because d-separations are the same for $\mathcal{D}$ and $\mathcal{D}^*$, and $\mathcal{D}^* \subset \mathcal{D}$, so there cannot be additional D-separations w.r.t. $\mathcal{D}^*$ compared to the D-separations w.r.t. $\mathcal{D}$. Therefore, removing this particular deterministic relation does not change the D-separation statements, and hence completeness follows also for this case. ■

We conjecture that for the more general case of arbitrary deterministic relations between dummy variables, e.g., Table 1 left, D-separation as defined by Spirtes et al. (2000) is still complete, but we leave the proof for future work.

The completeness of D-separation is important in the JCI context because it motivates a relaxation of the standard Causal Faithfulness assumption. In our setting, the standard assumption is too restrictive, so we relax it to allow for violations due to deterministic relations between the regime and the intervention variables. We define our relaxed version, that we call *D-Faithfulness assumption*, as follows:

Assumption 4 (D-Faithfulness) *For three disjoint sets of variables $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{W}$ and a probability distribution $\mathcal{P}$ that satisfies both the Causal Markov assumption for $\mathcal{G}$ and the set of deterministic relations $\mathcal{D}$, we assume that $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{W}\ [\mathcal{P}] \implies \boldsymbol{X} \perp_D \boldsymbol{Y} \mid \boldsymbol{W}\ [\mathcal{D}, \mathcal{G}]$.*

This assumption, in conjunction with the previous ones, implies that in JCI independences correspond one-to-one with D-separations, which paves the road for constraint-based causal discovery[2]. The completeness of D-separation suggests that this relaxation is "tight", i.e., we only relax the standard Causal Faithfulness assumption to allow for the extra independences that are due to the deterministic relations in Assumption 3, but not any more.

2. A consequence of D-Faithfulness is that intervention variables should be pairwise dependent (also when conditioning on a subset of them). This excludes some experimental design matrices, e.g. a matrix similar to Table 1, but where $P(R = r) = 0.25$ for all r. In practice, we can often alleviate this problem, e.g. by dropping some data points.

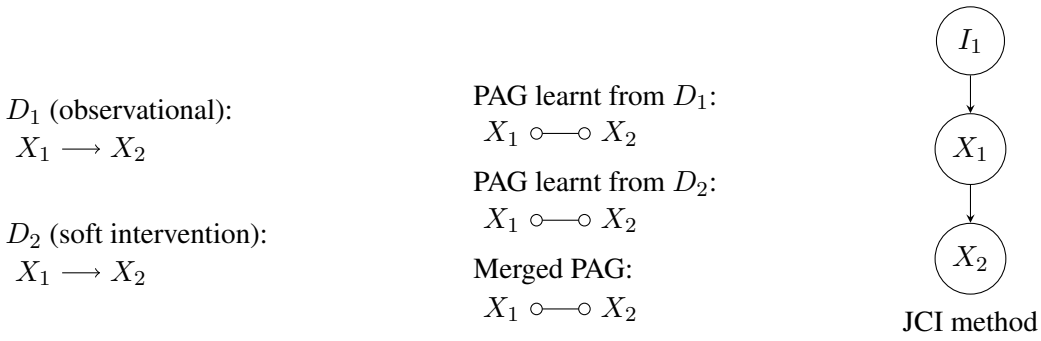


Figure 3: **A simple example in which JCI improves identifiability:** Consider two datasets with the same underlying DAG, one of which has a soft intervention (left). If we learn the causal graphs of X_1 and X_2 in each dataset separately and then merge them, e.g., as described by Triantafillou and Tsamardinos (2015), we cannot learn the causal direction but only that they are dependent (middle). JCI adds more variables and thus conditional independence tests that allow to distinguish the direction (right). Here, since $R = I_1$, w.l.o.g. we represent only I_1. See details in the main text.

Under Assumptions 1–4, we define *Joint Causal Inference (JCI)* as the problem of inferring the causal DAG $\mathcal{G}$ from the distributions $(\mathcal{P}_r)_{r=1,\dots,n}$ or from finite samples $(D_r)_{r=1,\dots,n}$ of those. Moreover, we call any causal discovery method that can solve a JCI instance a *JCI method*. We will show in Section 4.2 that the ideas behind some previous approaches, e.g., (Peters et al., 2015), can be seen as special cases of JCI.

4.1 Joint Causal Inference improves on the identifiability w.r.t. merging learnt structures

As already mentioned, formulating causal inference on multiple datasets as JCI offers several advantages with respect to the approaches in which the causal graphs are learnt separately from each dataset and then merged. One of the advantages is the improved identifiability. In this section, we show a few examples, where, for simplicity, we assume oracle inputs and model only the regime variable (rather than the regime variable and a single intervention variable).

In Figure 3 we show a simple example in which JCI improves identifiability. In the absence of information on the intervention targets, we cannot identify the causal direction between the variables when we learn the structures separately and then combine them. In the same case, a JCI method is able to correctly reconstruct the causal structure by using additional conditional independence tests with the intervention variable, specifically, $I_1 \not\perp\!\!\!\perp X_1$, $I_1 \not\perp\!\!\!\perp X_2$, $I_1 \perp\!\!\!\perp X_2 \,|\, X_1$, $I_1 \not\perp\!\!\!\perp X_1 \,|\, X_2$ and $X_1 \not\perp\!\!\!\perp X_2 \,|\, I_1$. Using the background knowledge from JCI that system variables cannot cause the intervention variable I_1, i.e., $X_1 \not\dashrightarrow I_1 \wedge X_2 \not\dashrightarrow I_1$, and that there are no latent confounders between I_1 and the system variables, we can infer $I_1 \to X_1$, $X_1 \to X_2$ and that there are no confounders with any JCI method supporting direct causal relations. Note that in this example, since there are only two datasets, $R = I_1$, so for simplicity we represent only I_1.

If for each datasets the targets of the intervention are known, then it is possible to retrieve their descendants (and non-descendants) by checking which variables change in each interventional dataset with respect to the observational case. This technique was successfully applied in (Magliacane et al., 2016) to retrieve a list of weighted ancestral relations that could be used as background

