Constraint-based Causal Discovery for Non-Linear Structural Causal Models with Cycles and Latent Confounders

Patrick Forré
Informatics Institute
University of Amsterdam
The Netherlands
p.d.forre@uva.nl

Joris M. Mooij
Informatics Institute
University of Amsterdam
The Netherlands
j.m.mooij@uva.nl

Abstract

We address the problem of causal discovery from data, making use of the recently proposed causal modeling framework of *modular structural causal models (mSCM)* to handle cycles, latent confounders and non-linearities. We introduce *σ-connection graphs (σ-CG)*, a new class of mixed graphs (containing undirected, bidirected and directed edges) with additional structure, and extend the concept of *σ-separation*, the appropriate generalization of the well-known notion of d-separation in this setting, to apply to σ-CGs. We prove the closedness of σ-separation under marginalisation and conditioning and exploit this to implement a test of σ-separation on a σ-CG. This then leads us to the first causal discovery algorithm that can handle non-linear functional relations, latent confounders, cyclic causal relationships, and data from different (stochastic) perfect interventions. As a proof of concept, we show on synthetic data how well the algorithm recovers features of the causal graph of modular structural causal models.

1 INTRODUCTION

Correlation does not imply causation. To go beyond spurious probabilistic associations and infer the asymmetric causal relations we need sufficiently powerful models. Structural causal models (SCMs), also known as structural equation models (SEMs), provide a popular modeling framework (see [12, 25, 26, 32]) that is up to this task. Still, the problem of causal discovery from data is notoriously hard. Theory and algorithms need to address several challenges like probabilistic settings, stability under interventions, combining observational and interventional data, latent confounders and marginalisation, selection bias and conditioning, faithfulness violations, cyclic causation like feedback loops and pairwise interactions, and non-linear functional relations in order to go beyond artificial simulation settings and become successful on real-world data.

Several algorithms for causal discovery have been introduced over the years. For the acyclic case without latent confounders, numerous constraint-based [25,32] and score-based approaches [6, 14, 17] exist. More sophisticated constraint-based [5, 25, 32, 34] and score-based approaches [4, 7, 8, 10, 11] can deal with latent confounders in the acyclic case. For the linear cyclic case, most algorithms assume no latent confounders [18, 27, 29], though some of the more recent ones allow for those [19, 30]. To the best of our knowledge, no algorithms have yet been proposed for the general non-linear cyclic case.

In this work we present a novel conditional independence constraint-based causal discovery algorithm that—up to the knowledge of the authors—is the first causal discovery algorithm that addresses most of the previously mentioned problems at once, notably non-linearities, cycles, latent confounders, and multiple interventional data sets, only excluding selection bias and faithfulness violations.

For this to work we build upon the theory of *modular structural causal models (mSCM)* introduced in [12]. mSCMs form a general and convenient class of structural causal models that can deal with cycles and latent confounders. The measure-theoretically rigorous presentation opens the door for general non-linear measurable functions and any kind of probability distributions (e.g. mixtures of discrete or continuous ones). mSCM are provably closed under any combination of perfect interventions and marginalisations (see [12]).

Unfortunately, it is known that the direct generalization of the *d-separation criterion* (also called m- or m^*-separation, see [9, 24, 25, 28, 33]), which relates the conditional independencies of the model to its underlying

graphical structure, does not apply in general if the structural equations are *non-linear* and the graph contains *cycles* (see [12, 31] or example 2.17).

Luckily, one key property of mSCMs is that the variables of the mSCM always entail the conditional independences implied by *σ-separation*, a non-naive generalization of the d/m/m*-separation (see [12]), which also works in the presence of cycles, non-linearities and latent confounders, and reduces to d-separation in the acyclic case.

To prove the σ-separation criterion, the authors of [12] have constructed an extensive theory for directed graphs with hyperedges. As a first contribution in this paper we give a simplified but equivalent definition of mSCMs plainly in terms of directed graphs and prove the σ-separation criterion directly under weaker assumptions.

As a second contribution we extend the definition of σ-separation to mixed graphs (including also bi- and undirected edges) by introducing additional structure. We will refer to this class of mixed graphs as *σ-connection graphs (σ-CG)*, since they are inspired by the d-connection graphs introduced by [19]. We prove that σ-CGs and σ-separation are closed under marginalisation and conditioning, in analogy with the d-connection graphs (d-CG) from [19].

The work of [19] provides an elegant approach to causal discovery using a weighted SAT solver to find the causal graph that is most compatible with (weighted) conditional independences in the data, encoding the notion of d-separation into *answer set programming (ASP)*, a declarative programming language with an expressive syntax for implementing discrete or integer optimization problems. Our third contribution is to adapt the approach of [19] by replacing d-separation by σ-separation and d-CGs by σ-CGs. The results mentioned above will then ensure all the needed properties to make the adapted algorithm of [19] applicable to general mSCMs, i.e., to non-linear causal models with cycles and latent confounders, under the additional assumptions of *no selection bias* and of *σ-faithfulness*.

Finally, as a proof of concept, we will show the effectiveness of our proposed algorithm in recovering features of the causal graphs of mSCMs from simulated data.

2 THEORY

2.1 MODULAR STRUCTURAL CAUSAL MODELS

Structural causal or equation models (SCM/SEM) usually start with a set of variables $(X_v)_{v\in V}$ attached to a graph G that satisfy (or are even defined by) equations of the form:

$$X_v = g_{\{v\}}(X_{\mathrm{Pa}^G(v)}, E_v),$$

with a function $g_{\{v\}}$ and noise variable E_v attached to each node $v \in V$. Here $\mathrm{Pa}^G(v)$ denotes the set of (direct causal) parents of v. In *linear* models the functions $g_{\{v\}}$ are linear functions, in *acyclic* models the nodes of V form an acyclic graph G, and under *causal sufficiency* the variables E_v are independent (i.e. "no latent confounders"). The functions $g_{\{v\}}$ are usually interpreted as *local causal mechanisms* that produce the values of X_v from the values of $X_{\mathrm{Pa}^G(v)}$ and E_v. These local mechanisms $g_{\{v\}}$ are—in the causal setting—assumed to be stable even when one intervenes upon some of the variables, i.e. one makes a causal *local compatibility* assumption. One important observation now is that one can also consider the *global mechanism* g that maps the values of the latent variables $(E_v)_{v\in V}$ to the values of the observed variables $(X_v)_{v\in V}$. The assumption of acyclicity or invertible linearity will then guarantee the *global compatibility* of all these mechanisms $g_{\{v\}}$ and g. However, if we now abstain from assuming acyclicity or linearity, the global compatibility does not follow from the local compatibility anymore (see figure 1). So in a general consistent causal setting this needs to be *guaranteed or assumed*.

