Constraint-Based Causal Discovery using Partial Ancestral Graphs in the presence of Cycles

Joris M. Mooij
Korteweg-de Vries Institute
University of Amsterdam
Amsterdam, The Netherlands

Tom Claassen
Institute for Computing and Information Sciences
Radboud University Nijmegen
Nijmegen, The Netherlands

Abstract

While feedback loops are known to play important roles in many complex systems, their existence is ignored in a large part of the causal discovery literature, as systems are typically assumed to be acyclic from the outset. When applying causal discovery algorithms designed for the acyclic setting on data generated by a system that involves feedback, one would not expect to obtain correct results. In this work, we show that—surprisingly—the output of the Fast Causal Inference (FCI) algorithm is correct if it is applied to observational data generated by a system that involves feedback. More specifically, we prove that for observational data generated by a simple and σ-faithful Structural Causal Model (SCM), FCI is sound and complete, and can be used to consistently estimate (i) the presence and absence of causal relations, (ii) the presence and absence of direct causal relations, (iii) the absence of confounders, and (iv) the absence of specific cycles in the causal graph of the SCM. We extend these results to constraint-based causal discovery algorithms that exploit certain forms of background knowledge, including the causally sufficient setting (e.g., the PC algorithm) and the Joint Causal Inference setting (e.g., the FCI-JCI algorithm).

1 INTRODUCTION

Causal discovery, i.e., establishing the presence or absence of causal relationships between observed variables, is an important activity in many scientific disciplines. Typical approaches to causal discovery from observational data are either score-based, or constraint-based (or a combination of the two). The more generally applicable constraint-based approach, which we focus on in this work, is based on exploiting information in conditional independences in the observed data to draw conclusions about the possible underlying causal structure.

Although many systems of interest in various application domains involve feedback loops or other types of cyclic causal relationships (for example, in economical, biological, chemical, physical, control and climatological systems), most of the existing literature on causal discovery from observational data ignores this and assumes from the outset that the underlying causal system is acyclic. Nonetheless, several algorithms have been developed specifically for the cyclic setting. For example, quite some work has been done for linear systems (e.g., Richardson and Spirtes, 1999; Lacerda et al., 2008; Hyttinen et al., 2010, 2012; Rothenhäusler et al., 2015).

More generally applicable are causal discovery algorithms that exploit conditional independence constraints, without assuming certain restrictions on the parameterizations of the causal models (such as linearity). Pioneering work in this area was done by Richardson (1996b), resulting in the CCD algorithm, the first constraint-based causal discovery algorithm shown to be applicable in a cyclic setting (see also Richardson, 1996a; Richardson and Spirtes, 1999). It was shown to be sound under the assumptions of causal sufficiency, the d-separation Markov property, and d-faithfulness. More recently, other algorithms that are sound under these assumptions (except for the requirement of causal sufficiency) were proposed (Hyttinen et al., 2014; Strobl, 2018).

However, it was already noted by Spirtes (1994, 1995) that the d-separation Markov property assumption can be too strong in general, and he proposed an alternative criterion, making use of the so-called "collapsed graph" construction. More recently, an alternative formulation in terms of the *σ-separation* criterion was introduced, and the corresponding Markov property was shown to

Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), PMLR volume 124, 2020.

hold in a very general setting (Forré and Mooij, 2017). Whereas the Markov property based on σ-separation applies under mild assumptions, the stronger Markov property based on d-separation is limited to more specific settings (e.g., continuous variables with linear relations, or discrete variables, or the acyclic case) (Forré and Mooij, 2017). As discussed in (Forré and Mooij, 2017; Bongers et al., 2020), the σ-separation Markov property seems appropriate for a wide class of cyclic structural causal models with non-linear functional relationships between non-discrete variables, for example structural causal models corresponding to the equilibrium states of dynamical systems governed by random differential equations (Bongers and Mooij, 2018).

Apart from a Markov property, constraint-based causal discovery algorithms need to make some type of faithfulness assumption. A natural extension of the common faithfulness assumption used in the acyclic setting is obtained by replacing d-separation by σ-separation, that we refer to as *σ-faithfulness*. Forré and Mooij (2018) proposed a constraint-based causal discovery algorithm that is sound and complete, assuming the σ-separation Markov property in combination with the σ-faithfulness assumption. However, their algorithm is limited in practice to about 5–7 variables because of the combinatorial explosion in the number of possible causal graphs with increasing number of variables. Interestingly, under the additional assumption of causal sufficiency, the CCD algorithm is also sound under these assumptions (as already noted in Section 4.5 of Richardson, 1996b). Other causal discovery algorithms (LCD (Cooper, 1997), ICP (Peters et al., 2016) and Y-structures (Mani, 2006)), all originally designed for the acyclic setting, have been shown to be sound also in the σ-separation setting (Mooij et al., 2020). The most general scenario (under the additional assumption of causal sufficiency, however) is addressed by the NL-CCD algorithm (Chapter 4 in Richardson, 1996b), which was shown to be sound under the assumptions of the σ-separation Markov property together with the (weaker) d-faithfulness assumption.

One of the classic algorithms for constraint-based causal discovery is the Fast Causal Inference (FCI) algorithm (Spirtes et al., 1995, 1999; Zhang, 2008b). It was designed for the acyclic case, assuming the d-separation Markov property in combination with the d-faithfulness assumption. Recently, it was observed that when run on data generated by cyclic causal models, the accuracy of FCI is actually comparable to its accuracy in the strictly acyclic setting (Figures 25, 26, 29, 31, 32 in Mooij et al., 2020). This is surprising, as it is commonly believed that the application domain of FCI is limited to acyclic causal systems, and one would expect such serious model misspecification to result in glaringly incorrect results.

In this work, we show that when FCI is applied on data from a cyclic causal system that satisfies the σ-separation Markov property and is σ-faithful, its output is still sound and complete. Furthermore, we derive criteria for how to read off various features from the partial ancestral graph output by FCI (specifically, the absence or presence of ancestral relations, direct relations, cyclic relations and confounders). This provides a practical causal discovery algorithm for that setting that is able to handle hundreds or even thousands of variables as long as the underlying causal model is sparse enough, and that is also applicable in the presence of latent confounders. It thus forms a significant improvement over the previous state-of-the-art in causal discovery for the σ-separation setting.

The results we derive in this work are not limited to FCI, but apply to any constraint-based causal discovery algorithm that solves the same task as FCI does, i.e., that estimates the directed partial ancestral graph from conditional independences in the data, e.g., FCI+ (Claassen et al., 2013) and CFCI (Colombo et al., 2012). Our results therefore make constraint-based causal discovery in the presence of cycles as practical as it is in the acyclic case, without requiring any modifications of the algorithms. Our work also provides the first characterization of the σ-Markov equivalence class of directed mixed graphs. We extend our results to variants of algorithms that exploit certain background knowledge, for example, causal sufficiency (e.g., the PC algorithm, Spirtes et al., 2000) or the Joint Causal Inference framework (e.g., the FCI-JCI algorithm, Mooij et al., 2020). For simplicity, we assume no selection bias in this work, but we expect that our results can be extended to allow for that as well.