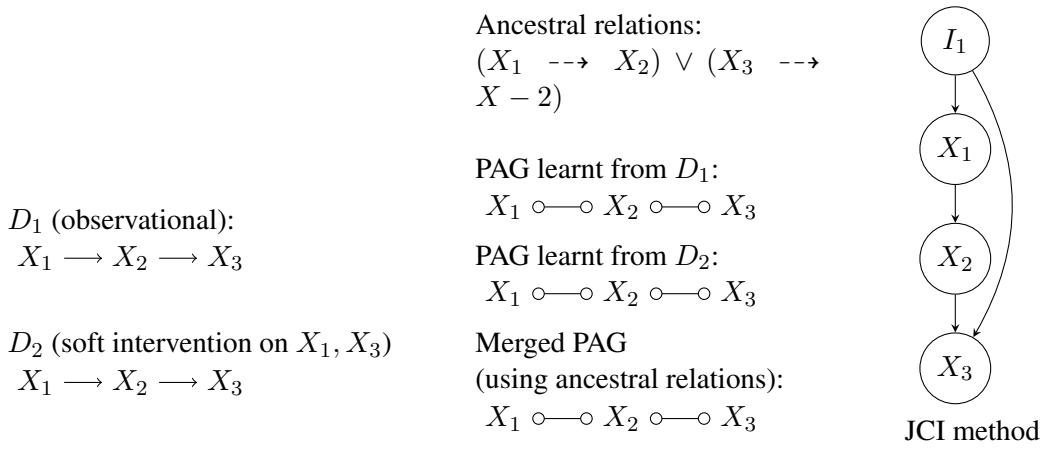


Figure 4: **A more complex example in which JCI improves identifiability:** If we have background knowledge on the intervention targets, e.g., we know that in one of the datasets X_1 and X_3 are intervened upon (left), we can use this information to extract some extra background knowledge in the form of ancestral relations. Merging separately learnt causal structures and this extra knowledge (as done e.g. by Magliacane et al., 2016) is still not enough information to recover the causal structure (middle). Instead, a JCI method can identify the true causal structure, more precisely the ADMG (right). Since $R = I_1$, we can represent only I_1. See details in the main text.

knowledge. For example, if one were to know that X_1 is the intervention target in D_2 in the example in Figure 3, the change in X_2 in D_2 with respect to D_1 would imply that $X_1 \dashrightarrow X_2$.

Although (weighted) ancestral relations help in simple cases as the previous example, in general they cannot reproduce the same results as JCI. An example is given in Figure 4. In this case, knowing that in dataset D_2 the intervention targets are X_1 and X_3, and observing that X_2 changes significantly, allows us to reconstruct that one of these targets causes X_2, which is not enough to reconstruct the causal graph by merging the PAGs learnt from each dataset separately. Instead, a JCI method that supports direct causal relations can take advantage of the additional conditional independence tests with the regime variable and infer the complete DAG:

Proposition 7 *In the example in Figure 4, a JCI method that supports direct causal relations can reconstruct correctly the underlying causal graph, more precisely the acyclic directed mixed graph (ADMG), from oracle independence test results.*

Proof For readability, we provide the proof in the Appendix. ■

4.2 Reformulation of related work as special cases of JCI

Local Causal Discovery (LCD) (Cooper, 1997) is a simple algorithm that searches for variables X, Y, W that satisfy the pattern $W \to X \to Y$, where W is a variable not caused by any other variable under consideration. We can apply LCD to multiple observational and experimental datasets with soft interventions by using R as W, since the regime variable is by assumption not caused by any other variable. Then LCD can be summarized as:

$$(R \not\perp\!\!\!\perp X) \wedge (X \not\perp\!\!\!\perp Y) \wedge (R \perp\!\!\!\perp Y \,|\, X) \implies (X \dashrightarrow Y).$$

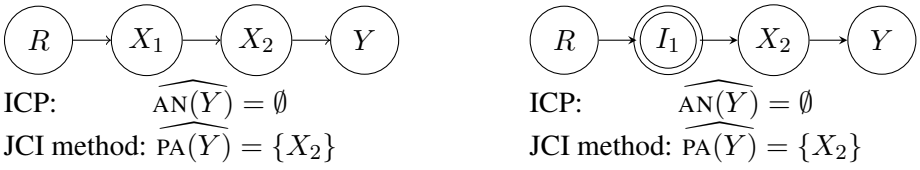


Figure 5: **Two examples in which ICP is overly conservative in the JCI setting:** In the left example, for variable Y ICP finds several sets $\boldsymbol{S}$ that satisfy $R \perp\!\!\!\perp Y \mid \boldsymbol{S}$, e.g., $\{X_1\}$, $\{X_2\}$ and $\{X_1, X_2\}$, but their intersection is $\emptyset$. Instead, a JCI method that supports direct causal relations will correctly infer that the single parent of Y is X_2. In the right example, naively adding intervention variables does not allow to estimate the ancestors of a variable Y, which would otherwise be estimated correctly.

This rule can be seen as a restricted case of the JCI setting, in which we can iteratively pick pairs of variables (X, Y) and apply the above rule to detect a subset of the causal graph.

An extension of this approach is used in Invariant Causal Prediction (ICP) (Peters et al., 2015). ICP also considers the regime variable R (which is called the *discrete environment variable* in that work), but it does not model the intervention variables (see Peters et al., 2015, Appendix). Given a target variable Y that is not directly intervened upon, we can reformulate the main idea behind ICP as the search for the intersection of all the sets $\boldsymbol{S}$ such that $R \perp\!\!\!\perp Y \mid \boldsymbol{S}$:

$$\boldsymbol{S}^* = \bigcap_{\boldsymbol{S}: R \perp\!\!\!\perp Y \mid \boldsymbol{S}} \boldsymbol{S}. \tag{1}$$

In the absence of confounders, $\boldsymbol{S}^*$ is a conservative estimate of a subset of the parents of Y, even when the Causal Faithfulness assumption is violated. If we cannot exclude the presence of confounders, as in the JCI setting, then ICP requires the Causal Faithfulness assumption to provide $\boldsymbol{S}^*$ as an estimate of a subset of the ancestors of Y. We can see this reformulation of ICP as a special case of JCI that extends LCD with a more conservative estimate. In principle, one could easily integrate the conservative estimate (1) in a JCI method to provide more accurate estimates for the top predictions, but we leave this for future work.

On the other hand, depending on the set of interventions in the available datasets, ICP may be overly conservative compared to JCI. Besides the restriction on the variable Y not to be directly intervened upon in any dataset, which is not necessary in JCI, there are some other cases in which ICP provides an overly conservative estimate of the set of ancestors. We show two examples in Figure 5. Specifically, in the left example the estimated set of ancestors for a variable Y that is two hops away from the intervened variable X_1 is empty, while a JCI method that supports direct causal relations can find the correct parent set $\{X_2\}$. Similarly, as shown in the right example, naively adding the intervention variables reduces the applicability of ICP to only estimate ancestors of variables that are directly intervened upon (e.g., X_2). This naive addition would allow ICP to learn the intervention targets, but not the structure of the causal graph.

5. A strategy for extending constraint-based methods for JCI

Joint Causal Inference provides some challenges for current constraint-based methods:

- faithfulness violations due to deterministic relations between the dummy variables,

- the availability of complex background knowledge (i.e., not limited to the presence/absence of edges or ancestral relations) on the dummy variables that can improve structure learning and recover from some of the faithfulness violations (e.g., R can only cause a system variable through an intervention variable).

There is some work on dealing with faithfulness violations in the PC algorithm (Lemeire et al., 2012), but it assumes causal sufficiency (in our context, no hidden variables in $\mathcal{G}$), and cannot handle background knowledge. Logic-based constraint-based algorithms, (e.g., Hyttinen et al., 2014; Triantafillou and Tsamardinos, 2015; Magliacane et al., 2016) can handle complex background knowledge and causal insufficiency, but the existing implementations cannot deal with faithfulness violations due to deterministic relations.

