Proceedings of Machine Learning Research vol 140:1–16, 2022 1st Conference on Causal Learning and Reasoning

Local Constraint-Based Causal Discovery under Selection Bias

Author names withheld

Editor: Under Review for CLeaR 2022

Abstract

We consider the problem of discovering causal effects from independence constraints when both selection bias and confounding are present. While the seminal FCI algorithm is sound and complete in this setup, no criterion for interpreting causal relations from its PAG output under selection bias is presently known. We instead focus on local constraint-based patterns between three or four variable patterns. The LCD algorithm is seen to fail in these conditions, while the four variable (Extended) Y-Structure is shown to be asymptotically sound in predicting causal effects when selection bias and latent confounding are present. We introduce a finite-sample scoring rule for Y-Structures that is shown to successfully predict ancestral causal effects in simulation experiments where selection bias is introduced. In a real-world setting, we show how Y-Structures perform well across different datasets, potentially circumventing spurious correlations due to selection bias.

Keywords: causal discovery, causal modeling, causal inference, observational and experimental data, selection bias

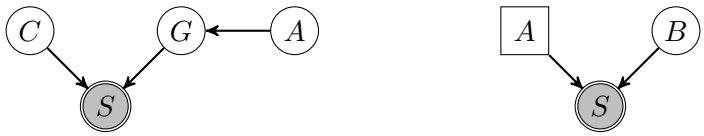
1. Introduction

The discovery of causal effects from data is central to many disciplines in science such as biology, economics and psychology. The goal of predicting the effect of new unseen interventions is vital when experiments are expensive, unethical or unfeasible. Algorithms that predict causal effects have been developed for various settings that match conditions found in the data such as including cyclic causal relations and allowing for latent confounding.

Selection bias is the preferential exclusion of samples in the data gathering process. This results in a set of samples that is not a representation of the underlying causal model, severely hindering methods for causal inference and discovery. In practice, the hypothesis that a selection bias mechanism is affecting samples at-hand is often disregarded as a possibility for convienience. Consider the following example, graphically represented in Fig 1(a).

Example 1 ***(Data-Driven Professor)*** *A professor tries to entice students to attend her course through a policy of handing out cookies C whenever she interacts with students. After the final exam, she hands out the survey sheets S, where students can score the quality of the class. Unbeknownst to the professor, students get good grades G as a direct causal effect of their innate abilities A (such as studying capacity and intelligence), and only students with good grades actually hand-in the class review sheet, which is privacy-sensitive and unobservable to the professor. In the next year, the professor updates her cookie policy C by increasing the number of cookies provided. Disregarding the possibility of selection bias, she observes a spurious improvement in the final grades G of those that handed in the exam. Additionally, she would even observe a spurious causal relation of her cookie policy C on the student's innate abilities A.*

© 2022 .



(a) Example. 1 (b) Example. 2

Figure 1: **(Examples)** Causal graph accompanying the examples in the main text. Here selection bias is represented by $\boldsymbol{S}$ unobserved.

Here, C and G are spuriously correlated by selection bias on S, and is detectable in a correlation. Crucially, C and A are additionally correlated in the data under selection, which can unknowingly lead to detecting causal effects when there are none or even lead to causal discovery in the wrong direction. Fig. 1(b) accompanies the following example.

Example 2 ***(Gene Regulatory Network)*** *Genes, the constituent that make up a gene regulatory network, influence (regulate) other genes by changing their expression level. In experiments, for some individual genes A are disabled, gene samples are grown and the expressions of another set genes B is measured to detect causal relations. However, the experimental growth process might destroy some of these genes, whereas others have a higher survivability S due to a redundant growth mechanism. This mechanism, which is unmeasured, can subsequently induce a spurious causal effect of A on B.*

In this example we see that selection can lead to spurious correlations in the data, even in the post-interventional distribution. While bias due to latent confounding can be mitigated through randomization, this is not the case for data under selection. In fact, Fig. 2 represents a *Randomized Controlled Trial* (RCT), where the randomization of A induces a spurious causal effect of B. There are however algorithms for causal discovery developed for this regime.