2 PRELIMINARIES

In Section A (Supplementary Material), we introduce our notation and terminology and provide the reader with a summary of the necessary definitions and results from the graphical causal modeling and discovery literature. For more details, we refer the reader to the literature (Pearl, 2009; Spirtes et al., 2000; Richardson and Spirtes, 2002; Zhang, 2006, 2008b,a; Bongers et al., 2020; Forré and Mooij, 2017). Here, we only give a short high-level overview of the key notions because of space constraints.

There exists a variety of graphical representations of causal models. Most popular are *directed acyclic graphs (DAGs)*, presumably because of their simplicity. DAGs are appropriate under the assumptions of causal sufficiency (i.e., there are no latent common causes of the observed variables), acyclicity (absence of feedback loops) and no selection bias (i.e., there is no implicit conditioning on a common effect of the observed variables). DAGs have many convenient properties, amongst which

a Markov property (which has different equivalent formulations, the most prominent one being in terms of the notion of d-separation) and a simple causal interpretation. A more general class of graphs are *acyclic directed mixed graphs (ADMGs)*. These make use of additional bidirected edges to represent latent confounding, and have a similarly convenient Markov property (sometimes referred to as m-separation) and causal interpretation. When also dropping the assumption of acyclicity (thereby allowing for feedback), one can make use of the more general class of *directed mixed graphs (DMGs)*. These graphs can be naturally associated with (possibly cyclic) structural causal models (SCMs) and can represent feedback loops. The corresponding Markov properties and causal interpretation are more subtle (Bongers et al., 2020) than in the acyclic case. Cyclic SCMs can be used, e.g., to describe the causal semantics of the equilibrium states of dynamical systems governed by random differential equations (Bongers and Mooij, 2018).

In this work, we will restrict ourselves to the subclass of *simple* SCMs, i.e., those SCMs for which any subset of the structural equations has a unique solution for the corresponding endogenous variables in terms of the other variables appearing in these equations. Simple SCMs admit (sufficiently weak) cyclic interactions but retain many of the convenient properties of acyclic SCMs (Bongers et al., 2020). They are a special case of modular SCMs (Forré and Mooij, 2017). In particular, they satisfy the σ-separation Markov property and their graphs have an intuitive causal interpretation.[1]

For acyclic constraint-based causal discovery, ADMGs provide a more fine-grained representation than necessary, because one can only recover the Markov equivalence class of ADMGs from conditional independences in observational data. A less expressive class of graphs, *maximal ancestral graphs (MAGs)*, was introduced by Richardson and Spirtes (2002). Each ADMG induces a MAG and each MAG represents a set of ADMGs. The mapping from ADMG to MAG preserves the d-separations and the (non-)ancestral relations. Contrary to ADMGs, MAGs have at most a single edge connecting any pair of distinct variables. One of the key properties that distinguishes MAGs from ADMGs is that Markov-equivalent MAGs have the same adjacencies. In addition to being able to handle latent variables, MAGs can also represent implicit conditioning on a subset of the variables, making use of undirected edges. Therefore, they can be used to represent both latent variables and selection bias.

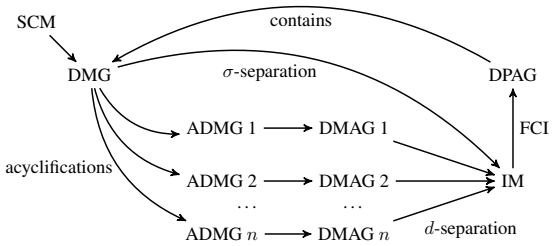

Figure 1: Relations between various representations.

It is often convenient when performing causal reasoning or discovery to be able to represent a set of hypothetical MAGs in a compact way. For these reasons, *partial ancestral graphs (PAGs)* were introduced (Zhang, 2006).[2] The usual way to think about a PAG is as an object that represents a set of MAGs. The (Augmented) Fast Causal Inference (FCI) algorithm (Spirtes et al., 1995, 1999; Zhang, 2008b) takes as input the conditional independences that hold in the data (assumed to be d-Markov and d-faithful w.r.t. a "true" ADMG), and outputs a PAG. As shown in seminal work (Spirtes et al., 1995, 1999; Ali et al., 2005; Zhang, 2008b), the FCI algorithm is sound and complete, and the PAG output by FCI represents the Markov equivalence class of the true ADMG.

In this work, we will for simplicity assume no selection bias. This means that we can restrict ourselves to MAGs without undirected edges, which we refer to as *directed MAGs (DMAGs)*, and PAGs without undirected or circle-tail edges, which we refer to as *directed PAGs* (DPAGs). Almost all proofs will be deferred to Section C (Supplementary Material) because of space constraints.

3 EXTENSIONS TO THE CYCLIC SETTING

The theory of MAGs and PAGs is rather intricate. A natural question is how this theory can be extended when the

[1]The σ-separation criterion is very similar to the d-separation criterion, with the only difference being that σ-separation has as an additional condition for a non-collider to block a path that it has to point to a node in a different strongly connected component. Two nodes in a DMG are said to be in the same strongly connected component if and only if they are both ancestor of each other.

[2]PAGs were originally introduced by Richardson (1996b) in order to represent the output of the CCD algorithm. It was conjectured by Richardson that PAGs could also be used to represent the output of the FCI algorithm, which was originally formulated in terms of Partially Oriented Inducing Path Graphs (POIPGs). This conjecture was proved subsequently by Spirtes. Richardson (p. 102, 1996b) notes: "It is an open question whether or not the set of symbols is sufficiently rich to allow us to represent the class of cyclic graphs with latent variables." In the present work we turned full circle by reinterpreting PAGs as representing properties of DMGs, and have thereby answered this question affirmatively.

assumption of acyclicity is dropped. This does not seem to be straightforward at first sight. An obvious approach would be to generalize the notion of MAGs by adding edge types that represent cycles. However, it would probably require a lot of effort to rederive and reformulate the known results about MAGs and PAGs in this more general setting. In this work, we take another approach: we represent a (possibly cyclic) DMG directly by a DPAG. In order to make this idea precise, we first need to extend the notion of inducing path to the cyclic setting. Our strategy is illustrated in Figure 1.

3.1 INDUCING PATHS

We propose the following generalization of the notion of inducing path (Definition 9) to the σ-separation setting:

Definition 1 *Let $\mathcal{G} = \langle \mathcal{V}, \mathcal{E}, \mathcal{F} \rangle$ be directed mixed graph (DMG). An* inducing path (walk) between two nodes $i, j \in \mathcal{V}$ *is a path (walk) in $\mathcal{G}$ between i and j on which every collider is in $\text{AN}_{\mathcal{G}}(\{i, j\})$, and each non-collider on the path (walk), except i and j, only has outgoing directed edges to neighboring nodes on the path (walk) that lie in the same strongly connected component of $\mathcal{G}$.*

This is indeed the proper generalization, since it has the following property.