Here we propose a simple but effective strategy for dealing with faithfulness violations due to deterministic relations. We rephrase the constraints of a constraint-based algorithm in terms of d-separations and d-connections, instead of independence test results. At testing time we decide for each independence test result which d-separation or d-connection can be soundly derived from it and provide these d-separations and d-connections as input to the modified constraint-based algorithm.

Before introducing the rules that we use to derive sound d-separations and d-connections from input independence test results, we first summarise the basic properties of conditional independence originally introduced by Dawid (1979), which we will use to prove an intermediate lemma. We follow the notation and ordering from a more recent publication (Constantinou and Dawid, 2015):

Proposition 8 *Let X, Y, Z, W be random variables. We write $W \preceq Y$ to denote that W is a function of Y, or in other words $W = f(Y)$ for a measurable function f. Then the following properties hold:*

1. $X \perp\!\!\!\perp Y \,|\, Z \iff Y \perp\!\!\!\perp X \,|\, Z$,
2. $X \perp\!\!\!\perp Y \,|\, X$,
3. $X \perp\!\!\!\perp Y \,|\, Z$ *and* $W \preceq Y \implies X \perp\!\!\!\perp W \,|\, Z$,
4. $X \perp\!\!\!\perp Y \,|\, Z$ *and* $W \preceq Y \implies X \perp\!\!\!\perp Y \,|\, (Z, W)$,
5. $X \perp\!\!\!\perp Y \,|\, Z \wedge X \perp\!\!\!\perp W \,|\, (Y, Z) \implies X \perp\!\!\!\perp (Y, W) \,|\, Z$.

Lemma 9 *For $X, Y, \boldsymbol{W}$ disjoint (sets) of random variables and random variable $F \preceq \boldsymbol{W}$:*

$$X \perp\!\!\!\perp Y \,|\, \boldsymbol{W} \iff X \perp\!\!\!\perp Y \,|\, \boldsymbol{W} \cup \{F\}.$$

Proof This is a simple consequence of the properties of conditional independence that we reviewed in Proposition 8. We first show one direction of the implication:

$$X \perp\!\!\!\perp (\boldsymbol{W}, Y) \,|\, (\boldsymbol{W}, Y) \text{ (Property 2)} \implies X \perp\!\!\!\perp F \,|\, (\boldsymbol{W}, Y) \text{ (Property 3)}.$$

$$\begin{aligned} X \perp\!\!\!\perp Y \,|\, \boldsymbol{W} \wedge X \perp\!\!\!\perp F \,|\, (\boldsymbol{W}, Y) &\implies X \perp\!\!\!\perp (Y, F) \,|\, \boldsymbol{W} && \text{(Property 5)} \\ &\implies X \perp\!\!\!\perp (Y, F) \,|\, (\boldsymbol{W}, F) && \text{(Property 4)} \\ &\implies X \perp\!\!\!\perp Y \,|\, (\boldsymbol{W}, F) && \text{(Property 3)}. \end{aligned}$$

And for the other direction:

$$X \perp\!\!\!\perp Y \,|\, (\boldsymbol{W}, F) \wedge X \perp\!\!\!\perp F \,|\, \boldsymbol{W} \implies X \perp\!\!\!\perp (F, Y) \,|\, \boldsymbol{W} \text{ (Property 5)}$$
$$\implies X \perp\!\!\!\perp Y \,|\, \boldsymbol{W} \quad \text{(Property 3)}.$$

■

We can now use Lemma 9 to prove a sound conversion from D-separation statements to d-separation statements:

Theorem 10 *For some set of variables $\boldsymbol{W}$, let $\text{DET}(\boldsymbol{W})$ denote the variables determined by (a subset of) $\boldsymbol{W}$ (see Definition 3). Let X, Y be two different variables that are disjoint from $\text{DET}(\boldsymbol{W})$. Under the Causal Markov and D-Faithfulness assumptions, the following holds:*

$$X \perp_D Y \,|\, \boldsymbol{W} \iff X \perp_d Y \,|\, \text{DET}(\boldsymbol{W}).$$

Proof The following equivalences hold:

$$X \perp_D Y \,|\, \boldsymbol{W} \iff X \perp\!\!\!\perp Y \,|\, \boldsymbol{W} \iff X \perp\!\!\!\perp Y \,|\, \text{DET}(\boldsymbol{W}) \text{ by Lemma (9)}$$
$$\iff X \perp_D Y \,|\, \text{DET}(\boldsymbol{W}) \iff X \perp_d Y \,|\, \text{DET}(\boldsymbol{W}).$$

The first equivalence follows from the Causal Markov and D-Faithfulness assumptions, while the second is based on Lemma (9). The third equivalence follows again from the Causal Markov and D-Faithfulness assumptions, while the last one is based on the definition of D-separation, which reduces to d-separation when conditioning on a set $\boldsymbol{Z}$ for which $\text{DET}(\boldsymbol{Z}) = \boldsymbol{Z}$. ■

Using the result from Theorem 10, we can now introduce our strategy for dealing with faithfulness violations due to deterministic relations. First we rephrase a constraint-based algorithm in terms of d-separations and d-connections, which is usually a trivial change, as shown in Section 6. Then we can convert the problem of possibly unfaithful independences to the problem of possibly incomplete input. Specifically, we can derive a subset of sound d-separations and d-connections from independence test results as follows:

Corollary 11 *Let $X, Y, \boldsymbol{W}$ be disjoint (sets) of variables. Let $\text{DET}(\boldsymbol{W})$ be the variables determined by (a subset of) $\boldsymbol{W}$ (see Definition 3). Under the Causal Markov and D-Faithfulness assumptions, the following holds:*

- $X \not\perp\!\!\!\perp Y \,|\, \boldsymbol{W} \implies X \not\perp_d Y \,|\, \boldsymbol{W}$,
- *If* $X, Y \notin \text{DET}(\boldsymbol{W})$: $\quad X \perp\!\!\!\perp Y \,|\, \boldsymbol{W} \implies X \perp_d Y \,|\, \text{DET}(\boldsymbol{W})$.

Proof The first implication follows from the Causal Markov assumption, while the second follows from the Causal Markov and D-Faithfulness assumptions, and Theorem 10. ■

Note that this procedure outputs d-separations only for a subset of independence test results, ignoring independences when X or $Y \in \text{DET}(\boldsymbol{W})$.

The simple strategy in Corollary 11 can be applied to any constraint-based method, providing that it can deal with partial inputs, i.e., missing results for certain independence tests. Logic-based methods (e.g., Hyttinen et al., 2014; Magliacane et al., 2016) can be run out-of-the-box with partial inputs, while other standard algorithms like FCI (Zhang, 2008) would require possibly non-trivial extensions. Anytime FCI (Colombo et al., 2012) allows one to ignore (in)dependences above a certain order, but up to that order they are all required to be available, so that algorithm would also require possibly non-trivial extensions.