Figure 1: Gear analogy. Left: A cyclic mechanism that is locally compatible but not globally. Center: For acyclic mechanisms local compatibility implies global compatibility. Right: A cyclic mechanism that is locally and globally compatible. This shows that the assumption of global compatibility is needed when cycles are present.

The definition of modular structural causal models (mSCM) and the mentioned list of desirable properties basically follow directly from *causal* postulates:

Postulate 2.1 (Causal postulates)**.** *The observed world appears as the projection of an extended world such that:*

1. *All latent and observed variables in this extended world are causally linked by a directed graph.*
2. *Every subsystem of this extended world can be expressed as the joint effect of its joint direct causes.*
3. *All these mechanisms are globally compatible.*

Special subsystems of interest are the loops of a graph.

Definition 2.2 (Loops)**.** *Let $G = (V, E)$ be a directed graph (with or without cycles).*

1. *A* loop *of G is a set of nodes $S \subseteq V$ such that for every two nodes $v_1, v_2 \in S$ there are two directed paths $v_1 \longrightarrow \cdots \longrightarrow v_2$ and $v_2 \longrightarrow \cdots \longrightarrow v_1$ in G with all the intermediate nodes also in S (if any). The single element sets $S = \{v\}$ are also considered as loops.*
2. *The* strongly connected component *of v in G is defined to be:*

$$\mathrm{Sc}^G(v) := \mathrm{Anc}^G(v) \cap \mathrm{Desc}^G(v),$$

   the set of nodes that are both ancestors and descendants of v (including v itself).
3. *Let $\mathcal{L}(G) := \{S \subseteq G \mid S \text{ a loop of } G\}$ be the* loop set *of G.*

Remark 2.3. *Note that the loop set $\mathcal{L}(G)$ contains all single element loops $\{v\} \in \mathcal{L}(G)$, $v \in V$, as the smallest loops and all strongly connected components $\mathrm{Sc}^G(v) \in \mathcal{L}(G)$, $v \in V$, as the largest loops, but also all non-trivial intermediate loops S with $\{v\} \subsetneq S \subsetneq \mathrm{Sc}^G(v)$ inside the strongly connected components (if existent). If G is acyclic then $\mathcal{L}(G)$ only consists of the single element loops: $\mathcal{L}(G) = \{\{v\} \mid v \in V\}$.*

The definition of mSCM is made in such a way that it will automatically incorporate the causal postulates 2.1. In the following, all spaces are meant to be equipped with σ-algebras, forming standard measurable spaces, and all maps to be measurable.

Definition 2.4 (Modular Structural Causal Model, [12])**.** *A* modular structural causal model (mSCM) *by definition consists of:*

1. *a set of nodes $V^+ = U \dot\cup V$, where elements of V correspond to observed variables and elements of U to latent variables,*
2. *an observation/latent space $\mathcal{X}_v$ for every $v \in V^+$, $\mathcal{X} := \prod_{v \in V^+} \mathcal{X}_v$,*
3. *a product probability measure $\mathbb{P} := \mathbb{P}_U = \otimes_{u \in U} \mathbb{P}_u$ on the latent space $\prod_{u \in U} \mathcal{X}_u$,*[1]
4. *a directed graph structure $G^+ = (V^+, E^+)$ with the properties:*[2]
   (a) $V = \mathrm{Ch}^{G^+}(U)$,
   (b) $\mathrm{Pa}^{G^+}(U) = \emptyset$,[3]
   (c) $\mathrm{Ch}^{G^+}(u_1) \nsubseteq \mathrm{Ch}^{G^+}(u_2)$ *for every two distinct* $u_1, u_2 \in U$,[3]

   where Ch^{G^+} and Pa^{G^+} stand for children and parents in G^+, resp.,
5. *a system of structural equations $g = (g_S)_{\substack{S \in \mathcal{L}(G^+) \\ S \subseteq V}}$:*

$$g_S : \prod_{v \in \mathrm{Pa}^{G^+}(S) \setminus S} \mathcal{X}_v \to \prod_{v \in S} \mathcal{X}_v,$$[4]

   that satisfy the following global compatibility *conditions: For every nested pair of loops $S' \subseteq S \subseteq V$ of G^+ and every element $x_{\mathrm{Pa}^{G^+}(S) \cup S} \in \prod_{v \in \mathrm{Pa}^{G^+}(S) \cup S} \mathcal{X}_v$ we have the implication:*

$$\begin{aligned} & g_S(x_{\mathrm{Pa}^{G^+}(S) \setminus S}) &&= x_S \\ \implies \quad & g_{S'}(x_{\mathrm{Pa}^{G^+}(S') \setminus S'}) &&= x_{S'}, \end{aligned}$$

   where $x_{\mathrm{Pa}^{G^+}(S') \setminus S'}$ and $x_{S'}$ denote the corresponding components of $x_{\mathrm{Pa}^{G^+}(S) \cup S}$.

The mSCM *can be summarized by the tuple $M = (G^+, \mathcal{X}, \mathbb{P}, g)$.*

Remark 2.5. *Given the mechanisms attached to the nodes $g_{\{v\}}$ the existence (and compatibility) of the other mechanisms g_S for non-trivial loops S can be guaranteed under certain conditions, e.g. trivially in the acyclic case, or if every cycle is contractive (see subsection 4.1), or more generally if the cycles are "uniquely solvable" (see [3, 12]).*

We are now going to define the actual random variables $(X_u)_{u \in V^+}$ attached to any mSCM.

Remark 2.6. *Let $M = (G^+, \mathcal{X}, \mathbb{P}, g)$ be a* mSCM *with $G^+ = (U \dot\cup V, E^+)$.*

[1] The assumption of independence of the noise variables here is not to be confused with causal sufficiency. The noise variables here might have two or more child nodes and thus can play the role of latent confounders. The independence assumption here also does not restrict the model class. If they were dependent we would just consider them as one variable and use a different graph that encoded this.

[2] Even though we allow for selfloops in the directed graph G^+ we note that the causal mechanisms g_S will depend only on $\mathrm{Pa}^{G^+}(S) \setminus S$, removing the self-dependence on the functional level. Otherwise, the functions g_S would not hold up to a direct interventional interpretation and one would want to replace them with functions that do.

[3] This assumption is only necessary to give the mSCM a "reduced/summarized" form. In practice one could allow for more latent variables and more complex latent structure.

[4] Note that the index set runs over all "observable loops" $S \subseteq V$, $S \in \mathcal{L}(G^+)$, which contains the usual single element sets $S = \{v\}$, which relate to the usual mechanisms $g_{\{v\}}$.