Proposition 1 *Let $\mathcal{G} = \langle \mathcal{V}, \mathcal{E}, \mathcal{F} \rangle$ be a DMG and i, j two distinct vertices in $\mathcal{G}$. Then the following are equivalent:*

(i) There is an inducing path in $\mathcal{G}$ between i and j;
(ii) There is an inducing walk in $\mathcal{G}$ between i and j;
(iii) $i \not\perp^{\sigma}_{\mathcal{G}} j \mid Z$ for all $Z \subseteq \mathcal{V} \setminus \{i, j\}$.

In words: there is an inducing walk (or path) between two nodes in a DMG if and only if the two nodes cannot be σ-separated by any subset of nodes that does not contain either of the two nodes.

3.2 REPRESENTING DMGs BY DPAGs

The following definition forms the key to our approach.

Definition 2 *Let $\mathcal{P}$ be a DPAG and $\mathcal{G}$ a DMG, both with vertex set $\mathcal{V}$. We say that $\mathcal{P}$* contains *$\mathcal{G}$ if all of the following hold:*

(i) two vertices i, j are adjacent in $\mathcal{P}$ if and only if there is an inducing path between i, j in $\mathcal{G}$;
(ii) if $i \ast\!\!\rightarrow j$ in $\mathcal{P}$ (i.e., $i \rightarrow j$ in $\mathcal{P}$ or $i \circ\!\!\rightarrow j$ in $\mathcal{P}$ or $i \leftrightarrow j$ in $\mathcal{P}$), then $j \notin \text{AN}_{\mathcal{G}}(i)$;
(iii) if $i \rightarrow j$ in $\mathcal{P}$ then $i \in \text{AN}_{\mathcal{G}}(j)$.

It is only a slight variation on how PAGs are traditionally interpreted, and agrees with the traditional (acyclic) interpretation when restricting the DMGs to be acyclic.

3.3 ACYCLIFICATIONS

Inspired by the "collapsed graph" construction of Spirtes (1994, 1995), Forré and Mooij (2017) introduced a notion of *acyclification* for a class of graphical causal models termed HEDGes, but the same concept can be defined for DMGs, which we will do here.

Definition 3 *Given a DMG $\mathcal{G} = \langle \mathcal{V}, \mathcal{E}, \mathcal{F} \rangle$. An* acyclification *of $\mathcal{G}$ is an ADMG $\mathcal{G}' = \langle \mathcal{V}, \mathcal{E}', \mathcal{F}' \rangle$ with*

(i) the same nodes $\mathcal{V}$;
(ii) for any pair of nodes $\{i, j\}$ such that $i \notin \text{SC}_{\mathcal{G}}(j)$:
  (a) $i \rightarrow j \in \mathcal{E}'$ iff there exists a node k such that $k \in \text{SC}_{\mathcal{G}}(j)$ and $i \rightarrow k \in \mathcal{E}$;
  (b) $i \leftrightarrow j \in \mathcal{F}'$ iff there exists a node k such that $k \in \text{SC}_{\mathcal{G}}(j)$ and $i \leftrightarrow k \in \mathcal{F}$;
(iii) for any pair of distinct nodes $\{i, j\}$ such that $i \in \text{SC}_{\mathcal{G}}(j)$: $i \rightarrow j \in \mathcal{E}'$ or $i \leftarrow j \in \mathcal{E}'$ or $i \leftrightarrow j \in \mathcal{F}'$.

In words: all strongly connected components are made fully-connected, edges between strongly connected components are preserved, and any edge into a node in a strongly connected component must be copied and made adjacent to all nodes in the strongly connected component. Note that a DMG may have multiple acyclifications. An example is given in Figure 2.

All acyclifications share certain "spurious" edges: the additional incoming directed and adjacent bidirected edges connecting nodes of two different strongly connected components. These have no *causal* interpretation but are necessary to correctly represent the σ-separation properties as d-separation properties. The skeleton of any acyclification $\mathcal{G}'$ of $\mathcal{G}$ equals the skeleton of $\mathcal{G}$ plus additional spurious adjacencies: the edges $i - j$ with $i \ast\!\!\rightarrow k$ and $k \in \text{SC}_{\mathcal{G}}(j)$, and the edges $i - j$ with $i \in \text{SC}_{\mathcal{G}}(j)$ where i and j are not adjacent in $\mathcal{G}$. These "spurious edges" added in any acyclification of a DMG $\mathcal{G}$ correspond with (non-trivial) inducing paths in $\mathcal{G}$.

The "raison d'être" for acyclifications is that they are σ-separation-equivalent to the original DMG, i.e., their σ-independence models agree:

Proposition 2 *For any DMG $\mathcal{G}$ and any acyclification $\mathcal{G}'$ of $\mathcal{G}$, $\text{IM}_{\sigma}(\mathcal{G}) = \text{IM}_{\sigma}(\mathcal{G}') = \text{IM}_{d}(\mathcal{G}')$.*

One particular acyclification that we will make use of repeatedly will be denoted $\mathcal{G}^{\text{acy}}$, and is obtained by replacing all strongly connected components of $\mathcal{G}$ by fully-connected bidirected components without any directed edges (i.e., if $i \in \text{SC}_{\mathcal{G}}(j)$ then $i \leftrightarrow j$ in $\mathcal{G}'$, but neither $i \rightarrow j$ nor $j \rightarrow i$ in $\mathcal{G}'$). Another useful set of acyclifications is obtained by replacing all strongly connected components of $\mathcal{G}$ by arbitrary fully-connected DAGs, and optionally adding an arbitrary set of bidirected edges.

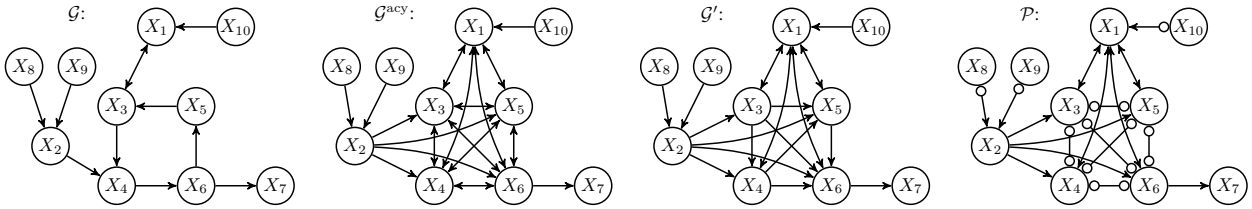

Figure 2: From left to right: Directed mixed graph $\mathcal{G}$, two of its acyclifications ($\mathcal{G}^{\text{acy}}$ and $\mathcal{G}'$) and the DPAG output by FCI $\mathcal{P} = \mathcal{P}_{\texttt{FCI}}(\text{IM}_\sigma(\mathcal{G})) = \mathcal{P}_{\texttt{FCI}}(\text{IM}_d(\mathcal{G}')) = \mathcal{P}_{\texttt{FCI}}(\text{IM}_d(\mathcal{G}^{\text{acy}}))$.