The seminal FCI algorithm (Spirtes et al., 1995, 1999) is famously sound and complete under both confounding and selection bias. It returns a PAG, an equivalence class of all the set of independence relations in the data represented by a set of ancestral graphs (Zhang, 2008a). When selection bias is excluded, a criterion for reading off ancestral causal effects and confounding was conjectured in Zhang (2006) and Zhang (2008a), shown to be sound in Roumpelaki et al. (2016) and generalized to cycles in Mooij and Claassen (2020). However, a procedure for extracting (non)ancestral relations from this equivalence class when selection bias is a possibility is yet unknown.

Contributions In this work, we investigate how constraint-based causal discovery is affected under selection bias in addition to latent confounding. We take a bottom-up approach and investigate local methods that reason from (in)dependence relations between small groups of three or four variables, namely LCD (Cooper, 1997) and Y-Structures (Mani et al., 2006; Mooij and Cremers, 2015). These algorithms contain simpler causal semantics, allowing for a more direct interpretation of causal information of the output under selection bias. Moreover, less reasoning over combinations of statistical testing output is required, reducing the risk of propagating statistical errors. Lastly, these approaches have a computational advantage, trading off the completeness of global methods such as FCI for speed, and are easier adapted to high-dimensional data by pre-selecting plausible edges.

Our contributions are as follows.

- For three variables, we investigate the failure modes of LCD under the selection mechanism. We find that there are no other constraint-based approaches available for this setting.
- For four variables we find that the (Extended) Y-Structures pattern search is valid under when the data is gathered under selection, and we introduce a finite-sample scoring method.
- We introduce a simulator for empirically assessing model performance under selection and without.
- Using real-world data, we show a successful application of (Extended) Y-Structures, implying that a selection mechanism could plausibly underlay the data at-hand.

Related work The recovery of causal effects from selection bias, where the aim is an inference or estimation of a causal effect from data under (partial) selection when the underlying graphical conditions are known, has been studied extensively in literature (Bareinboim and Pearl, 2012; Bareinboim et al., 2014; Bareinboim and Tian, 2015; Correa and Bareinboim, 2017; Correa et al., 2018, 2019a,b). Selection bias is also related to data missingness as a special case of Missing Not At Random (MNAR) (Mohan and Pearl, 2021).

Constraint-based methods causal discovery methods reason from an input of conditional independent relations derived from data to infer causal relations. The formative PC algorithm (Spirtes et al., 2000) identifies all causal effects from independence statements when all relevant variables have been observed. When selection bias and latent confounding are present, the FCI algorithm (Spirtes et al., 1995, 1999) is sound and, augmented with additional rules, complete (Zhang, 2008b). Extensions of FCI include optimizing for computationally efficiency (Colombo et al., 2012; Claassen et al., 2013), soundness under interruptions at any stage (Spirtes, 2001) and when cycles are present (Forré and Mooij, 2020; Mooij and Claassen, 2020). A sound and complete reformulation of FCI in terms of minimal (in)dependence statements and one additional rule is presented in Claassen and Heskes (2011).

2. Background

We first set out background theory. The main ingredients are graphical models with explicit selection and context nodes (representing environments).

Notation wise, we write single variables as capitals, i.e. X, and sets of variables as boldface, i.e. $\boldsymbol{X}$. In the context of graphical models, we interchangeably use the terms 'variable' and 'node', implying a random variable associated to a node.

2.1. Graphical Models

A *Directed Mixed Graph* (DMG) is a graph $\mathcal{G} = (\boldsymbol{V}, E)$ consisting of a node set $\boldsymbol{V}$ representing random variables with some distribution $\mathbb{P}(\boldsymbol{V})$ and an edge set E consisting of directed and bidirected edges between pairs of different nodes. In the canonical causal graphical interpretation, directed edges represent *direct* causal relations with respect to $\boldsymbol{V}$, and X is a *direct cause* of Y if there exists a directed edge from X towards Y. The bidirected edges represent latent confounding, and can be viewed as originating from an underlying graph where these nodes have been marginalized out.

Node pairs connected by any edge are *adjacent*, and a *path* is a sequence of adjacent edges, where each node occurs at most once. There is an *ancestral* relation from X to Y if there exists a path between X and Y where all edges are directed and have an arrowhead into Y. We refer to X as an *ancestor* of Y, and Y as a *descendant* of X. A *collider* is a node X along a path with edge configuration $\circ\!\!\rightarrow X \leftarrow\!\!\circ$, where the circle in $\circ\!\!\rightarrow$ indicates either an arrowhead or a tail.