6. Ancestral Causal Inference With Determinism (ACID)

We implement the strategy in Corollary 11 in Ancestral Causal Inference with Determinism (ACID) as a determinism-tolerant extension of Ancestral Causal Inference (ACI), a recently introduced logic-based method (Magliacane et al., 2016). Before describing how ACID differs from ACI, we will briefly describe ACI itself.

6.1 Ancestral Causal Inference (ACI)

ACI reconstructs ancestral structures (combinations of "indirect" causal relations), also in the presence of latent variables and statistical errors. Ancestral structures are formally defined as:

Definition 12 *An **ancestral structure** is any relation on the observed variables that satisfies the non-strict partial order axioms:*

$$(\text{reflexivity}) : X \dashrightarrow X, \tag{2}$$

$$(\text{transitivity}) : X \dashrightarrow Y \ \wedge \ Y \dashrightarrow Z \implies X \dashrightarrow Z, \tag{3}$$

$$(\text{antisymmetry}) : X \dashrightarrow Y \wedge Y \dashrightarrow X \implies X = Y. \tag{4}$$

The underlying causal DAG induces a unique ancestral structure on the observed variables: the transitive closure of the direct causal relations (directed edges) in the DAG.

ACI encodes the ancestral structure definition and five other causal reasoning rules:

Lemma 13 *For X, Y, Z, U, $\boldsymbol{W}$ disjoint (sets of) variables:*

1. $(X \perp\!\!\!\perp Y \mid \boldsymbol{W}) \wedge (X \not\dashrightarrow \boldsymbol{W}) \implies X \not\dashrightarrow Y$,
2. $(X \perp\!\!\!\perp Y \mid \boldsymbol{W}) \wedge (X \not\perp\!\!\!\perp Y \mid \boldsymbol{W} \cup Z) \implies (X \not\perp\!\!\!\perp Z \mid \boldsymbol{W}) \wedge (Z \not\dashrightarrow \{X, Y\} \cup \boldsymbol{W})$,
3. $(X \not\perp\!\!\!\perp Y \mid \boldsymbol{W}) \wedge (X \perp\!\!\!\perp Y \mid \boldsymbol{W} \cup Z) \implies (X \not\perp\!\!\!\perp Z \mid \boldsymbol{W}) \wedge (Z \dashrightarrow \{X, Y\} \cup \boldsymbol{W})$,
4. $(X \not\perp\!\!\!\perp Y \mid \boldsymbol{W}) \wedge (X \perp\!\!\!\perp Y \mid \boldsymbol{W} \cup Z) \wedge (X \perp\!\!\!\perp Z \mid \boldsymbol{W} \cup U) \implies X \perp\!\!\!\perp Y \mid \boldsymbol{W} \cup U$,
5. $(Z \not\perp\!\!\!\perp X \mid \boldsymbol{W}) \wedge (Z \not\perp\!\!\!\perp Y \mid \boldsymbol{W}) \wedge (X \perp\!\!\!\perp Y \mid \boldsymbol{W}) \implies X \not\perp\!\!\!\perp Y \mid \boldsymbol{W} \cup Z$.

These rules are shown to be sound assuming the Causal Markov and Causal Faithfulness assumptions. Causal discovery is then reformulated as an optimization problem where a loss function is optimized over possible ancestral structures. Given a list of weighted inputs, e.g. a set of conditional independences weighted by their confidence, the loss function sums the weights of all the inputs that are violated in a candidate ancestral structure. In addition, ACI provides a method for scoring causal predictions, which roughly approximates their marginal probability.

6.2 ACID

Out-of-the-box ACI is not able to deal with the faithfulness violations due to deterministic relations, and thus cannot be used for JCI. Therefore, we propose *Ancestral Causal Inference with Determinism (ACID)*, which extends ACI following the strategy discussed in Section 5. We reformulate the logical rules of ACI in terms of *d-separation*, completely decoupling them from any assumption on the relation between (in)dependences and d-separations/connections, e.g., Causal Faithfulness.

These new rules, that we call the *ACID rules*, are almost identical to the original ACI rules, except that the independences $\perp\!\!\!\perp$ are substituted by d-separations $\perp_d$, and the dependences $\not\perp\!\!\!\perp$ by d-connections $\not\perp_d$, as we show in the following:

Lemma 14 *For X, Y, Z, U, $\boldsymbol{W}$ disjoint (sets of) variables:*

1. $(X \perp_d Y \mid \boldsymbol{W}) \wedge (X \not\dashrightarrow \boldsymbol{W}) \implies X \not\dashrightarrow Y$,
2. $(X \perp_d Y \mid \boldsymbol{W}) \wedge (X \not\perp_d Y \mid \boldsymbol{W} \cup Z) \implies (X \not\perp_d Z \mid \boldsymbol{W}) \wedge (Z \not\dashrightarrow \{X, Y\} \cup \boldsymbol{W})$,
3. $(X \not\perp_d Y \mid \boldsymbol{W}) \wedge (X \perp_d Y \mid \boldsymbol{W} \cup Z) \implies (X \not\perp_d Z \mid \boldsymbol{W}) \wedge (Z \dashrightarrow \{X, Y\} \cup \boldsymbol{W})$,
4. $(X \not\perp_d Y \mid \boldsymbol{W}) \wedge (X \perp_d Y \mid \boldsymbol{W} \cup Z) \wedge (X \perp_d Z \mid \boldsymbol{W} \cup U) \implies X \perp_d Y \mid \boldsymbol{W} \cup U$,
5. $(Z \not\perp_d X \mid \boldsymbol{W}) \wedge (Z \not\perp_d Y \mid \boldsymbol{W}) \wedge (X \perp_d Y \mid \boldsymbol{W}) \implies X \not\perp_d Y \mid \boldsymbol{W} \cup Z$.

Proof The proofs of these rules are slight modifications of the proofs of the ACI rules. For completeness, we provide them in the Appendix. ■

So far, this only changes the interpretation of the implemented rules, but no change of the code is required. What changes are the inputs: only the sound d-separations and d-connections that can be derived with Corollary 11 are used as inputs for ACID. Similarly to other logic-based methods, the ACID rules are sound also with partial inputs (i.e. when some d-separation information may not be available). On the other hand, using partial inputs may reduce the completeness of causal discovery. We consider this a minor issue, since our focus is on prediction accuracy, and ACI is already known not to be complete in the general case, but has nevertheless been shown to obtain state-of-the-art accuracies (Magliacane et al., 2016).