Other important properties of acyclifications are:

Proposition 3 *Let $\mathcal{G}$ be a DMG and i, j two nodes in $\mathcal{G}$.*

(i) If $i \in \text{AN}_{\mathcal{G}}(j)$ then there exists an acyclification $\mathcal{G}'$ of $\mathcal{G}$ with $i \in \text{AN}_{\mathcal{G}'}(j)$;
(ii) If $i \notin \text{AN}_{\mathcal{G}}(j)$ then $i \notin \text{AN}_{\mathcal{G}'}(j)$ for all acyclifications $\mathcal{G}'$ of $\mathcal{G}$;
(iii) There is an inducing path between i and j in $\mathcal{G}$ if and only if there is an inducing path between i and j in $\mathcal{G}'$ for any acyclification $\mathcal{G}'$ of $\mathcal{G}$.

3.4 SOUNDNESS AND COMPLETENESS

In the acyclic setting, the FCI algorithm was shown to be sound and complete (Zhang, 2008b). The notion of acyclifications, together with their elementary properties (Propositions 2 and 3) allows us to easily extend these soundness and completeness results to the σ-separation setting (allowing for cycles).

Consider FCI as a mapping $\mathcal{P}_{\texttt{FCI}}$ from independence models (on variables $\mathcal{V}$) to DPAGs (with vertex set $\mathcal{V}$), which maps the independence model of a DMG $\mathcal{G}$ to the DPAG $\mathcal{P}_{\texttt{FCI}}(\text{IM}_\sigma(\mathcal{G}))$.

Theorem 1 *In the σ-separation setting (but without selection bias), FCI is*

(i) sound*: for all DMGs $\mathcal{G}$, $\mathcal{P}_{FCI}(\text{IM}_\sigma(\mathcal{G}))$ contains $\mathcal{G}$;*
(ii) arrowhead complete*: for all DMGs $\mathcal{G}$: if $i \notin \text{AN}_{\tilde{\mathcal{G}}}(j)$ for any DMG $\tilde{\mathcal{G}}$ that is σ-Markov equivalent to $\mathcal{G}$, then there is an arrowhead $i \leftrightarrow\!\!* \; j$ in $\mathcal{P}_{FCI}(\text{IM}_\sigma(\mathcal{G}))$;*
(iii) tail complete*: for all DMGs $\mathcal{G}$, if $i \in \text{AN}_{\tilde{\mathcal{G}}}(j)$ in any DMG $\tilde{\mathcal{G}}$ that is σ-Markov equivalent to $\mathcal{G}$, then there is a tail $i \to j$ in $\mathcal{P}_{FCI}(\text{IM}_\sigma(\mathcal{G}))$;*
(iv) Markov complete*: for all DMGs $\mathcal{G}_1$ and $\mathcal{G}_2$, $\mathcal{G}_1$ is σ-Markov equivalent to $\mathcal{G}_2$ iff $\mathcal{P}_{FCI}(\text{IM}_\sigma(\mathcal{G}_1)) = \mathcal{P}_{FCI}(\text{IM}_\sigma(\mathcal{G}_2))$.*

Proof sketch: The main idea is the following (see also Figure 1). For all DMGs $\mathcal{G}$, $\text{IM}_\sigma(\mathcal{G}) = \text{IM}_d(\mathcal{G}')$ for any acyclification $\mathcal{G}'$ of $\mathcal{G}$ (Proposition 2). Hence FCI maps any acyclification $\mathcal{G}'$ of $\mathcal{G}$ to the same DPAG $\mathcal{P}_{\texttt{FCI}}(\text{IM}_\sigma(\mathcal{G}))$, and thereby any conclusion we draw about these acyclifications can be transferred back to a conclusion about $\mathcal{G}$ by means of Proposition 3. A complete proof is given in Section C of the Supplementary Material. □

Note that these definitions of soundness and completeness reduce to their acyclic counterparts (Zhang, 2008b) when restricting to ADMGs. In particular, the soundness and Markov completeness properties together imply that the DPAG $\mathcal{P}_{\texttt{FCI}}(\text{IM}_\sigma(\mathcal{G}))$ output by FCI, when given as input the σ-independence model of a DMG $\mathcal{G}$, represents the σ-Markov equivalence class of $\mathcal{G}$. In other words, FCI provides a characterization of the σ-Markov equivalence class of a DMG. This is, to the best of our knowledge, the first such characterization.

In order to read off the independence model from the DPAG $\mathcal{P}_{\texttt{FCI}}(\text{IM}_\sigma(\mathcal{G}))$, one can follow the same procedure as in the acyclic case: first construct a representative DMAG (for details, see Zhang (2008b)) and then apply the d-separation criterion to this DMAG. While the soundness of FCI (Theorem 1(i)) allows us to read off some (non-)ancestral relations from the DPAG output by FCI, this is by far not all causal information that is identifiable from the σ-Markov equivalence class. In the following sections, we will discuss how various causal features can be identified from DPAGs.

3.5 IDENTIFIABLE (NON-)ANCESTRAL RELATIONS

Zhang (2006) conjectured the soundness and completess of a criterion to read off all invariant ancestral relations from a complete DPAG, i.e., to identify the ancestral relations that are present in all Markov equivalent ADMGs that are represented by a complete DPAG. Roumpelaki et al. (2016) proved soundness of the criterion.[3] We extend Theorem 3.1 in (Roumpelaki et al., 2016) to DPAGs and DMGs:

[3]They 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.

Proposition 4 *Let $\mathcal{G}$ be a DMG, and let $\mathcal{P}$ be a DPAG that contains $\mathcal{G}$, and such that all unshielded triples in $\mathcal{P}$ have been oriented according to FCI rule $\mathcal{R}0$ (Zhang, 2008b) using $\mathrm{IM}_\sigma(\mathcal{G})$. For two nodes $i \neq j \in \mathcal{P}$: If*

- *there is a directed path from i to j in $\mathcal{P}$, or*
- *there exist uncovered possibly directed paths (see Definition 13) from i to j in $\mathcal{P}$ of the form $i, u, \ldots, j$ and $i, v, \ldots, j$ such that u, v are distinct non-adjacent nodes in $\mathcal{P}$,*

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

As an example, from the (complete) DPAG in Figure 2 it follows that $X_2 \in \mathrm{AN}_\mathcal{G}(X_4)$, and $X_2 \in \mathrm{AN}_\mathcal{G}(X_7)$.