A *Markov property* relates properties of the graph to (conditional) independences in the corresponding distribution. When cycles are present, the σ-separation Markov property is a non-trivial generalization of the d-separation Markov property (Forré and Mooij, 2017), see Section **??** in the Appendix for a full definition. Often a causal graph is associated to a *Structural Causal Model* (SCM), where equations represent causal mechanisms as functions of observed endogenous variables and independent exogenous noise variables, see e.g. Pearl (2009). Effectively, the DMGs with the σ-separation property that we use in this work are a graphical representation of *Simple* SCMs, a convenient subclass of all cyclic SCMs that allows for certain cyclic interactions while maintaining intuitive causal semantics (Bongers et al., 2021). In this work, we also assume *σ-faithfulness* (which we will also refer to as "faithfulness"), that is, that there are no other conditional independences than those entailed by the σ-separation Markov property.

2.2. Interventional Data and Background Knowledge

We use the recently introduced *Joint Causal Inference* (JCI) framework (Mooij et al., 2020) to model certain background knowledge, interventions and other manipulations with *context variables*, in addition to the endogenous *system variables* and assumptions on the causal relations between these. The central aim of JCI is to define a meta-system of combined system and context variables that can be used for a variety of causal methods.

The full JCI framework requires up to four assumptions, two of which are necessary for systems with only a single context variable:

Assumption 0 *(*Joint SCM*) The data-generating process underlying $\boldsymbol{C}$ and $\boldsymbol{X}$ is modeled by a single underlying MAG.*

Assumption 1 *(*Exogenity*) No system variable $X \in \boldsymbol{X}$ is an ancestor of any context variable $C \in \boldsymbol{C}$.*

We refer to this combined modeling assumption as JCI-1. For the remainder of this work we have at most one context variable, which we always denote as C and graphically represented by a squared node (see 1(*b*)) We direct the reader to Mooij et al. (2020) for the JCI assumptions related to larger context models that are outside the scope of this work.

2.3. Selection Bias and Causal Relations

We will mainly use the ADMG framework, allowing for latent confounding, with possible inclusion of JCI-1 context nodes to encode background knowledge. However, ADMGs are not closed under conditioning, such that whenever a selection bias mechanism requires modelling, we add an unmeasured conditioned node for the selection bias variable $\boldsymbol{S}$ explicitly, represented by a shaded gray node (see Fig. 1(*a*)). We sometimes use $\boldsymbol{S} = 0$ and $\boldsymbol{S} = 1$ to indicate that a single model or dataset is not affected or affected by selection bias respectively.

For ADMGs (and DAGs, in the following), X is a *parent* of Y if there exists a directed edge is with an arrowhead on Y. In the canonical causal interpretation, directed edges represent *direct* causal relations, where the relationship $X \to Y$ implies direct causation. A *collider* is a node X along a path with an edge configuration $\ldots, e_{in}, X, e_{out}, \ldots$, where edges e_{in} and e_{out} have an arrowhead towards X. There is an *ancestral* relation from X to Y if there exists a path between X and Y where all adjacencies are directed and into Y. We refer to X as an *ancestor* of Y, and Y as a *descendant* of X. There is an *anterior* relation from X to Y in a ADMG with a selection variable if X is anterior of Y in the corresponding MAG. (Richardson and Spirtes, 2002) Note that the set of anteriors and anterior relations is a superset of all ancestors and ancestral relations respectively, and the definitions match when selection is not present. The sets of parents / ancestors / descendants of X are denoted as $\text{PA}(X)$ / $\text{AN}(X)$ / $\text{DE}(X)$ respectively, where the same definitions hold for sets $\boldsymbol{X}$ disjunctively.

3. Causal Discovery from Constraints

In this work we are primarily interested in causal discovery, predicting (part of) the true and unbiased causal relations underlying the causal system, from data that is under selection bias.

Without selection bias, the external manipulation of a variable X (such as randomization and various types of interventions) breaks confounding and potentially allows for the discovery of an ancestral causal relation to some target Y. Whenever Y exhibits a meaningful change from its observational distribution under some intervention of X, we find that $X \in \text{AN}(Y)$.