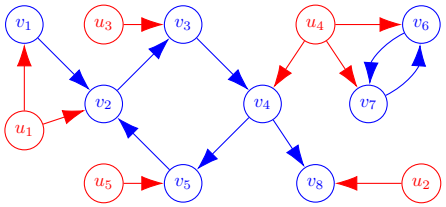

Figure 2: The graph G^+ of a modular structural causal model (mSCM). The observed variables $v_i \in V$ are in blue and the latent variables $u_j \in U$ are in red. We have the four observed strongly connected components: $\{v_1\}$, $\{v_2, v_3, v_4, v_5\}$, $\{v_6, v_7\}$, $\{v_8\}$.

1. *The latent variables $(X_u)_{u\in U}$ are given by the canonical projections $E_u : \prod_{u'\in U} \mathcal{X}_{u'} \to \mathcal{X}_u$ and are jointly $\mathbb{P}$-independent (by 3). Sometimes we will write $(E_u)_{u\in U}$ instead of $(X_u)_{u\in U}$ to stress their interpretation as error/noise variables.*
2. *The observed variables $(X_v)_{v\in V}$ are inductively defined by:*

$$X_v := g_{S,v}\big((X_w)_{w\in \mathrm{Pa}^{G^+}(S)\setminus S}\big),$$

where $S := \mathrm{Sc}^{G^+}(v) := \mathrm{Anc}^{G^+}(v) \cap \mathrm{Desc}^{G^+}(v)$ and where the second index v refers to the v-component of g_S. Note that the inductive definition is possible because when "aggregating" each of the biggest cycles $\mathrm{Sc}^{G^+}(v)$ into one node then only an acyclic graph is left, which can be totally ordered.
3. *By the compatibility condition for g we then have that for every $S \in \mathcal{L}(G^+)$ with $S \subseteq V$ the following equality holds:*

$$X_S = g_S(X_{\mathrm{Pa}^{G^+}(S)\setminus S}),$$

where we put $X_A := (X_v)_{v\in A}$ for subsets A.

As a consequence of the convenient definition 2.4 all the following desirable constructions (like marginalisations and interventions) are easily seen to be well-defined (for proofs see [12]). Note that already defining these constructions was a key challenge in the theory of causal models in the presence of cycles (see [3, 12], a.o.).

Definition 2.7. *Let $M = (G^+, \mathcal{X}, \mathbb{P}, g)$ be a* mSCM *with $G^+ = (U \dot\cup V, E^+)$.*

1. *By plugging the functions g_S into each other we can define the* marginalised mSCM *M' w.r.t. a subset $W \subseteq V$. For example, when marginalizing out $W = \{w\}$ we can define (for the non-trivial case $w \in \mathrm{Pa}^{G^+}(S) \setminus S$):*

$$g_{S',v}(x_{\mathrm{Pa}^{G'}(S')\setminus S'}) := g_{S,v}(x_{\mathrm{Pa}^{G^+}(S)\setminus(S\cup\{w\})}, g_{\{w\}}(x_{\mathrm{Pa}^{G^+}(w)\setminus\{w\}})),$$

where G' is the marginalised graph of G^+, $S' \subseteq V'$ is any loop of G' and S the corresponding induced loop in G^+.

2. *For a subset $I \subseteq V$ and a value $x_I \in \prod_{v\in I} \mathcal{X}_v$ we define the* intervened *graph G' by removing all the edges from parents of I to I. We put $X_u^{\mathrm{do}(x_I)} := X_u$ for $u \in U$ and $X_I^{\mathrm{do}(x_I)} := x_I$ and inductively ($S := \mathrm{Sc}^{G'}(v)$):*

$$X_v^{\mathrm{do}(x_I)} := g_{S,v}(X^{\mathrm{do}(x_I)}_{\mathrm{Pa}^{G'}(S)\setminus S}).$$

By selecting all functions g_S where S is still a loop in the intervened graph G' we get the post-interventional mSCM *M'. These constructions give us all interventional distributions, e.g. (cf. [25]):*

$$\mathbb{P}(X_A|\,\mathrm{do}(x_I), X_B) := \mathbb{P}(X_A^{\mathrm{do}(x_I)}|X_B^{\mathrm{do}(x_I)}).$$

Instead of fixing $X_I^{\mathrm{do}(x_I)}$ to a value x_I we could also specify a distribution $\mathbb{P}'_I$ for it ("randomization"). In this way we define stochastic interventions $\mathrm{do}(\xi_I)$ with an independent random variable ξ_I taking values in $\mathcal{X}_I$ and get a $\mathbb{P}^{\mathrm{do}(I)}$ similarly.

2.2 Σ-SEPARATION IN MSCMS AND Σ-CONNECTION GRAPHS

We now introduce σ-separation as a generalization of d-separation directly on the level of mixed graphs. To make the definition stable under marginalisation and conditioning we need to carry extra structure. The resulting graphs will be called σ-connection graphs (σ-CG), where the name is inspired by [19]. An example that shall clarify the difference between d- and σ-separation is given later in figure 2 and table 1.

Definition 2.8 (σ-Connection Graphs (σ-CG))**.** *A σ-connection graph (σ-CG) is a mixed graph G with a set of nodes V and directed ($\to$), undirected ($-$) and bidirected ($\leftrightarrow$) edges, together with an equivalence relation $\sim_\sigma$ on V that has the property that every equivalence class $\sigma(v)$, $v \in V$, is a loop in the underlying directed graph structure: $\sigma(v) \in \mathcal{L}(G)$. Undirected self-loops ($v - v$) are allowed, (bi)-directed self-loops ($v \to v$, $v \leftrightarrow v$) are not.*

In particular, every node is assigned to a unique fixed loop $\sigma(v)$ in G with $v \in \sigma(v)$ and two of such loops $\sigma(v_1)$, $\sigma(v_2)$ are either identical or disjoint. The reason for why we need such structure is illustrated in figure 5.