Zhang (2006, p. 137) provides a sound and complete criterion to read off definite non-ancestors from a complete DPAG, assuming acyclicity. We can directly extend this criterion to DPAGs and DMGs:

Proposition 5 *Let $\mathcal{G}$ be a DMG, and let $\mathcal{P}$ be a DPAG that contains $\mathcal{G}$. For two nodes $i \neq j \in \mathcal{P}$: if there is no possibly directed path from i to j in $\mathcal{P}$ then $i \notin \mathrm{AN}_\mathcal{G}(j)$.*

As an example, from the DPAG in Figure 2 we can read off that X_8 cannot be ancestor of X_1 in $\mathcal{G}$, nor the other way around. However, $X_3 \circ\!\!\circ X_6 \rightarrow X_7$ is a possibly directed path in the DPAG, and so X_3 may be (and in this case is) ancestor of X_7 in $\mathcal{G}$.

3.6 IDENTIFIABLE NON-CONFOUNDED PAIRS

While in ADMGs and DMGs confounding is indicated by bidirected edges, in DPAGs confounding can also "hide" behind directed edges. The following notion is of key importance in this regard:

Definition 4 (Zhang (2008a)) *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*. The same notion applies to a DPAG, but is then called* definitely visible *(and its negation* possibly invisible*).*

For example, in the DPAG in Figure 2, edge $X_6 \rightarrow X_7$ is definitely visible (by virtue of $X_2 \rightarrow X_6$), as are all edges $X_2 \rightarrow \{X_3, X_4, X_5, X_6\}$ (by virtue of $X_8 \circ\!\!\rightarrow X_2$, or $X_9 \circ\!\!\rightarrow X_2$).

The notion of (in)visibility is closely related with confounding, as shown in Lemma 9 and 10 in Zhang (2008a). To generalize this, we make use of the following Lemma.

Lemma 1 *Let $\mathcal{P}$ be a DPAG that contains DMG $\mathcal{G}$, and let $k \ast\!\!\rightarrow i$ be an edge in $\mathcal{P}$ that is into i. Then there exists an inducing walk in $\mathcal{G}$ between k and i that is into i. If $k \leftrightarrow i$ in $\mathcal{P}$, then there exists an inducing walk in $\mathcal{G}$ between k and i that is both into k and into i.*

This allows us to generalize Lemma 9 in (Zhang, 2008a) to the cyclic setting (with almost identical proof).

Lemma 2 *Let $\mathcal{P}$ be a DPAG, and $i \rightarrow j$ a directed edge in $\mathcal{P}$. If $i \rightarrow j$ is definitely visible in $\mathcal{P}$, then for all DMGs $\mathcal{G}$ contained in $\mathcal{P}$, there exists no inducing walk between i and j in $\mathcal{G}$ that is into i.*

This provides us with a sufficient condition to read off unconfounded pairs of nodes from DPAGs:

Proposition 6 *Let $\mathcal{P}$ be a DMAG and $\mathcal{G}$ be a DMG contained in $\mathcal{P}$. Let $i \neq j$ be two nodes in $\mathcal{P}$. If i and j are not adjacent in $\mathcal{P}$, or if there is a directed edge $i \rightarrow j$ in $\mathcal{P}$ that is definitely visible in $\mathcal{P}$, then $i \leftrightarrow j$ is absent from $\mathcal{G}$.*

For example, from the DPAG in Figure 2 one can infer that there is no bidirected edge $X_2 \leftrightarrow X_7$ in the underlying DMG $\mathcal{G}$, as the two nodes are not adjacent in the DPAG, and also that there is no bidirected edge between X_2 and any node in $\{X_3, X_4, X_5, X_6\}$ in $\mathcal{G}$, as all these edges are definitely visible in the DPAG.

3.7 IDENTIFYING DIRECT (NON-)CAUSES

Contrary to DMGs, a directed edge in a DPAG does not necessarily correspond with a *direct* causal relation. The following proposition provides sufficient conditions to identify the absence of a directed edge from the DPAG.

Proposition 7 *Let $\mathcal{P}$ be a DPAG that contains a DMG $\mathcal{G}$. For two nodes $i \neq j$ in $\mathcal{P}$, if $i \leftrightarrow\!\!\ast\, j$ in $\mathcal{P}$, or i and j are not adjacent in $\mathcal{P}$, then $i \rightarrow j$ is not present in $\mathcal{G}$.*

The following proposition was inspired by Theorem 3 in Borboudakis et al. (2012) and provides sufficient conditions to conclude the presence of a directed edge from the DPAG.

Proposition 8 *Let $\mathcal{P}$ be a DPAG that contains a DMG $\mathcal{G}$. For two nodes $i \neq j$ in $\mathcal{P}$, if $i \rightarrow j$ in $\mathcal{P}$ and:*

(i) there does not exist a possibly directed path from i to j in $\mathcal{P}$ that avoids the edge $i \rightarrow j$, or

(ii) if there is no inducing walk between i and j in $\mathcal{G}$ that is both into i and j (for example, because $i \rightarrow j$ is definitely visible in $\mathcal{P}$), and for all vertices k such that there is a possibly directed path $i \ast\!\!\rightarrow k \ast\!\!\rightarrow j$ from i to j in $\mathcal{P}$, the edge $k \rightarrow j$ is

definitely visible in the DPAG $\mathcal{P}^*$ obtained from $\mathcal{P}$ by replacing the edge between k and j by $k \to j$,

then $i \to j$ is present in $\mathcal{G}$.

As an example, the edge $X_2 \to X_3$ in the DPAG in Figure 2 cannot be identified as being present in $\mathcal{G}$ because both conditions are not satisfied: (i) because of the possibly directed path $X_2 \to X_4 \circ\!\!\circ X_3$, (ii) because of the same path where the edge $X_4 \to X_3$ would be possibly invisible if oriented in that way. Also the edge $X_1 \to X_3$ in the DPAG cannot be identified as being present in $\mathcal{G}$. The edge $X_6 \to X_7$ in the DPAG, on the other hand, is identifiably present in $\mathcal{G}$.

3.8 IDENTIFIABLE NON-CYCLES

Strongly connected components in the DMG end up as a specific pattern in the DPAG. This can be used as a sufficient condition for identifying the absence of certain cyclic causal relations in a complete DPAG.