The goal is to discover these ancestral relations from data gathered under a selection bias mechanism $\boldsymbol{S}$. The existence of such a relation underlying the data is quite strong, and would imply that we can predict the effect of interventions had we observed a full random sample. When selection is present, the intervention of X can potentially detect an *anterior* relation instead of an ancestral relation. To exclude this possibility, we limit the goal to detect *purely ancestral* causal relations, i.e. $X \in \text{AN}(Y) \wedge X \notin \text{AN}(S)$.

The remainder of this section introduces logical rules for causal inference. We then apply these results in three and four variable constraint-based discovery of ancestral relations in the presence of selection bias.

Assumptions As our main topic is constraint-based causal discovery, throughout the remainder of this work we assume causal *faithfulness*, relating each (conditional) independence in the data to a non-adjacency in the graph. Combined with the Markov property, this implies (conditional) independences from the data map one-to-one to instances of *d*-separation in the causal graph. There is no *causal sufficiency*, i.e. latent variables are allowed.

3.1. Logical Causal Inference Rules

In Claassen and Heskes (2011), the concept of a *minimal conditional independence* for disjoint sets of variables $\{X\}, \{Y\}, \boldsymbol{W}, \boldsymbol{Z}$ is introduced as:[1]

$$X \perp\!\!\!\perp Y \mid \boldsymbol{W} \cup [\boldsymbol{Z}] := (X \not\perp\!\!\!\perp Y \mid \boldsymbol{W}) \wedge (\forall \boldsymbol{Z'} \subsetneq \boldsymbol{Z} : X \perp\!\!\!\perp Y \mid \boldsymbol{W} \cup \boldsymbol{Z'}), \tag{1}$$

1. Without faithfulness, the statements in this section hold for graphical notions of minimal *d*-separations and *d*-connections instead of minimal independences and dependences.

where $\boldsymbol{W}$ represents selection all other conditioned variables. Similarly, a *minimal conditional dependence* is defined as:

$$X \not\perp\!\!\!\perp Y \mid \boldsymbol{W} \cup [\boldsymbol{Z}] := (X \perp\!\!\!\perp Y \mid \boldsymbol{W}) \wedge (\forall \boldsymbol{Z'} \subsetneq \boldsymbol{Z} : X \not\perp\!\!\!\perp Y \mid \boldsymbol{W} \cup \boldsymbol{Z'}). \tag{2}$$

Here, the intuition for a minimal independence (dependence) given $[\boldsymbol{Z}]$ is that $\boldsymbol{Z}$ contains the set of variables that each have a meaningful role in separating (connecting) X from Y. Indeed, for all proper subsets of $\boldsymbol{Z}$, X and Y are dependent when conditioning on $\boldsymbol{W}$ in (1), analogously for proper subsets in (2).

The following useful Lemma was introduced by Claassen and Heskes (2011) as part of a logical reformulation of FCI, mapping minimal (in)dependences to ancestral causal relations, and generalized to DMGs with the σ-separation Markov Property by Mooij et al. (2020).

Lemma 1 (Claassen and Heskes (2011)) *Let $\{X\}, \{Y\}, \boldsymbol{W}, \boldsymbol{Z}$ be disjoint sets of variables on a DMG. Then the following statements hold*[2]*:*

$$X \perp\!\!\!\perp Y \mid \boldsymbol{W} \cup [\boldsymbol{Z}] \implies \boldsymbol{Z} \in \textsc{an}(X \cup Y \cup \boldsymbol{W}), \tag{3}$$

$$X \not\perp\!\!\!\perp Y \mid \boldsymbol{W} \cup [\boldsymbol{Z}] \implies \boldsymbol{Z} \notin \textsc{an}(X \cup Y \cup \boldsymbol{W}). \tag{4}$$

Here, (3) states that finding a minimal independence results implies at least one ancestral relation. However, a minimal dependence for Z in (4) implies that Z is not ancestral to *all* of X, Y and conditioning set $\boldsymbol{W}$, which can include a selection node.

Here, (3) states that finding a minimal independence results in at least one implicated ancestral relation. However, a minimal dependence for Z in (4) implies that Z is not ancestral to *all* of X, Y and conditioning set W, which can include the selection node.

Crucially, Claassen and Heskes (2011) showed that repeated applications of these two rules, together with a third "*inferred blocking*"-rule over a minimum of 4 variables that is outside the scope of this work, are a sound and complete reformulation of FCI. Additional inferred statements about ancestral relations can be reached by reasoning over the list of all logical consequences that originate from repeated applications of the logical rules and transitivity, acyclicity and irreflexivity properties of the implied relations. This algorithm is referred to as *Logical Causal Inference* (LoCI).