Definition 2.9 (σ-Open Path in a σ-CG)**.** *Let G be a σ-CG with set of nodes V and $Z \subseteq V$ a subset. Consider*

a path π in G with $n \geq 1$ nodes:

$$v_1 \;\overleftrightarrows\; \cdots \;\overleftrightarrows\; v_n.^{5}$$

*The path will be called Z-σ-*open *if:*

1. *the* endnodes $v_1, v_n \notin Z$, *and*
2. *every triple of adjacent nodes in π that is of the form:*
   (a) collider*:*
   $$v_{i-1} \;\overleftrightarrow{\blacktriangleright}\; v_i \;\overleftrightarrow{\blacktriangleleft}\; v_{i+1},$$
   satisfies $v_i \in Z$,
   (b) (non-collider) left chain*:*
   $$v_{i-1} \leftarrow v_i \;\overleftrightarrows\; v_{i+1},$$
   satisfies $v_i \notin Z$ or $v_i \in Z \cap \sigma(v_{i-1})$,
   (c) (non-collider) right chain*:*
   $$v_{i-1} \;\overleftrightarrows\; v_i \rightarrow v_{i+1},$$
   satisfies $v_i \notin Z$ or $v_i \in Z \cap \sigma(v_{i+1})$,
   (d) (non-collider) fork*:*
   $$v_{i-1} \leftarrow v_i \rightarrow v_{i+1},$$
   satisfies $v_i \notin Z$ or $v_i \in Z \cap \sigma(v_{i-1}) \cap \sigma(v_{i+1})$,
   (e) (non-collider) with undirected edge*:*
   $$v_{i-1} \;\text{—}\; v_i \;\overleftrightarrows\; v_{i+1},$$
   $$v_{i-1} \;\overleftrightarrows\; v_i \;\text{—}\; v_{i+1},$$
   satisfies $v_i \notin Z$.

The difference between σ- and d-separation lies in the additional conditions involving $Z \cap \sigma(v_{i\pm1})$. The intuition behind them is that the dependence structure *inside* a loop $\sigma(v_i)$ is so strong that non-colliders can only be blocked by conditioning if an edge is pointing out of the loop (see example 2.17 and table 1).

Similar to d-separation we can now define σ-separation in a σ-CG.

Definition 2.10 (σ-Separation in a σ-CG)**.** *Let G be a σ-CG with set of nodes V. Let $X, Y, Z \subseteq V$ be subsets.*

1. *We say that X and Y are* σ-connected by Z or not σ-separated by Z *if there exists a path π (with some $n \geq 1$ nodes) in G with one endnode in X and one endnode in Y that is Z-σ-open. In symbols this statement will be written as follows:*
   $$X \overset{\sigma}{\not\perp\!\!\!\perp}_{G} Y \mid Z.$$
2. *Otherwise, we will say that X and Y are* σ-separated *by Z and write:*
   $$X \overset{\sigma}{\perp\!\!\!\perp}_{G} Y \mid Z.$$

Remark 2.11. *1. The* finest/trivial *σ-CG structure of a mixed graph G is given by $\sigma(v) := \{v\}$ for all $v \in V$. In this way σ-separation in G coincides with the usual notion of d-separation in a d-connection graph (d-CG) G (see [19]). We will take this as the definition of d-separation and d-CG in the following.*

2. *The* coarsest *σ-CG structure of a mixed graph G is given by $\sigma(v) := \mathrm{Sc}^G(v) := \mathrm{Anc}^G(v) \cap \mathrm{Desc}^G(v)$ w.r.t. the underlying directed graph. Note that the definition of strongly connected component totally ignores the bi- and undirected edges of the σ-CG.*

3. *In any σ-CG we will always have that σ-separation implies d-separation, since every Z-d-open path is also Z-σ-open because $\{v\} \subseteq \sigma(v)$.*

4. *If a σ-CG G is acyclic (implying $\mathrm{Sc}^G(v) = \{v\}$) then σ-separation coincides with d-separation.*

We now want to "hide" or marginalise out the latent nodes $u \in U$ from the graph of any mSCM and represent their induced dependence structure with bidirected edges.

Definition 2.12 (Induced σ-CG of a mSCM)**.** *Let $M = (G^+, \mathcal{X}, \mathbb{P}, g)$ be a* mSCM *with $G^+ = (U \dot\cup V, E^+)$. The* induced σ-CG G of M*, also referred to as the* causal graph *G of M is defined as follows:*

1. *The nodes of G are all $v \in V$, i.e. all observed nodes of G^+.*
2. *G contains all the directed edges of G^+ whose endnodes are both in V, i.e. observed.*
3. *G contains the bidirected edge $v \leftrightarrow w$ with $v, w \in V$ if and only if $v \neq w$ and there exists a $u \in U$ with $v, w \in \mathrm{Ch}^{G^+}(u)$, i.e. v and w have a common latent confounder.*
4. *G contains no undirected edges.*
5. *We put $\sigma(v) := \mathrm{Sc}^G(v) = \mathrm{Anc}^G(v) \cap \mathrm{Desc}^G(v)$.*

Remark 2.13. *Caution must be applied when going from G^+ to G: It is possible that three observed nodes v_1, v_2, v_3 have one joint latent common cause u_1, which can be read off G^+. This information will get lost when going from G^+ to G, as we will represent this with three bidirected edges. G will nonetheless capture the conditional independence relations (see Theorem 2.14).*

[5] The stacked edges are meant to be read as an "OR" at each place independently. We also allow for repeated nodes in the paths.

We now present the most important ingredient for our constraint-based causal discovery algorithm, namely a generalized directed global Markov property that relates the underlying causal graph (σ-CG) G of any mSCM M to the conditional independencies of the observed random variables $(X_v)_{v\in V}$ via a σ-separation criterion.

Theorem 2.14 (σ-Separation Criterion, see Corollary B.3)**.** *The observed variables $(X_v)_{v\in V}$ of any* mSCM *M satisfy the* σ-separation criterion *w.r.t. the induced σ-CG G. In other words, for all subsets $W, Y, Z \subseteq V$ we have the implication:*

$$W \overset{\sigma}{\underset{G}{\perp\!\!\!\perp}} Y \,|\, Z \implies X_W \underset{\mathbb{P}}{\perp\!\!\!\perp} X_Y \,|\, X_Z.$$

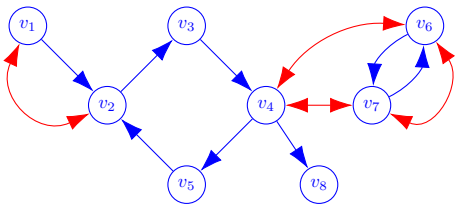

Figure 3: The induced σ-connection graph (causal graph) of the modular structural causal model (mSCM) of figure 2, with σ-equivalence classes $\{\{v_1\}, \{v_2, v_3, v_4, v_5\}, \{v_6, v_7\}, \{v_8\}\}$.