6.3 ACID-JCI

To improve the identifiability and accuracy of the predictions, we also add as background knowledge a series of logical rules describing the causal structure of the regime and intervention variables that apply in the JCI setting:

Lemma 15 *Under the JCI assumptions, for any set of variables $\boldsymbol{W}$:*

1. $\forall i \in \mathcal{I}$ *s.t.* $I_i \notin \boldsymbol{W} : (R \dashrightarrow I_i) \wedge (R \not\perp_d I_i \,|\, \boldsymbol{W})$,
2. $\forall j \in \mathcal{X}$ *s.t.* $X_j \notin \boldsymbol{W} : R \dashrightarrow X_j \implies \exists i \in \mathcal{I}$ *s.t.* $I_i \dashrightarrow X_j$,
3. $\forall j \in \mathcal{X}$ *s.t.* $X_j \notin \boldsymbol{W} : R \perp_d X_j \,|\, \boldsymbol{W} \cup \{I_k\}_{k \in \mathcal{I}}$.
4. $\forall i \in \mathcal{I}, j \in \mathcal{X} : (X_j \not\dashrightarrow R) \wedge (X_j \not\dashrightarrow I_i)$,
5. $\forall j \in \mathcal{X} : R \not\perp_d X_j \implies R \dashrightarrow X_j$,
6. $\forall i \in \mathcal{I}, j \in \mathcal{X} : I_i \not\perp_d X_j \,|\, R \implies I_i \dashrightarrow X_j$,
7. $\forall j, k \in \mathcal{X}$ *s.t.* $j \neq k$ *and* $X_j, X_k \notin \boldsymbol{W} : X_j \perp_d X_k \,|\, \boldsymbol{W} \implies X_j \perp_d X_k \,|\, \boldsymbol{W} \cup R$,
8. $\forall i \in \mathcal{I}, \forall j, k \in \mathcal{X}$ *s.t.* $j \neq k$ *and* $X_j, X_k \notin \boldsymbol{W} : X_j \perp_d X_k \,|\, \boldsymbol{W} \implies X_j \perp_d X_k \,|\, \boldsymbol{W} \cup I_i$.

Proof The propositions follow directly from the JCI assumptions and background knowledge:

1. R causes the intervention variables directly;
2. R cannot cause system variables directly, but only through intervention variables;
3. The intervention variables suffice to block any path between R and any other variable;
4. System variables cannot cause any I_i or R;
5. There are no confounders between R and the system variables;
6. There are no possible confounders between the intervention variables and system variables other than R;
7. Adding R to the separating set cannot open paths;
8. Adding an intervention variable to the separating set cannot open paths.

■

Adding this background knowledge provides a simple means to ruling out several spurious candidate causal structures that do not satisfy the JCI modeling assumptions. The integration of such complex knowledge is one of the main advantages of logic-based methods. We will refer to the combination of ACID with these rules as ACID-JCI.

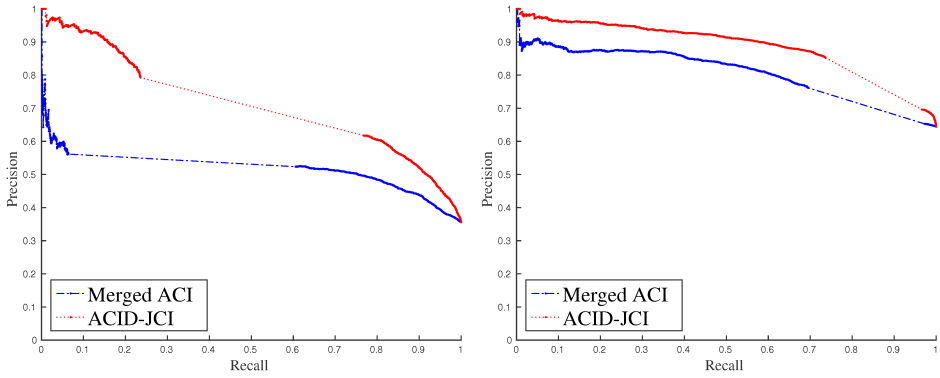
If some of the intervention targets are known, we can also use this information as background knowledge. For example, if we know that the inhibitor $I_{Akt-Inh}$ targets the protein Akt, we can simply add the background knowledge $I_{Akt-Inh} \dashrightarrow Akt$. Given the flexibility of logic-based approaches, we can also include more complex cases. For example, for two mutually exclusive targets A and B for intervention I_1, we can add the background knowledge: $I_1 \dashrightarrow A \vee I_1 \dashrightarrow B$, $I_1 \dashrightarrow A \iff I_1 \not\dashrightarrow B$ and $I_1 \dashrightarrow B \iff I_1 \not\dashrightarrow A$.

7. Evaluation

We evaluate ACID on simulated data in the JCI setting. The simulator builds up on a simulator used in related work (Hyttinen et al., 2014; Magliacane et al., 2016) and implements soft interventions on unknown targets. For each combination of the number of system variables p and interventions i, we generate randomly 1000 linear acyclic models with latent variables and Gaussian noise, and simulate soft interventions on random targets. We then sample $N = 500$ data points for each model, randomly distributed between the i experimental datasets and the observational dataset, perform independence tests and weight the (in)dependence statements using the weighting schemes described by Magliacane et al. (2016).

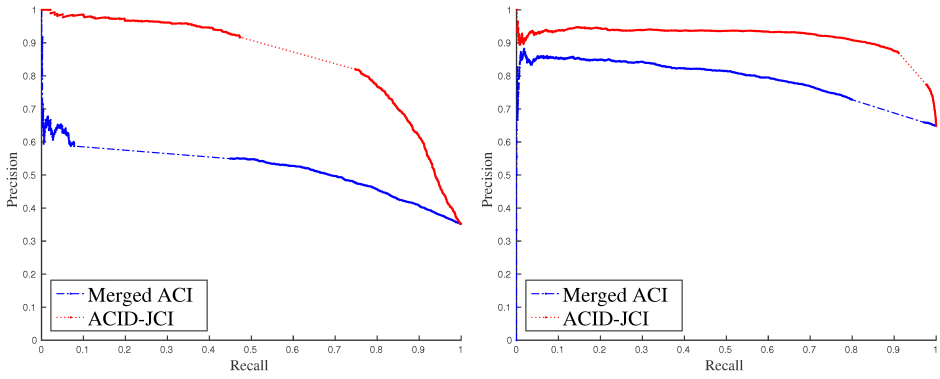
In our setting, we evaluate the causal discovery methods applied to datasets with unknown-target soft interventions. Other existing constraint-based methods (e.g., Hyttinen et al., 2014; Triantafillou and Tsamardinos, 2015) do not apply to this setting as they assume perfect interventions with known targets. Moreover, we wish to evaluate the net effect of using JCI with respect to merging separately learnt causal structures, while factoring out the effect due to different algorithms.