3.2. Three Variables: LCD

Local Causal Discovery (LCD) (Cooper, 1997) is a constraint-based method for three variables, combining (in)dependence testing with ancestral background knowledge. A *local* method, when more than three variables are included in the data, a search across all three variable subsets matching LCD conditions below is performed. Allowing for latent variables, the JCI-1 formulation of LCD requires that the data includes (at least) one context variable C subject to JCI-1 assumptions.[3] For each triple of variables $\langle C, X, Y \rangle$, if

$$C \perp\!\!\!\perp Y \mid [X] \tag{5}$$

holds in the data then $X \in \textsc{an}(Y)$, $Y \notin \textsc{an}(X)$ and X and Y are unconfounded (such that $\mathbb{P}(Y \mid \mathrm{do}(X)) = \mathbb{P}(Y \mid X)$). Soundness follows of the first two statements follows almost directly.

2. Part of the first implication was already introduced by Spirtes et al. (1999) as $X \perp\!\!\!\perp Y \mid \boldsymbol{S} \cup [\boldsymbol{W} \cup \boldsymbol{Z}] \implies \boldsymbol{Z} \in \textsc{an}(X \cup Y \cup \boldsymbol{S})$
3. Technically the original assumption of LCD is slightly stronger, only requiring $X \notin \textsc{an}(C)$. The proofs below also hold for this case.

Proposition 2 *Without selection bias but with possible confounding, LCD as described above is sound.*

Proof The constraints (5) correspond to a minimal independence (1) for X, such that $X \in \text{AN}(C \cup Y)$ by Lemma 1. Then eliminating $X \in \text{AN}(C)$ by the JCI-1 assumption for C implies $X \in \text{AN}(Y)$. By acyclicity $Y \notin \text{AN}(X)$. As C and Y are non-adjacent because $C \not\perp\!\!\!\perp Y$ (and faithfulness) but $C \perp\!\!\!\perp Y | X$ implies that C and X are adjacent and there is a non-collider path through X. Because $X \notin Y$ and no selection bias is present, the edge is either confounded or directed into X, such that the end-point in X is an arrowhead. As we cannot have a collider on X this excludes a bidirected edge between X and Y. ■

In Fig 2, all ADMGs that are valid under LCD conditions are given, matching the conditions in the proof above. In each one, X is ancestral of Y and the pair is unconfounded.

When the data is gathered under selection bias $\boldsymbol{S}$, the application of the (in)dependence tests of LCD in (5) is essentially replaced with

$$C \perp\!\!\!\perp Y \,|\, X \cup \boldsymbol{S}. \tag{6}$$

Clearly, these tests combined with JCI-1 background knowledge for C do not result in a purely ancestral relation. Still, a practitioner might be oblivious to the selection bias mechanism acting on the data at-hand, and we investigate LCD further in this condition.

In Fig. 3, all non-trivial ADMGs with selection node are given that adhere to (6), as is easily checked. We only include graphs here where $\boldsymbol{S}$ contributes to a spurious correlation i.e. when it is (a descendant of) a collider node. Fig. 3(*a*) shows a failure mode of LCD where S induces a spurious adjacency and X and Y are non-adjacent otherwise. Curiously, in Fig. 3(*b*) we have the opposite $Y \in \text{AN}(X)$ of the target causal relation in the true graph. See also Example 1. On the other hand, in Fig. 3(*c*) and 3(*d*) for instances of ADMGs where the selection mechanism could contribute to the adjacency between C and X. This could potentially increase the prevalence of LCD triples in $\boldsymbol{S} = 1$ and increase the recall of its predictions in the data when the selection mechanism is 'switched on' compared to the same causal model without selection.

We generalize the result for LCD under selection bias to the following impossibility.

Proposition 3 *When selection bias is present, there exists no sound method for discovering a purely ancestral causal relation from a combination of (in)dependence tests and ancestral JCI-1 background knowledge for three variables.*

Proof Sketch of proof by enumeration. By completeness of LoCI and the limit of three variables we are limited to checking for repeated consequences of Lemma 1 in combinations with of JCI-1 background knowledge, which is of the form of excluding certain ancestral relations. Three variables is sufficient for at most one minimum (in)dependence relation, and exploring both cases with combined non-ancestral relations does not result in a purely ancestral relation. ■

3.3. Four Variables: Y-Structures

Given the result in the previous section for three variables, we now turn to the four variable case.