Table 1: d- and σ-separation in the σ-connection graph G from figure 3. σ-separation implies d-separation, but i.g. d-separation encodes more conditional independencies than σ-separation:

d-separation	σ-separation
$\{v_2\} \perp\!\!\!\perp^d_G \{v_4\} \,\|\, \{v_3, v_5\}$	$\{v_2\} \not\perp\!\!\!\perp^\sigma_G \{v_4\} \,\|\, \{v_3, v_5\}$
$\{v_1\} \perp\!\!\!\perp^d_G \{v_6\}$	$\{v_1\} \perp\!\!\!\perp^\sigma_G \{v_6\}$
$\{v_1\} \perp\!\!\!\perp^d_G \{v_6\} \,\|\, \{v_3, v_5\}$	$\{v_1\} \not\perp\!\!\!\perp^\sigma_G \{v_6\} \,\|\, \{v_3, v_5\}$
$\{v_1\} \not\perp\!\!\!\perp^d_G \{v_8\}$	$\{v_1\} \not\perp\!\!\!\perp^\sigma_G \{v_8\}$
$\{v_1\} \perp\!\!\!\perp^d_G \{v_8\} \,\|\, \{v_3, v_5\}$	$\{v_1\} \not\perp\!\!\!\perp^\sigma_G \{v_8\} \,\|\, \{v_3, v_5\}$
$\{v_1\} \perp\!\!\!\perp^d_G \{v_8\} \,\|\, \{v_4\}$	$\{v_1\} \perp\!\!\!\perp^\sigma_G \{v_8\} \,\|\, \{v_4\}$

If we want to infer the causal graph (σ-CG) G of a mSCM from data with help of conditional independence tests in practice we usually need to assume also the reverse implication of the σ-separation criterion from 2.14 for the observational distribution $\mathbb{P}(X_V)$ and the relevant interventional distributions $\mathbb{P}(X_V | \operatorname{do}(\xi_I))$ (see 2.7). This will be called *σ-faithfulness*.

Definition 2.15 (σ-faithfulness)**.** *We will say that the tuple $(G, \mathbb{P})$ is* σ-faithful *if for every three subsets $W, Y, Z \subseteq V$ we have the equivalence:*

$$W \overset{\sigma}{\underset{G}{\perp\!\!\!\perp}} Y \,|\, Z \iff X_W \underset{\mathbb{P}}{\perp\!\!\!\perp} X_Y \,|\, X_Z.$$

Remark 2.16 (Strong σ-completeness, cf. [22])**.** *We do believe that the* generic *non-linear mSCM is σ-faithful to G, as the needed conditional dependence structure in all our simulated (sufficiently non-linear) cases was observed (cf. example 2.17). But proving such a* strong σ-completeness *result is difficult (and to our knowledge only done for multinomial and linear Gaussian DAG models, see [22]) and left for future research. Further note that the class of linear models, which even follow d-separation, would i.g. not be considered σ-faithful. Since linear models are of measure zero in the bigger class of general mSCMs this would not contradict our conjecture.*

Example 2.17. *Consider a directed four-cycle, e.g. the subgraph $\{v_2, v_3, v_4, v_5\}$ from figure 2, where all other observed nodes are assumed to be absent. Consider only the non-linear causal mechanisms given by ($i = 3, 4, 5$):*

$$\begin{aligned} g_{\{v_2\}}(X_5, E_1) &:= \tanh(0.9 \cdot X_5 + 0.5) + E_1, \\ g_{\{v_i\}}(X_{i-1}, E_i) &:= \tanh(0.9 \cdot X_{i-1} + 0.5) + E_i, \end{aligned}$$

where E_1, E_3, E_4, E_5 are assumed to be independent. The equations $X_i = g_{\{v_i\}}(X_{i-1}, E_i)$ and the one for X_2 will imply the conditional dependence

$$X_2 \underset{\mathbb{P}}{\not\perp\!\!\!\perp} X_4 \,|(X_3, X_5),$$

which can be checked by computations and/or simulations. As one can read off table 1, d-separation fails to express the dependence, in contrast to σ-separation, which captures it correctly.

2.3 MARGINALISATION AND CONDITIONING IN Σ-CONNECTION GRAPHS

Inspired by [19] we will define marginalisation and conditioning operations on σ-connection graphs (σ-CG) and prove the closedness of σ-separation (and thus its criterion) under these operations. These are key results to extend the algorithm of [19] to the setting of mSCMs.

Definition 2.18 (Marginalisation of a σ-CG)**.** *Let G be a σ-CG with set of nodes V and $w \in V$, $W := \{w\}$. We define the* marginalised σ-CG G^W *with set of nodes $V \setminus W$ via the rules for $v_1, v_2 \in V \setminus W$:*
$v_1 \overset{a}{\circ\!\!-\!\!\circ}{}_b\, v_2 \in G^W$ with arrow heads/tails a and b if and only if there exists:

1. $v_1 \overset{a}{\circ\!\!-\!\!\circ}{}_b\, v_2$ *in G, or*
2. $v_1 \overset{a}{\circ\!\!-\!\!\circ}{}\, w \overset{q}{\circ\!\!-\!\!\circ}{}_b\, v_2$ *in G, or*
3. $v_1 \overset{a}{\circ\!\!-\!\!\circ}{}_q\, w \overset{}{\circ\!\!-\!\!\circ}{}_b\, v_2$ *in G, or*
4. $v_1 \overset{a}{\circ\!\!\rightarrow} w \,-\!-\, w \overset{}{\leftarrow\!\!\circ}{}_b\, v_2$ *in G.*

Note that directed paths in G have no colliders, so loops in G map to loops in G^W (if not empty). Thus we have the induced σ-CG structure $\sim_\sigma$ on G^W.

Definition 2.19 (Conditioning of a σ-CG)**.** *Let G be a σ-CG with set of nodes V and $c \in V$, $C := \{c\}$. We define the* conditioned σ-CG G_C *with set of nodes $V \setminus C$ via the rules for $v_1, v_2 \in V \setminus C$:*
$v_1 \; {}_a\!\circ\!\!-\!\!\circ_b \; v_2 \in G_C$ if and only if there exists:

1. $v_1 \; {}_a\!\circ\!\!-\!\!\circ_b \; v_2$ *in G, or*
2. $v_1 \; {}_a\!\circ\!\!\rightarrow c \leftarrow\!\!\circ_b \; v_2$ *in G, or*
3. $v_1 \; {}_a\!\leftarrow c \leftarrow\!\!\circ_b \; v_2$ *in G, $\sigma(v_1) = \sigma(c)$, or*
4. $v_1 \; {}_a\!\circ\!\!\rightarrow c \rightarrow_b \; v_2$ *in G, $\sigma(c) = \sigma(v_2)$, or*
5. $v_1 \; {}_a\!\leftarrow c \rightarrow_b \; v_2$ *in G, $\sigma(v_1) = \sigma(c) = \sigma(v_2)$.*

Note that directed paths in $\sigma(v)$ in G condition to directed paths, so loops in $\sigma(v)$ in G map to loops in G_C (if not empty). Thus we have a well-defined induced σ-CG structure $\sim_\sigma$ on G_C.