Therefore, we choose to compare the ancestral structure predicted by ACID-JCI with a naive baseline based on ACI, "Merged ACI". In this baseline we merge ancestral structures learnt on each dataset separately with ACI by averaging the scores of the causal predictions over the datasets. As inputs to both algorithms we provide the same weighted independence test results (up to maximum



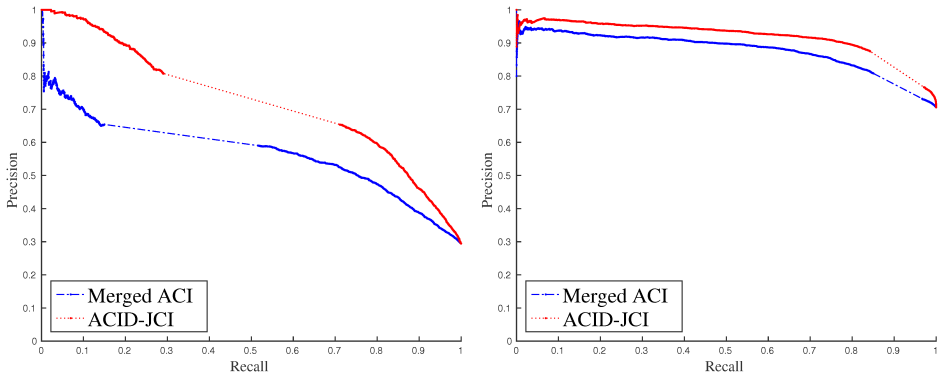
(a) PR ancestral, p = 4, i = 1

(b) PR non-ancestral, p = 4, i = 1



(c) PR ancestral, p = 4, i = 3

(d) PR non-ancestral, p = 4, i = 3

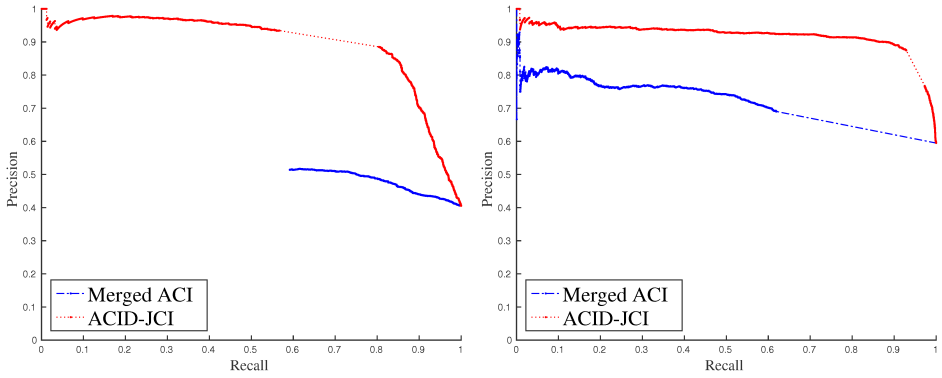


(e) PR ancestral, p = 6, i = 1

(f) PR non-ancestral, p = 6, i = 1

Figure 6: Precision-recall curves on synthetic data for p system variables and i interventions: ancestral predictions (left column) and non-ancestral predictions (right column). Using JCI substantially improves the accuracy.

order), computed with a test based on partial correlation and Fisher's z-transform with significance threshold $\alpha = 0.05$, and the frequentist weighting scheme described by Magliacane et al. (2016).



(a) PR ancestral, p = 3, i = 4 (b) PR non-ancestral, p = 3, i = 4

Figure 7: Precision-recall curves on synthetic data for $p = 3$ system variables and $i = 4$ interventions: ancestral predictions (left column) and non-ancestral predictions (right column). Also in this setting, JCI substantially improves the accuracy.

In Figures 6 and 7 we report the precision and recall (PR) curves for predicting ancestral relations ("indirect" causal relations) and non-ancestral relations (the absence of such a causal relation) for different settings of p system variables and i interventions. We can see from these figures that ACID-JCI improves significantly on the accuracy of the predictions with respect to the baseline. As expected, the more interventional datasets are available, the better the accuracy for ACID-JCI, as we can see in the case of $p = 4$, $i = 1$ vs. $p = 4$, $i = 3$ in Figure 6. When there are only few datasets, e.g., $i = 1$, the performance of both methods improves as the number of system variables p increases, and the gap between the method gets smaller, as we can see in the case of $p = 4$, $i = 1$ and $p = 6$, $i = 1$ in Figure 6. On the other hand, when there are only few variables, e.g., $p = 3$ in Figure 7, ACID-JCI is able to predict correctly several ancestral relations that are not identifiable otherwise. In particular, in the case of $p = 3$ there are no predicted ancestral relations for standard methods like "Merged ACI", which explains why the PR curve for ancestral relations in Figure 7(a) starts at 0.5 recall. Instead, ACID-JCI is able to accurately predict these relations, illustrating that the *Joint* Causal Inference framework not only leads to statistical advantages but also enlarges the set of identifiable ancestral relations compared to methods that deal with each dataset separately.

8. Conclusions and discussion

In this paper, we presented Joint Causal Inference (JCI), a powerful formulation of causal discovery over multiple datasets that has been unexploited so far by constraint-based methods. Current constraint-based methods cannot be applied out-of-the-box to JCI because of faitfulness violations, so we proposed a simple strategy for dealing with this type of faithfulness violations. We implemented this strategy in ACID, a determinism-tolerant extension of a recently proposed causal discovery method, and applied ACID to JCI, showing its benefits in an evaluation on simulated data.

A limitation of JCI is that the assumption of a unique underlying causal DAG precludes certain types of interventions. There are several techniques to extend our formulation of the problem to perfect interventions, or other interventions that induce new independences. For example, given

an observational dataset, we could identify the datasets with perfect interventions by noticing the additional independences, perform causal inference on each of them separately and merge the predictions in a similar way to the methods presented by Hyttinen et al. (2014) and by Triantafillou and Tsamardinos (2015). In future work, we plan to investigate these techniques, as well as techniques to include fat-hand interventions that induce new dependences between the intervention targets.

Moreover, we plan to investigate other possible strategies or extensions to existing algorithms for dealing with faithfulness violations due to deterministic relations. Finally, although very accurate and flexible, logic-based methods as HEJ (Hyttinen et al., 2014) and ACI (Magliacane et al., 2016) are limited in the number of possible variables they can handle. JCI introduces additional variables, reducing their scalability even more. We plan to investigate improvements to the execution times of methods like ACID.

ACKNOWLEDGMENTS

SM, JMM and TC were supported by NWO, the Netherlands Organization for Scientific Research (VIDI grant 639.072.410). SM was also supported by the Dutch programme COMMIT/ under the Data2Semantics project. TC was also supported by NWO grant 612.001.202 (MoCoCaDi), and EU-FP7 grant agreement n.603016 (MATRICS).

REFERENCES

Giorgos Borboudakis and Ioannis Tsamardinos. Towards robust and versatile causal discovery for business applications. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-16)*, pages 1435–1444, San Francisco, CA, USA, 2016.

Tom Claassen and Tom Heskes. Causal discovery in multiple models from different experiments. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems, (NIPS-10)*, pages 415–423, Vancouver, British Columbia, Canada, 2010.

Diego Colombo, Marloes H. Maathuis, Markus Kalisch, and Thomas S. Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, 40(1): 294–321, 2012.