Proposition 9 *Let $\mathcal{G}$ be a DMG and denote by $\mathcal{P} = \mathcal{P}_{FCI}(\mathrm{IM}_\sigma(\mathcal{G}))$ the corresponding complete DPAG output by FCI. Let $i \neq j$ be two nodes in $\mathcal{P}$. If $j \in \mathrm{SC}_{\mathcal{G}}(i)$, then $i \circ\!\!\circ j$ in $\mathcal{P}$, and for all nodes k: $k \to i$ in $\mathcal{P}$ iff $k \to j$ in $\mathcal{P}$, and $k \leftrightarrow i$ in $\mathcal{P}$ iff $k \leftrightarrow j$ in $\mathcal{P}$, and $k \circ\!\!\to i$ in $\mathcal{P}$ iff $k \circ\!\!\to j$ in $\mathcal{P}$.*

Hence, any pair of nodes that does not fit this pattern cannot be part of a cycle in $\mathcal{G}$. For example, in the complete DPAG in Figure 2, only the nodes in $\{X_3, X_4, X_5, X_6\}$ might be part of a cycle. For all other pairs of nodes, it follows from Proposition 9 that they cannot be part of a cycle. This sufficient condition is also necessary:

Proposition 10 *Let $\mathcal{G}$ be a DMG and denote by $\mathcal{P} = \mathcal{P}_{FCI}(\mathrm{IM}_\sigma(\mathcal{G}))$ the corresponding complete DPAG output by FCI. Let $i \neq j$ be two nodes in $\mathcal{P}$. If there is an edge $i \circ\!\!\circ j$ in $\mathcal{P}$, and all nodes k for which $k \ast\!\!\to i$ is in $\mathcal{P}$ also have an edge of the same type $k \ast\!\!\to j$ (i.e., the two edge marks at k are the same) in $\mathcal{P}$, then there exists a DMG $\tilde{\mathcal{G}}$ with $j \in \mathrm{SC}_{\tilde{\mathcal{G}}}(i)$ that is σ-Markov equivalent to $\mathcal{G}$, but also a DMG $\mathcal{H}$ with $j \notin \mathrm{SC}_{\mathcal{H}}(i)$ that is σ-Markov equivalent to $\mathcal{G}$.*

In other words, under the conditions of this proposition, it is not identifiable from $\mathcal{P}$ alone whether j and i are part of a causal cycle.

4 EXTENSIONS FOR BACKGROUND KNOWLEDGE

In this section, we discuss extensions of our results to situations in which available causal background knowledge is taken into account by causal discovery algorithms.

Assume that we have certain background knowledge, formalized as a Boolean function Ψ on the set of all DMGs (indicating for each DMG whether it satisfies the background knowledge). For example, one type of background knowledge commonly considered in the literature (probably mainly for reasons of simplicity) is *causal sufficiency*, which can be formalized by $\Psi(\mathcal{G}) = 1$ iff $\mathcal{G}$ contains no bidirected edges, and $\Psi(\mathcal{G}) = 0$ otherwise. A less trivial example of background knowledge are the JCI Assumptions, which play a central role in the Joint Causal Inference framework (Mooij et al., 2020) for performing causal discovery from multiple datasets that correspond with measurements of a system in different contexts (for example, a combination of observational and different interventional datasets). The latter example will be discussed in more detail in Section 4.3.

4.1 SOUNDNESS AND COMPLETENESS

We first extend the standard notions of soundness and completeness to a setting that involves cycles and background knowledge (but no selection bias).

Definition 5 *Under background knowledge Ψ, a mapping Φ from independence models to DPAGs is called:*

- sound *if for all DMGs $\mathcal{G}$ with $\Psi(\mathcal{G}) = 1$: $\Phi(\mathrm{IM}_\sigma(\mathcal{G}))$ contains $\mathcal{G}$;*
- arrowhead complete *if for all DMGs $\mathcal{G}$ with $\Psi(\mathcal{G}) = 1$: if $i \notin \mathrm{AN}_{\tilde{\mathcal{G}}}(j)$ for any DMG $\tilde{\mathcal{G}}$ with $\Psi(\tilde{\mathcal{G}}) = 1$ that is σ-Markov equivalent to $\mathcal{G}$, then there is an arrowhead $i \leftrightarrow\!\!\ast\ j$ in $\Phi(\mathrm{IM}_\sigma(\mathcal{G}))$;*
- tail complete *if for all DMGs $\mathcal{G}$ with $\Psi(\mathcal{G}) = 1$: if $i \in \mathrm{AN}_{\tilde{\mathcal{G}}}(j)$ in any DMG $\tilde{\mathcal{G}}$ with $\Psi(\tilde{\mathcal{G}}) = 1$ that is σ-Markov equivalent to $\mathcal{G}$, then there is a tail $i \to j$ in $\Phi(\mathrm{IM}_\sigma(\mathcal{G}))$;*
- Markov complete *if for all DMGs $\mathcal{G}_1, \mathcal{G}_2$ with $\Psi(\mathcal{G}_1) = \Psi(\mathcal{G}_2) = 1$: $\mathcal{G}_1$ is σ-Markov equivalent to $\mathcal{G}_2$ iff $\Phi(\mathrm{IM}_\sigma(\mathcal{G}_1)) = \Phi(\mathrm{IM}_\sigma(\mathcal{G}_2))$.*

It is called complete if it is both arrowhead complete and tail complete.

Note that this reduces to the standard notions (Zhang, 2008b) if $\Psi(\mathcal{G}) = 1$ iff $\mathcal{G}$ is acyclic, while it also reduces to the notions in Theorem 1 if no background knowledge is used (i.e., $\Psi(\mathcal{G}) = 1$ for all $\mathcal{G}$).

We assume that the background knowledge is *compatible with the acyclification* in the following sense:

Assumption 1 *For all DMGs $\mathcal{G}$ with $\Psi(\mathcal{G}) = 1$, the following three conditions hold:*

(i) There exists an acyclification $\mathcal{G}'$ of $\mathcal{G}$ with $\Psi(\mathcal{G}') = 1$;

(ii) For all nodes i, j in $\mathcal{G}$: if $i \in \mathrm{AN}_{\mathcal{G}}(j)$ then there

exists an acyclification $\mathcal{G}'$ of $\mathcal{G}$ with $\Psi(\mathcal{G}') = 1$ such that $i \in \text{AN}_{\mathcal{G}'}(j)$;

(iii) For all nodes i, j in $\mathcal{G}$: if $i \notin \text{AN}_{\mathcal{G}}(j)$ then $i \notin \text{AN}_{\mathcal{G}'}(j)$ for all acyclifications $\mathcal{G}'$ of $\mathcal{G}$ with $\Psi(\mathcal{G}') = 1$.

For example, the background knowledge of "causal sufficiency" satisfies this assumption, as well as the background knowledge of "acyclicity".

The following result is straightforward given all the definitions, but is also quite powerful, as it allows us to directly generalize existing acyclic soundness and completeness results (for certain background knowledge) to the σ-separation setting.

Theorem 2 *Let Ψ be background knowledge that satisfies Assumption 1 and let Φ be a mapping from independence models to DPAGs. Then:*

(i) If Φ is sound for background knowledge Ψ under the additional assumption of acyclicity, then Φ is sound for background knowledge Ψ.

(ii) If Φ is arrowhead (tail) complete for background knowledge Ψ under the additional assumption of acyclicity, then Φ is arrowhead (tail) complete for background knowledge Ψ.

(iii) If Φ is sound and arrowhead complete for background knowledge Ψ under the additional assumption of acyclicity, then Φ is Markov complete.

In the remainder of this section, we will apply this result to two types of background knowledge: causal sufficiency, and the JCI assumptions.