The proofs of the following theorem, stating the closedness of σ-separation under marginalisation and conditioning, can be found in the supplementary material (Theorem A.1 and Theorem A.2). See also figure 5.

Theorem 2.20. *Let G be a σ-CG with set of nodes V and $X, Y, Z \subseteq V$ any subsets. For any nodes $w, c \in V \setminus (X \cup Y \cup Z)$, $W := \{w\}$, $C := \{c\}$, we then have the equivalences:*

$$X \overset{\sigma}{\underset{G^W}{\perp\!\!\!\perp}} Y \mid Z \iff X \overset{\sigma}{\underset{G}{\perp\!\!\!\perp}} Y \mid Z,$$
$$X \overset{\sigma}{\underset{G_C}{\perp\!\!\!\perp}} Y \mid Z \iff X \overset{\sigma}{\underset{G}{\perp\!\!\!\perp}} Y \mid Z \cup C.$$

Corollary 2.21. *Let G be a σ-CG with set of nodes V and $X, Y, Z \subseteq V$ pairwise disjoint subsets and $W := V \setminus (X \cup Y \cup Z)$. Then we have the equivalence:*

$$X \overset{\sigma}{\underset{G}{\perp\!\!\!\perp}} Y \mid Z \iff X \overset{\sigma}{\underset{G_Z^W}{\perp\!\!\!\perp}} Y,$$

where G_Z^W is any σ-CG with set of nodes $X \cup Y$ obtained by marginalising out all the nodes from W and conditioning on all the nodes from Z in any order. This means that if $X = \{x\}$ and $Y = \{y\}$ then x and y are σ-separated by Z in G if and only if x and y are not connected by any edge in the σ-CG G_Z^W.

It is also tempting to introduce an intervention operator directly on the level of σ-CGs. However, since the interplay between conditioning and intervention is complicated (e.g. they do not commute i.g.) we do not investigate this further in this paper. The intervention operator on the level of mSCMs will be enough for our purposes as we assume no pre-interventional selection bias and then only encounter observational or post-interventional conditioning, which is covered by our framework.

3 ALGORITHM

In this section, we propose an algorithm for causal discovery that is based on the theory in the previous section. Given that theory, our proposed algorithm is a straightforward modification of the algorithm by [19]. The main idea is to formulate the causal discovery problem as an optimization problem that aims at finding the causal graph that best matches the data at hand. This is done by encoding the rules for conditioning, marginalisation, and intervention (see below) on a σ-CG into Answer Set Programming (ASP), an expressive declarative programming language based on stable model semantics that supports optimization [15, 20]. The optimization problem can then be solved by employing an off-the-shelf ASP solver.

3.1 CAUSAL DISCOVERY WITH Σ-CONNECTION GRAPHS

Let $M = (G^+, \mathcal{X}, \mathbb{P}, g)$ be a mSCM with $G^+ = (U \dot\cup V, E^+)$ and $I \subseteq V$ a subset. Consider a (stochastic) perfect intervention $\mathrm{do}(\xi_I)$ that enforces $X_I = \xi_I$ for an independent random variable ξ_I taking values in $\mathcal{X}_I$. Denote the (unique) induced distribution of the intervened mSCM $M_{\mathrm{do}(\xi_I)}$ by $\mathbb{P}^{\mathrm{do}(I)}$, and the causal graph (i.e., induced σ-CG of the intervened mSCM on the observed variables) by $G_{\mathrm{do}(I)} = (G^+_{\mathrm{do}(I)})^U$.

Under $\mathbb{P}^{\mathrm{do}(I)}$, the observed variables $(X_v)_{v \in V}$ satisfy the σ-separation criterion w.r.t. $G_{\mathrm{do}(I)}$ by Theorem 2.14. For the purpose of causal discovery, we will in addition assume σ-faithfulness (Definition 2.15), i.e., that each conditional independence between observed variables is due to a σ-separation in the causal graph. Taken together, and by applying Corollary 2.21, we get for all subsets $W, Y, Z \subseteq V$ the equivalences:

$$\begin{aligned} X_W \underset{\mathbb{P}^{\mathrm{do}(I)}}{\perp\!\!\!\perp} X_Y \mid X_Z \iff & W \overset{\sigma}{\underset{G_{\mathrm{do}(I)}}{\perp\!\!\!\perp}} Y \mid Z \\ \iff & W \overset{\sigma}{\underset{(G_{\mathrm{do}(I)})_Z^{V \setminus W \cup Y \cup Z}}{\perp\!\!\!\perp}} Y. \end{aligned} \tag{1}$$

If $W = \{w\}$ and $Y = \{y\}$ consist of a single node each, the latter can be easily checked by testing whether w is non-adjacent to y in $(G_{\mathrm{do}(I)})_Z^{V \setminus W \cup Y \cup Z}$.

3.2 CAUSAL DISCOVERY AS AN OPTIMIZATION PROBLEM

Following [19], we formulate causal discovery as an optimization problem where a certain loss function is optimized over possible causal graphs. This loss function sums the weights of all the inputs that are violated assuming a certain underlying causal graph.

The input for the algorithm is a list $S = \big((w_j, y_j, Z_j, I_j, \lambda_j)\big)_{j=1}^{n}$ of weighted conditional independence statements. Here, the weighted statement $(w_j, y_j, Z_j, I_j, \lambda_j)$ with $w_j, y_j \in V$, $Z_j, I_j \subseteq V$, and $\lambda_j \in \overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, +\infty\}$ encodes that $X_{w_j} \perp\!\!\!\perp_{\mathbb{P}^{\mathrm{do}(X_{I_j})}} X_{y_j} \,|\, X_{Z_j}$ holds with confidence λ_j, where a finite value of λ_j gives a "soft constraint" and a value of $\lambda_j = \pm\infty$ imposes a "hard constraint". Positive weights encode that we have empirical support *in favor* of the independence, whereas negative weights encode empirical support *against* the independence (in other words, in favor of *dependence*).