Panayiota Constantinou and Philip Dawid. Extended Conditional Independence and Applications in Causal Inference. *arXiv preprint arXiv:1512.00245*, 2015.

Gregory F. Cooper. A simple constraint-based algorithm for efficiently mining observational databases for causal relationships. *Data Mining and Knowledge Discovery*, 1(2):203–224, 1997.

Gregory F. Cooper and Changwon Yoo. Causal discovery from a mixture of experimental and observational data. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI-99)*, pages 116–125, Stockholm, Sweden, 1999.

Philip Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society B*, 41 (1):1–31, 1979.

Philip Dawid. Influence diagrams for causal modelling and inference. *International Statistical Review*, 70 (2):161–189, 2002.

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-07)*, San Juan, Puerto Rico, 2007.

Frederick Eberhardt. A sufficient condition for pooling data. *Synthese*, 163(3):433–442, 2008.

Doris Entner, Patrik O. Hoyer, and Peter Spirtes. Data-driven covariate selection for nonparametric estimation of causal effects. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics, (AISTATS-13)*, Scottsdale, AZ, USA, 2013.

Dan Geiger, Thomas Verma, and Judea Pearl. Identifying independence in Bayesian networks. *Networks*, 20(5):507–534, 1990.

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.

Antti Hyttinen, Frederick Eberhardt, and Matti Järvisalo. Constraint-based causal discovery: Conflict resolution with answer set programming. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence, (UAI-14)*, pages 340–349, Quebec City, Quebec, Canada, 2014.

Patrick Kemmeren, Katrin Sameith, Loes A.L. van de Pasch, Joris J. Benschop, Tineke L. Lenstra, Thanasis Margaritis, Eoghan O'Duibhir, 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.

Vincenzo Lagani, Ioannis Tsamardinos, and Sofia Triantafillou. Learning from mixture of experimental data: A constraint–based approach. In *Proceedings of Artificial Intelligence: Theories and Applications - 7th Hellenic Conference on AI, (SETN-12)*, Lamia, Greece, 2012.

Jan Lemeire, Stijn Meganck, Francesco Cartella, and Tingting Liu. Conservative independence-based causal structure learning in absence of adjacency faithfulness. *International Journal of Approximate Reasoning*, 53(9):1305–1325, 2012.

Sara Magliacane, Tom Claassen, and Joris M. Mooij. Ancestral causal inference. In *In Proceedings of Advances in Neural Information Processing Systems, (NIPS-16)*, pages 4466–4474, Barcelona, Spain, 2016.

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-05)*, Bridgetown, Barbados, 2005.

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-13)*, pages 431–439, 2013.

Judea Pearl. *Causality: models, reasoning and inference*. Cambridge University Press, 2009.

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, 2015.

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

Jin Tian and Judea Pearl. Causal discovery from changes. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence, (UAI-01)*, Seattle, Washington, USA, 2001.

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.

Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873–1896, 2008.

Appendix A. Proofs

Proposition 16 *In the example in Figure 4, a JCI method that supports direct causal relations can reconstruct correctly the underlying causal graph, more precisely the acyclic directed mixed graph (ADMG), from oracle independence test results.*

Proof In this proof, we proceed in two steps: first we reconstruct the ancestral relations, then show that a subset of these relations are actually direct causal relations and finally we show the absence of confounders. Given their simplicity, for the first step we use the rules from ACI that we will show in detail in Lemma 13 in Section 6.

From $I_1 \not\perp\!\!\!\perp X_1$, the background knowledge that there are no latent confounders between I_1 and the system variables and that $X_1 \not\dashrightarrow I_1$, we can infer that $I_1 \dashrightarrow X_1$. Reasoning in a similar way, we can also infer that $I_1 \dashrightarrow X_2$ and $I_1 \dashrightarrow X_3$. From $I_1 \not\perp\!\!\!\perp X_2$ and $I_1 \perp\!\!\!\perp X_2 \,|\, X_1$ we can use rule (3) in Lemma 13, which implies that $X_1 \dashrightarrow \{I_1, X_2\}$. Given the background knowledge $X_1 \not\dashrightarrow I_1$ we can infer that $X_1 \dashrightarrow X_2$. Similarly, from $X_1 \not\perp\!\!\!\perp X_3 \,|\, I_1$, $X_1 \perp\!\!\!\perp X_3 \,|\, \{I_1, X_2\}$ we can infer that $X_2 \dashrightarrow \{I_1, X_1, X_3\}$. Since $X_1 \not\dashrightarrow I_1$ and $X_2 \not\dashrightarrow X_1$ (from the acyclicity of ancestral relations), then this must imply that $X_2 \dashrightarrow X_3$. Since ancestral relations are transitive, from $X_1 \dashrightarrow X_2$ and $X_2 \dashrightarrow X_3$ we get that $X_1 \dashrightarrow X_3$. Moreover, because of the acyclicity of ancestral relations, these statements imply that $X_2 \not\dashrightarrow X_1$, $X_3 \not\dashrightarrow X_1$, $X_3 \not\dashrightarrow X_2$.

We now go through the ancestral relations we found and check which of them are direct causal relations:

- $X_1 \dashrightarrow X_2$: since $I_1 \dashrightarrow X_1$ and $X_2 \dashrightarrow X_3$, X_1 cannot cause X_2 indirectly through I_1 or X_3 (from the acyclicity of causal relations), so $X_1 \to X_2$ directly;
- $X_2 \dashrightarrow X_3$: since $I_1 \dashrightarrow X_2$ and $X_1 \dashrightarrow X_2$, X_2 cannot cause X_3 indirectly through I_1 or X_1 (acyclicity), so $X_2 \to X_3$ directly;
- $X_1 \dashrightarrow X_3$: this relation is not direct, because $X_1 \perp\!\!\!\perp X_3 \,|\, \{X_2, I_1\}$, so they are not adjacent in the causal graph;
- $I_1 \dashrightarrow X_1$: since $X_1 \dashrightarrow X_2$ and $X_1 \dashrightarrow X_3$, I_1 cannot cause X_1 indirectly through X_2 or X_3 (from the acyclicity of causal relations), so $I_1 \to X_1$ directly;
- $I_1 \dashrightarrow X_2$: this relation is not direct, because $I_1 \perp\!\!\!\perp X_2 \,|\, X_1$, so they are not adjacent in the causal graph;
- $I_1 \dashrightarrow X_3$: given all the previous causal relations, the background knowledge that there are no confounders between I_1 and the other variables, $X_1 \not\perp\!\!\!\perp X_3 \,|\, X_2$ and $X_1 \perp\!\!\!\perp X_3 \,|\, \{X_2, I_1\}$ the only possible graph includes a direct causal relation $I_1 \to X_3$.