The Extended Y-Structure (Mani et al., 2006) and the Y-Structure (Mooij and Cremers, 2015) algorithms, which we collectively refer to as Y-Structures, do not require a JCI assumption. Each

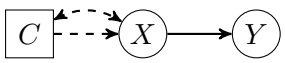


Figure 2: **(LCD, no selection bias)** The three ADMGs that adhere to LCD constraints without selection bias. Dashed edges indicate that at least one of the depicted edges must be present.

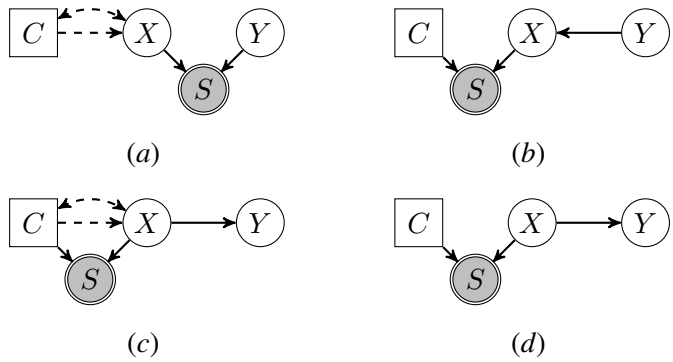


Figure 3: **(LCD, selection bias)** The non-trivial ADMGs of LCD that are allowed in addition to Fig. 2 when selection bias is present, represented by $\boldsymbol{S}$. Failure modes are presented in (a) and (b), where $X \notin \text{AN}(Y)$. The graphs in (c) and (d) are additional modes of LCD where $X \in \text{AN}(Y)$.

method consists of a search for a pattern of (in)dependence constraints between a quadruple of variables. In Mani et al. (2006) and Mooij and Cremers (2015), it was shown that Y-Structures imply an ancestral causal relation which is unconfounded, when selection bias is not present. Cooper (1997) showed (but not proved) that the constraints from an Extended Y-Structure imply that selection bias can be excluded.

Here we show soundness of Extended Y-Structures under selection bias.

Proposition 4 (Extended Y-Structure) *Let $\langle V, W, X, Y\rangle$ be an ordered tuple of disjoint variables in $\boldsymbol{V}$ and $\boldsymbol{S}$ be the set of selection bias variables. If*

$$\begin{aligned} V &\perp\!\!\!\perp Y \mid [X] \cup \boldsymbol{S} \\ V &\not\perp\!\!\!\perp W \mid [X] \cup \boldsymbol{S}, \end{aligned} \tag{7}$$

then $X \in \text{AN}(Y)$, $Y \notin \text{AN}(X)$, $X \notin \text{AN}(\boldsymbol{S})$ and X and Y are unconfounded.

Proof The application of Lemma 1 to $V \perp\!\!\!\perp Y \mid [X] \cup \boldsymbol{S}$ implies that $X \in \text{AN}(V \cup Y \cup \boldsymbol{S})$, and applying it to $V \not\perp\!\!\!\perp W \mid [X] \cup \boldsymbol{S}$ gives $X \notin \text{AN}(V \cup W \cup \boldsymbol{S})$, i.e. $X \notin \text{AN}(\boldsymbol{S})$. Substitution leads directly to $X \in \text{AN}(Y)$.

We now consider paths between V and Y, and note that all such paths must be σ-separated when conditioning on $X \cup \boldsymbol{S}$, as $V \perp\!\!\!\perp Y \mid X \cup \boldsymbol{S}$. But there must exist an open path $\mathcal{P}$ between V and Y given only $\boldsymbol{S}$, as $V \not\perp\!\!\!\perp Y \mid \boldsymbol{S}$ in (7). Now $V \perp\!\!\!\perp W \mid \boldsymbol{S}$ implies that all paths between V and W are σ-separated when conditioning on $\boldsymbol{S}$ only, such that $\mathcal{P}$ cannot contain W. In addition, V is non-adjacent to both Y and W. We consider different cases for the path $\mathcal{P}$. We provide the proof assuming that $\boldsymbol{S} = \{S\}$, that is, there is just a single selection node. (The proof strategy can be extended to allow for multiple selection nodes but it becomes a bit more involved to write it down).