As in [19], we define a loss function that measures the amount of evidence *against* the hypothesis that the data was generated by an mSCM with causal graph G, by simply summing the absolute weights of the input statements that conflict with G under the σ-Markov and σ-faithfulness assumptions:

$$\mathcal{L}(G, S) := \sum_{(w_j, y_j, Z_j, I_j, \lambda_j) \in S} \lambda_j (\mathbb{1}_{\lambda_j > 0} - \mathbb{1}_{w_j \perp\!\!\!\perp^{\sigma}_{G_{\mathrm{do}(I_j)}} y_j \,|\, Z_j}) \tag{2}$$

where $\mathbb{1}$ is the indicator function. This loss function differs from the one used in [19] in that we use σ-separation instead of d-separation. Causal discovery can now be formulated as the optimization problem:

$$G^* = \operatorname*{arg\,min}_{G \in \mathbb{G}(V)} \mathcal{L}_R(G, S) \tag{3}$$

where $\mathbb{G}(V)$ denotes the set of all possible causal graphs with variables V.

The optimization problem (3) may have multiple optimal solutions, because the underlying causal graph may not be identifiable from the inputs. Nonetheless, some of the features of the causal graph (e.g., the presence or absence of a certain directed edge) may still be identifiable. We employ the method proposed by [21] for scoring the confidence that a certain feature f is present by calculating the difference between the optimal losses under the additional hard constraints that the feature f is present vs. that the feature f is absent in G.

In our experiments, we will use the weights proposed in [21]: $\lambda_j = \log p_j - \log \alpha$, where p_j is the p-value of a statistical test with independence as null hypothesis, and α is a significance level (e.g., 1%). This test is performed on the data measured in the context of the (stochastic) perfect intervention I_j. These weights have the desirable property that independences typically get a lower absolute weight than strong dependences. For the conditional independence test, we use a standard partial correlation test after marginal rank-transformation of the data so as to obtain marginals with standard-normal distributions.

3.3 FORMULATING THE OPTIMIZATION PROBLEM IN ASP

In order to calculate the loss function (2), we make use of Corollary 2.21 to reduce the σ-separation test to a simple non-adjacency test in a conditioned and marginalised σ-CG, as in (1). We do this by encoding σ-CGs, Theorem 2.20 and the marginalisation and conditioning operations on σ-CGs (Definitions 2.18 and 2.19) in ASP. The details of the encoding are provided in the Supplementary Material.[6] The optimization problem in (3) can then be solved straightforwardly by running an off-the-shelf ASP solver with as input the encoding and the weighted independence statements.

A more precise statement of the following result is provided in the Supplementary Material. The proof is basically the same as the one given in [21].

Theorem 3.1. *The algorithm for scoring features f is sound for oracle inputs and asymptotically consistent under mild assumptions.*

4 EXPERIMENTS

4.1 CONSTRUCTING MSCMS AND SAMPLING FROM MSCMS

To construct a modular structural causal model (mSCM) in practice we need to specify the compatible system of functions $(g_S)_{S \in \mathcal{L}(G)}$. The following Theorem is helpful (and a direct consequence of Banach's fixed point theorem).

Theorem 4.1. *Consider the functions $g_{\{v\}}$ for the trivial loops $\{v\} \in \mathcal{L}(G)$, $v \in V$ and assume the following contractivity condition:*

> *For every non-trivial loop $S \in \mathcal{L}(G)$ and for every value $x_{\mathrm{Pa}^{G^+}(S) \setminus S}$ the multi-dimensional function:*
>
> $$x'_S \mapsto \Big(g_{\{v\}}(x'_{S \cap \mathrm{Pa}^{G^+}(v) \setminus \{v\}}, x_{\mathrm{Pa}^{G^+}(v) \setminus S})\Big)_{v \in S}$$
>
> *is a contraction, i.e. Lipschitz continuous with Lipschitz constant $L(x_{\mathrm{Pa}^{G^+}(S) \setminus S}) < 1$ w.r.t. a suitable norm $||\cdot||$.*

Then all the functions g_S for the non-trivial loops $S \in \mathcal{L}(G)$ exist, are unique and $g = (g_S)_{S \in \mathcal{L}(G)}$ forms a

[6] The full source code for the algorithm and to reproduce our experiments is available under an open source license from `https://github.com/caus-am/sigmasep`.

globally compatible system.
More constructively, for every value $x_{\mathrm{Pa}^{G^+}(S)\setminus S}$ *and initialization* $x_S^{(0)}$ *the iteration scheme (using vector notations):*

$$x_S^{(t+1)} := (g_{\{v\}})_{v\in S}(x_S^{(t)}, x_{\mathrm{Pa}^{G^+}(S)\setminus S})$$

converges to a unique limit vector x_S *(for* $t \to \infty$ *and independent of* $x_S^{(0)}$*). g_S is then given by putting:*

$$g_S(x_{\mathrm{Pa}^{G^+}(S)\setminus S}) := x_S.$$

This provides us with a method for constructing very general non-linear mSCMs (e.g. neural networks, see Section C in Supplementary Material) and to sample from them: by sampling x_U from the external distribution and then apply the above iteration scheme until convergence for all loops, yielding the limit x_V as one data point.

4.2 RESULTS ON SYNTHETIC DATA

In our experiments we will—due to computational restrictions—only allow for $d = 5$ observed nodes and $k = 2$ additional latent confounders. We sample edges independently with a probability of $p = 0.3$. We model the non-linear function $g_{\{v\}}$ as a neural network with $\tanh$ activation, bias terms that have a normal distribution with mean -0.5 and standard deviation 0.2, and weights sampled uniformly from the L1-unit ball to satisfy the contraction condition of Theorem 4.1 (see also Supplementary Material, Section C). We simulate 0–5 single-variable interventions with random (unique) targets. For each intervened model we sample from standard-normal noise terms and compute the observations. To also detect weak dependencies in cyclic models we allow for $n = 10^4$ samples in each such model for each allowed intervention. We then run all possible conditional independence tests between every pair of single nodes and calculate their p-values. We used $\alpha = 10^{-3}$ as the threshold between dependence and independence. For computational reasons we restrict to partial correlation tests of marginal Gaussian rank-transforms of the data. These tests are then fed into the ASP solver together with our encoding of the optimization problem (3). We query the ASP solver for the confidence for the absence or presence of each possible directed and bidirected edge. We simulate 300 models and aggregate results, using the confidence scores to compute ROC- and PR-curves for features. Figure 4 shows that, as expected, our algorithm recovers more directed edges of the underlying causal graph in the simulation setting as the number of single-variable interventions increases. More results (ROC- and PR-curves for directed edges and confounders for different numbers of single-variable interventions and for different encodings) are provided in the Supplementary Material.