Finally, we show the absence of confounders in the system:

- I_1, X_1 and I_1, X_2 and I_1, X_3 are unconfounded (JCI background knowledge);
- X_1, X_3 are unconfounded, because $X_1 \perp\!\!\!\perp X_3 \,|\, \{X_2, I_1\}$;
- X_1, X_2 are unconfounded, because $I_1 \perp\!\!\!\perp X_2 \,|\, X_1$;
- X_2, X_3 are unconfounded, because $X_1 \perp\!\!\!\perp X_3 \,|\, \{X_2, I_1\}$.

■

For completeness we provide the proofs for the ACID rules, which are just slight modifications of the proofs for the ACI (Magliacane et al., 2016) rules. These proofs were already implicitly based on d-separation concepts, with two additional implicit steps that were based on the Causal Markov and Faithfulness Assumptions: one in which each conditional independence implies a d-separation, and another one in which each d-separation implies a conditional independence. In this reformulation we can make these steps explicit and decouple them from the rules.

Lemma 17 *For X, Y, Z, U, $\boldsymbol{W}$ disjoint (sets of) variables of a DAG (or ADMG) $\mathcal{G}$:*

1. $(X \perp_d Y \mid \boldsymbol{W}) \wedge (X \not\dashrightarrow \boldsymbol{W}) \implies X \not\dashrightarrow Y$,
2. $(X \perp_d Y \mid \boldsymbol{W}) \wedge (X \not\perp_d Y \mid \boldsymbol{W} \cup Z) \implies (X \not\perp_d Z \mid \boldsymbol{W}) \wedge (Z \not\dashrightarrow \{X, Y\} \cup \boldsymbol{W})$,
3. $(X \not\perp_d Y \mid \boldsymbol{W}) \wedge (X \perp_d Y \mid \boldsymbol{W} \cup Z) \implies (X \not\perp_d Z \mid \boldsymbol{W}) \wedge (Z \dashrightarrow \{X, Y\} \cup \boldsymbol{W})$,
4. $(X \not\perp_d Y \mid \boldsymbol{W}) \wedge (X \perp_d Y \mid \boldsymbol{W} \cup Z) \wedge (X \perp_d Z \mid \boldsymbol{W} \cup U) \implies X \perp_d Y \mid \boldsymbol{W} \cup U$,
5. $(Z \not\perp_d X \mid \boldsymbol{W}) \wedge (Z \not\perp_d Y \mid \boldsymbol{W}) \wedge (X \perp_d Y \mid \boldsymbol{W}) \implies X \not\perp_d Y \mid \boldsymbol{W} \cup Z$.

Proof

1. This is a strengthened version of rule $\mathcal{R}2$(i) in (Entner et al., 2013): note that the additional assumptions made there ($Y \not\dashrightarrow \boldsymbol{W}$, $Y \not\dashrightarrow X$) are redundant and not actually used in their proof. For completeness, we give the proof here. If $X \dashrightarrow Y$, then there is a directed path from X to Y in $\mathcal{G}$. As all paths between X and Y in $\mathcal{G}$ are blocked by $\mathbf{W}$, the directed path from X to Y must contain a node $W \in \boldsymbol{W}$. Hence $X \dashrightarrow W$, a contradiction with $X \not\dashrightarrow \mathbf{W}$.

2. If $(X \perp_d Y \mid \boldsymbol{W}) \wedge (X \not\perp_d Y \mid \boldsymbol{W} \cup Z)$ then there exists a path π in $\mathcal{G}$ between X and Y such that each noncollider on π is not in $\boldsymbol{W} \cup \{Z\}$, every collider on π is ancestor of $\boldsymbol{W} \cup \{Z\}$, and there exists a collider on π that is ancestor of Z but not of $\boldsymbol{W}$. Let C be the collider on π closest to X that is ancestor of Z but not of $\boldsymbol{W}$. Note that

   (a) The path $X \cdots C \to \cdots \to Z$ is d-connected given $\boldsymbol{W}$.

   (b) $Z \not\dashrightarrow \boldsymbol{W}$ (because otherwise $C \dashrightarrow Z \dashrightarrow \boldsymbol{W}$, a contradiction).

   (c) $Z \not\dashrightarrow Y$ (because otherwise the path $X \cdots C \to \cdots \to Z \to \cdots \to Y$ would be d-connected given $\boldsymbol{W}$, a contradiction).

   Hence we conclude that $X \not\perp_d Z \mid \boldsymbol{W}$, $Z \not\dashrightarrow \boldsymbol{W}$, $Z \not\dashrightarrow Y$, and by symmetry also $Z \not\dashrightarrow X$.

3. Suppose $(X \not\perp_d Y \mid \boldsymbol{W}) \wedge (X \perp_d Y \mid \boldsymbol{W} \cup Z)$. Then there exists a path π in $\mathcal{G}$ between X and Y, such that each noncollider on π is not in $\boldsymbol{W}$, each collider on π is an ancestor of $\boldsymbol{W}$, and Z is a noncollider on π. Note that

   (a) The subpath $X \ldots Z$ must be d-connected given $\boldsymbol{W}$.

   (b) Z has at least one outgoing edge on π. Follow this edge further along π until reaching either X, Y, or the first collider. When a collider is reached, follow the directed path to $\boldsymbol{W}$. Hence there is a directed path from Z to X or Y or to $\boldsymbol{W}$, i.e., $Z \dashrightarrow \{X, Y\} \cup \boldsymbol{W}$.

4. If in addition, $X \perp_d Z \mid \boldsymbol{W} \cup U$, then U must be a noncollider on the subpath $X \dots Z$. Therefore, $X \perp_d Y \mid \boldsymbol{W} \cup U$.

5. Assume that $Z \not\perp_d X \mid \boldsymbol{W}$ and $Z \not\perp_d Y \mid \boldsymbol{W}$. Then there must be paths π between Z and X and ρ between Z and Y in $\mathcal{G}$ such that each noncollider is not in $\boldsymbol{W}$ and each collider is ancestor of $\boldsymbol{W}$. Let U be the node on π closest to X that is also on ρ (this could be Z). Then we have a path $X \cdots U \cdots Y$ such that each collider (except U) is ancestor of $\boldsymbol{W}$ and each noncollider (except U) is not in $\boldsymbol{W}$. This path must be blocked given $\boldsymbol{W}$ as $X \perp_d Y \mid \boldsymbol{W}$. If U would be a noncollider on this path, it would need to be in $\boldsymbol{W}$ in order to block it; however, it must then also be a noncollider on π or ρ and hence cannot be in $\boldsymbol{W}$. Therefore, U must be a collider on this path and cannot be ancestor of $\boldsymbol{W}$. We have to show that U is ancestor of Z. If U were a collider on π or ρ, it would be ancestor of $\boldsymbol{W}$, a contradiction. Hence U must have an outgoing arrow pointing towards Z on π and ρ. If we encounter a collider following the directed edges, we get a contradiction, as that collider, and hence U, would be ancestor of $\boldsymbol{W}$. Hence U is ancestor of Z, and therefore, $X \not\perp_d Y \mid \boldsymbol{W} \cup Z$.

■