4.2 CAUSAL SUFFICIENCY

We consider the (commonly assumed) background knowledge of "causal sufficiency". This is formalized by $\Psi(\mathcal{G}) = 1$ iff DMG $\mathcal{G}$ contains no bidirected edges. For the acyclic setting, the well-known PC algorithm (Spirtes et al., 2000), adapted with Meek's orientation rules (Meek, 1995a), was shown to be sound and complete. It outputs a so-called Complete Partially Directed Acyclic Graph (CPDAG), which can be interpreted also as a DPAG (by replacing all undirected edges $i - j$ by bicircle edges $i \circ\!\!-\!\!\circ j$). Because this particular background knowledge satisfies Assumption 1, we can apply Theorem 2 to extend the existing acyclic soundness and completeness results to the cyclic setting:

Corollary 1 *The PC algorithm with Meek's orientation rules is sound, arrowhead complete, tail complete and Markov complete (in the σ-separation setting without selection bias).*

We can therefore also apply Propositions 4, 5, to read off the absence or presence of indirect causal relations from the DPAG (obtained from the CPDAG) output by the PC algorithm. Note that the presence or absence of direct causal relations can be easily read off from the DPAG in this case as they are in one-to-one correspondence with directed edges in the DPAG.

4.3 JOINT CAUSAL INFERENCE

Recently, Mooij et al. (2020) proposed FCI-JCI, an extension of FCI that enables causal discovery from data measured in different contexts (for example, if observational data as well as data corresponding to various interventions is available). This is a particular implementation of the general Joint Causal Inference (JCI) framework. For a detailed treatment, we refer the reader to (Mooij et al., 2020); here we only give a brief summary of the JCI assumptions that we need to extend our results on FCI to FCI-JCI.

Definition 6 (JCI Assumptions) *The data-generating mechanism for a system in a context is described by a simple SCM $\mathcal{M}$ with two types of endogenous variables:* system *variables $\{X_i\}_{i \in \mathcal{I}}$ and* context *variables $\{C_k\}_{k \in \mathcal{K}}$. Its graph $\mathcal{G}(\mathcal{M})$ has nodes $\mathcal{I} \cup \mathcal{K}$ (corresponding to system variables and context variables, respectively). The following (optional) JCI Assumptions can be made about the graph $\mathcal{G} := \mathcal{G}(\mathcal{M})$:*

(1) Exogeneity*: No system variable causes any context variable, i.e., $\forall_{k \in \mathcal{K}} \forall_{i \in \mathcal{I}} : i \to k \notin \mathcal{G}$.*

(2) Randomization*: No pair of context and system variable is confounded, i.e., $\forall_{k \in \mathcal{K}} \forall_{i \in \mathcal{I}} : i \leftrightarrow k \notin \mathcal{G}$.*

(3) Genericity*: The induced subgraph $\mathcal{G}(\mathcal{M})_{\mathcal{K}}$ on the context variables is of the following special form: $\forall_{k \neq k' \in \mathcal{K}} : k \leftrightarrow k' \in \mathcal{G} \wedge k \to k' \notin \mathcal{G}$.*

The following Lemma is key to our extensions to the cyclic σ-separation setting.

Lemma 3 *If subset $\{1\}$, $\{1, 2\}$, or $\{1, 2, 3\}$ of the JCI Assumptions holds for a DMG $\mathcal{G}$, then the same subset of assumptions holds for any acyclification of $\mathcal{G}$.*

This trivially implies that these different combinations of the JCI Assumptions satisfy Assumption 1. That allows us to extend the existing acyclic soundness and completeness results for FCI-JCI to the cyclic setting.

FCI-JCI was shown to be sound under the assumption of acyclicity (Theorem 35, Mooij et al., 2020). This gives with Theorem 2:

Corollary 2 *For the background knowledge consisting of JCI Assumptions $\emptyset$, $\{1\}$, $\{1, 2\}$ or $\{1, 2, 3\}$, the corresponding version of FCI-JCI is sound (in the σ-separation setting without selection bias).*

We can therefore also apply Propositions 5 and 6 to read off the absence of indirect causal relations and confounding from the DPAG output by the FCI-JCI algorithm, and Propositions 7 and 8 to read off the absence or presence of direct causal relations. Furthermore, it is clear from its definition that all unshielded triples in the DPAG that FCI-JCI outputs have been oriented according to FCI rule $\mathcal{R}0$. Therefore, we can also apply Proposition 4 to read off the presence of indirect causal relations from the DPAG output by the FCI-JCI algorithm.

Under all three JCI assumptions, stronger results have been derived. In particular, completeness of FCI-JCI has been shown (Theorem 38 Mooij et al., 2020) under the background knowledge of all three JCI Assumptions in the acyclic setting. This gives with Theorem 2:

Corollary 3 *For the background knowledge consisting of JCI Assumptions $\{1, 2, 3\}$, the FCI-JCI algorithm is arrowhead complete, tail complete and Markov complete (in the σ-separation setting without selection bias).*

An important feature of Joint Causal Inference under JCI Assumptions $\{1, 2, 3\}$ is that the direct (non-)targets of interventions need not be known, but can be discovered from the data. The sufficient condition provided in Proposition 42 of Mooij et al. (2020) can be easily generalized to the σ-separation setting as well by observing that under JCI Assumptions $\{1, 2, 3\}$, there cannot be an inducing walk between a system node and a context node that is into both, and then applying Proposition 7 and Proposition 8. For details, see Proposition 12 in Section B of the Supplementary Material.

Furthermore, also Proposition 9 that allows one to identify the absence of cycles can be extended to FCI-JCI under JCI Assumptions $\{1, 2, 3\}$. For details, see Proposition 13 in Section B of the Supplementary Material.

5 DISCUSSION AND CONCLUSION

We have shown that, surprisingly, the FCI algorithm and several of its variants that were designed for the acyclic setting need not be adapted but directly apply also in the cyclic setting under the assumptions of the σ-Markov property, σ-faithfulness, and the absence of selection bias. Furthermore, we have provided sufficient conditions to identify causal features from the DPAG output by FCI and its variants. For convenience, we state this as a corollary, collecting several of our results.

Corollary 4 *Let $\mathcal{M}$ be a simple (possibly cyclic) SCM with graph $\mathcal{G}(\mathcal{M})$ and assume that its distribution $\mathbb{P}_{\mathcal{M}}(\boldsymbol{X})$ is σ-faithful w.r.t. the graph $\mathcal{G}(\mathcal{M})$. When using consistent conditional independence tests on an i.i.d. sample of observational data from the induced distribution $\mathbb{P}_{\mathcal{M}}(\boldsymbol{X})$ of $\mathcal{M}$, FCI provides a consistent estimate $\hat{\mathcal{P}}$ of the DPAG $\mathcal{P}_{\texttt{FCI}}(\mathrm{IM}_{\sigma}(\mathcal{G}(\mathcal{M})))$ that represents the σ-Markov equivalence class of $\mathcal{G}(\mathcal{M})$. From the estimated DPAG $\hat{\mathcal{P}}$, we obtain consistent estimates for: (i) the absence/presence of (possibly indirect) causal relations according to $\mathcal{M}$ via Propositions 4 and 5; (ii) the absence of confounders according to $\mathcal{M}$ via Proposition 6; (iii) the absence/presence of direct causal relations according to $\mathcal{M}$ via Propositions 7 and 8; (iv) the absence of causal cycles according to $\mathcal{M}$ via Proposition 9.*

A similar conclusion can be formulated for the FCI-JCI algorithm (see Section B of the Supplementary Material). Obviously, our results apply also in the acyclic setting (where σ-separation reduces to d-separation).