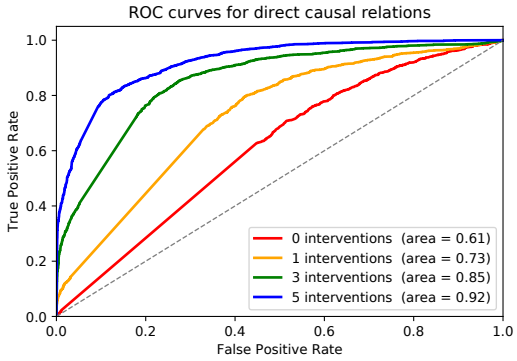

Figure 4: The ROC curves for identifying directed edges. See also figures 6 and 7 in the Supplementary Material.

5 CONCLUSION

We introduced σ-connection graphs (σ-CG) as a generalization of the d-connection graphs (d-CG) of [19] and extended the notion of σ-separation that was introduced in [12] to σ-CGs. We showed how σ-CGs behave under marginalisation and conditioning. This provides a graphical representation of how conditional independencies of modular structural causal models (mSCMs) behave under these operations. We provided a sufficient condition that allows constructing mSCMs and sampling from them. We extended the algorithm of [19] to deal with the more generally applicable notion of σ-separation instead of d-separation, thereby obtaining the first algorithm for causal discovery that can deal with cycles, nonlinearities, latent confounders and a combination of data sets corresponding to observational and different interventional settings. We illustrated the effectiveness of the algorithm on simulated data. In this work, we restricted attention to (stochastic) perfect ("surgical") interventions, but a straightforward extension to deal with other types of interventions and to generalize the idea of randomized controlled trials can be obtained by applying the JCI framework [23]. In future work we wish to improve our algorithm to also handle selection bias, become more scalable and apply it to real world data sets.

Acknowledgements

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

[1] F. Barthe, O. Guédon, S. Mendelson, and A. Naor. A probabilistic approach to the geometry of the l_p^n-ball. *Ann. Probab.*, 33(2):480–513, 2005.

[2] V.I. Bogachev. *Measure Theory. Vol. I, II*. Springer, 2007.

[3] S. Bongers, J. Peters, B. Schölkopf, and J. M. Mooij. Theoretical aspects of cyclic structural causal models. *arXiv.org preprint*, arXiv:1611.06221v2 [stat.ME], 2018.

[4] T. Claassen and T. Heskes. Bayesian probabilities for constraint-based causal discovery. In *IJCAI-13*, pages 2992–2996, 2013.

[5] T. Claassen, J.M. Mooij, and T. Heskes. Learning Sparse Causal Models is not NP-hard. In *UAI-13*, pages 172–181, 2013.

[6] G. F. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.

[7] M. Drton, M. Eichler, and T. Richardson. Computing maximum likelihood estimates in recursive linear models with correlated errors. *Journal of Machine Learning Research*, 10:2329–2348, 2009.

[8] R. Evans and T. Richardson. Maximum likelihood fitting of acyclic directed mixed graphs to binary data. In *UAI-10*, 2010.

[9] R.J. Evans. Graphs for margins of Bayesian networks. *Scand. J. Stat.*, 43(3):625–648, 2016.

[10] R.J. Evans. Margins of discrete Bayesian networks. *arXiv:1501.02103*, pages 1–41, 2017. Submitted to Annals of Statistics.

[11] R.J. Evans and T.S. Richardson. Markovian acyclic directed mixed graphs for discrete data. *Ann. Statist.*, 42(4):1452–1482, 08 2014.

[12] P. Forré and J. M. Mooij. Markov properties for graphical models with cycles and latent variables. arXiv:1710.08775, 2017.

[13] M. Gebser, R. Kaminski, B. Kaufmann, and T. Schaub. *Clingo* = ASP + control: Extended report. Technical report, University of Potsdam, 2014. `http://www.cs.uni-potsdam.de/wv/pdfformat/gekakasc14a.pdf`.

[14] D. Geiger and D. Heckerman. Learning Gaussian networks. In *UAI-94*, pages 235–243, 1994.

[15] M. Gelfond. Answer sets. In *Handbook of Knowledge Representation*, pages 285–316. 2008.

[16] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[17] D. Heckerman, C. Meek, and G. Cooper. A Bayesian approach to causal discovery. In C. Glymour and G. F. Cooper, editors, *Computation, Causation, and Discovery*, pages 141–166. MIT Press, 1999.

[18] A. Hyttinen, F. Eberhardt, and P.O. Hoyer. Causal discovery for linear cyclic models. In *Proceedings of the Fifth European Workshop on Probabilistic Graphical Models*, 2010.

[19] A. Hyttinen, F. Eberhardt, and M. Järvisalo. Constraint-based causal discovery: Conflict resolution with answer set programming. In *UAI-14*, pages 340–349, 2014.

[20] V. Lifschitz. What is Answer Set Programming? In *AAAI Conference on Artificial Intelligence*, pages 1594–1597, 2008.

[21] S. Magliacane, T. Claassen, and J.M. Mooij. Ancestral causal inference. In *NIPS-16*, pages 4466–4474. 2016.

[22] C. Meek. Strong completeness and faithfulness in Bayesian networks. In *UAI-95*, pages 411–418, 1995.

[23] J. M. Mooij, S. Magliacane, and T. Claassen. Joint causal inference from multiple contexts. *arXiv.org preprint*, https://arxiv.org/abs/1611.10351v3 [cs.LG], March 2018.

[24] J. Pearl. Fusion, propagation and structuring in belief networks. Technical Report 3, UCLA Computer Science Department, 1986. Technical Report 850022 (R-42).

[25] J. Pearl. *Causality: Models, reasoning, and inference*. Cambridge University Press, 2nd edition, 2009.

[26] J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference: Foundation and Learning Algorithms*. MIT press, 2017.

[27] T. Richardson. A discovery algorithm for directed cyclic graphs. In *UAI-96*, pages 454–461. 1996.

[28] T. Richardson. Markov properties for acyclic directed mixed graphs. *Scand. J. Stat.*, 30(1):145–157, 2003.

[29] T. Richardson and P. Spirtes. Automated discovery of linear feedback models. In C. Glymour and G. F. Cooper, editors, *Computation, Causation, and Discovery*, pages 253–304. MIT Press, 1999.

[30] D. Rothenhäusler, C. Heinze, J. Peters, and N. Meinshausen. BACKSHIFT: Learning causal cyclic graphs from unknown shift interventions. In *NIPS-15*, pages 1513–1521. 2015.

[31] P. Spirtes. Directed cyclic graphical representations of feedback models. In *UAI-95*, pages 491–499, 1995.

[32] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, 2000.

[33] T.S. Verma and J. Pearl. Causal Networks: Semantics and Expressiveness. *UAI-90*, 4:69–76, 1990.

[34] J. 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.