One important limitation of the σ-faithfulness assumption is that it excludes the linear and discrete cases. In pioneering work Richardson (1996b) already proposed a constraint-based causal discovery algorithm (NL-CCD) that made use of the σ-separation Markov assumption, while assuming only the d-faithfulness assumption (which is weaker than the σ-faithfulness assumption). In future work, we plan to investigate this setting as well, as well as the possibility of extending our results to a setting that does not rule out selection bias.

Acknowledgements

We are indebted to Jiji Zhang for contributing the proof of Proposition 11. We thank the reviewers for their constructive feedback that helped us improve this paper. This work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement 639466).

References

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

Ali, R. A., Richardson, T. S., Spirtes, P., and Zhang, J. (2005). Towards characterizing Markov equivalence classes for directed acyclic graphs with latent variables. In *Proceedings of the 21th Conference on Uncertainty in Artificial Intelligence (UAI-05)*, pages 10–17.

Bongers, S., Forré, P., Peters, J., Schölkopf, B., and Mooij, J. M. (2020). Foundations of structural causal models with cycles and latent variables. *arXiv.org preprint*, arXiv:1611.06221v3 [stat.ME].

Bongers, S. and Mooij, J. M. (2018). From random differential equations to structural causal models: the stochastic case. *arXiv.org preprint*, arXiv:1803.08784v2 [cs.AI].

Borboudakis, G., Triantafillou, S., and Tsamardinos, I. (2012). Tools and algorithms for causally interpreting directed edges

in maximal ancestral graphs. In *Proceedings of the Sixth European Workshop on Probabilistic Graphical Models (PGM 2012)*, pages 35–42.

Claassen, T., Mooij, J. M., and Heskes, T. (2013). Learning sparse causal models is not NP-hard. In *Proceedings of the 29th Annual Conference on Uncertainty in Artificial Intelligence (UAI-13)*, pages 172–181.

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

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

Forré, P. and Mooij, J. M. (2017). Markov properties for graphical models with cycles and latent variables. *arXiv.org preprint*, arXiv:1710.08775 [math.ST].

Forré, P. and Mooij, J. M. (2018). Constraint-based causal discovery for non-linear structural causal models with cycles and latent confounders. In *Proceedings of the 34th Annual Conference on Uncertainty in Artificial Intelligence (UAI-18)*.

Hyttinen, A., Eberhardt, F., and Hoyer, P. (2010). Causal discovery for linear cyclic models with latent variables. In *Proceedings of the Fifth European Workshop on Probabilistic Graphical Models (PGM 2010)*, pages 153–160.

Hyttinen, A., Eberhardt, F., and Hoyer, P. (2012). Learning linear cyclic causal models with latent variables. *Journal of Machine Learning Research*, 13:3387–3439.

Hyttinen, A., Eberhardt, F., and Järvisalo, M. (2014). Constraint-based causal discovery: Conflict resolution with answer set programming. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence, (UAI-14)*, pages 340–349.

Lacerda, G., Spirtes, P., Ramsey, J., and Hoyer, P. O. (2008). Discovering cyclic causal models by independent components analysis. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI-08)*.

Mani, S. (2006). *A Bayesian Local Causal Discovery Framework*. PhD thesis, University of Pittsburg.

Meek, C. (1995a). Causal inference and causal explanation with background knowledge. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence (UAI-95)*, pages 403–411.

Meek, C. (1995b). Strong completeness and faithfulness in Bayesian networks. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (UAI-95)*, pages 411–419.

Mooij, J. M., Magliacane, S., and Claassen, T. (2020). Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, 21(99):1–108.

Pearl, J. (1986). A constraint propagation approach to probabilistic reasoning. In *Proceedings of the First Conference on Uncertainty in Artificial Intelligence (UAI-85)*, pages 357–370.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference*. Cambridge University Press.

Peters, J., Bühlmann, P., and Meinshausen, N. (2016). Causal inference using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society, Series B*, 78(5):947–1012.

Richardson, T. (1996a). A discovery algorithm for directed cyclic graphs. In *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence (UAI-96)*.

Richardson, T. and Spirtes, P. (1999). Automated discovery of linear feedback models. In *Computation, Causation, and Discovery*, pages 253–304. MIT Press.

Richardson, T. S. (1996b). *Models of Feedback: Interpretation and Discovery*. PhD thesis, Carnegie-Mellon University.

Richardson, T. S. and Spirtes, P. (2002). Ancestral graph Markov models. *The Annals of Statistics*, 30(4):962–1030.

Rothenhäusler, D., Heinze, C., Peters, J., and Meinshausen, N. (2015). BACKSHIFT: Learning causal cyclic graphs from unknown shift interventions. In *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, pages 1513–1521.

Roumpelaki, A., Borboudakis, G., Triantafillou, S., and Tsamardinos, I. (2016). Marginal causal consistency in constraint-based causal learning. In *Proceedings of the UAI 2016 Workshop on Causation: Foundation to Application*, number 1792 in CEUR Workshop Proceedings, pages 39–47.

Spirtes, P. (1994). Conditional independence in directed cyclic graphical models for feedback. Technical Report CMU-PHIL-54, Carnegie Mellon University.

Spirtes, P. (1995). Directed cyclic graphical representations of feedback models. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (UAI-95)*, pages 499–506.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction, and Search*. MIT press, 2nd edition.

Spirtes, P., Meek, C., and Richardson, T. S. (1995). Causal inference in the presence of latent variables and selection bias. In *Proceedings of the 11th Annual Conference on Uncertainty in Artificial Intelligence (UAI-95)*, pages 499–506.

Spirtes, P., Meek, C., and Richardson, T. S. (1999). An algorithm for causal inference in the presence of latent variables and selection bias. In *Computation, Causation and Discovery*, chapter 6, pages 211–252. The MIT Press.

Spirtes, P. and Verma, T. (1992). Equivalence of causal models with latent variables. Technical Report CMU-PHIL-33, Carnegie Mellon University.

Strobl, E. V. (2018). A constraint-based algorithm for causal discovery with cycles, latent variables and selection bias. *International Journal of Data Science and Analytics*, 8:33–56.

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

Zhang, J. (2006). *Causal Inference and Reasoning in Causally Insufficient Systems*. PhD thesis, Carnegie Mellon University.

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

Zhang, J. (2008b). 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.