Suppose $\mathcal{P}$ ends with an edge between Y and S, i.e. $V \dots S \circ\!\!-\!\!\circ Y$, where $\circ\!\!-\!\!\circ$ can be any of the edges $\leftarrow$, $\rightarrow$ and $\leftrightarrow$. Then the only option is through X, as a path of the form $V \circ\!\!-\!\!\circ S \circ\!\!-\!\!\circ Y$ cannot be closed by conditioning on X. Therefore $\mathcal{P}$ must be of the form $V \circ\!\!\rightarrow X \leftarrow\!\!\circ S \circ\!\!-\!\!\circ Y$, where the orientations are due to $X \notin \text{AN}(V \cup \boldsymbol{S})$. This is a contradiction, because X is a collider that is not ancestor of $\boldsymbol{S}$ and hence blocks the path. Hence, $\mathcal{P}$ must be of the form $V \dots X \circ\!\!-\!\!\circ Y$. Now $X \notin \text{AN}(V \cup W \cup \boldsymbol{S})$, so all edges connecting to X involving W, V and $\boldsymbol{S}$ are with an arrowhead on X, such that $\mathcal{P}$ must be of the form $V \cdots \circ\!\!\rightarrow X \circ\!\!-\!\!\circ Y$, excluding $X \leftrightarrow Y$ and $X \leftarrow Y$, because otherwise we would again obtain a contradiction by X being a collider. ■

Here unconfoundedness, $X \in \text{AN}(Y)$ and $Y \notin \text{AN}(X)$ together imply that the causal effect can be estimated from observational data by using $\mathbb{P}(Y \mid \text{do}(X), \boldsymbol{S}) = \mathbb{P}(Y \mid X, \boldsymbol{S})$ (where $\boldsymbol{S}$ is thus possibly empty) for the Extended Y-Structure. This follows from the do-calculus for simple SCMs (Forré and Mooij, 2020).

The Y-Structure method adds two additional constraints $W \perp\!\!\!\perp Y \mid [X] \cup \boldsymbol{S}$ in addition to (7) required by the Extended Y-Structure, symmetrizing the equations for W and V (Mooij and Cremers, 2015). Soundness follows from Prop. 4.

4. Experiments

We first present the results of several simulation experiments and a real-world experiments.

4.1. Methods and Estimators

The practical estimators of LCD, Y-Structures and Extended Y-Structures, denoted as **`LCD`**, **`YSt`** and **`YSt-Ext`** respectively, use a simple partial correlation test under linear-Gaussian assumptions.

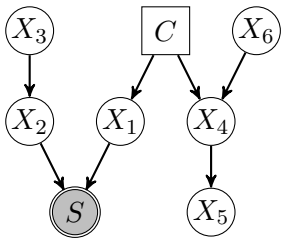
Additionally, we use Invariant Causal Prediction (ICP) (Peters et al., 2016) as a causal baseline for comparison, one of the state-of-the-art methods in this setting (Meinshausen et al., 2016), ICP is a method for predicting the direct causes $\text{PA}(Y)$ of a target variable Y. It tests in a sophisticated way if the conditional distribution $Y|\text{PA}(Y)$ remains invariant under changes of the environment. In Mooij et al. (2020) a reformulation of ICP is given in terms of ancestral relations by assuming causal faithfulness. In that formulation, LCD is sound under latent confounding but not selection bias, in a fashion to LCD. For the practical estimator **`ICP`**, we use its standard implementation in the `InvariantCausalPrediction` package in `R`.

Finite Sample Scoring Given the finite sample setting, we assign a score to each cause effect prediction in each of the methods that is an indication of the confidence level in the result. For **`LCD`**, we follow Mooij et al. (2020) in using $-\log(p_{CY})$, where p_{CY} is the p-value associated to the null-hypothesis of dependence between context variable C and target variable Y. For **`ICP`** we use the maximum of p-values for the predicted ancestors.

In Y-Structures type estimators **`YSt`** and **`YSt-Ext`**, for each potential ancestral relation $X \in \text{AN}(sY)$ we compute

$$\max\left(\min_{V',W'}\left(-\log(p_{V'Y}), -\log(p_{W'Y})\right)\right) \tag{8}$$

and for all patterns $\langle V', W', X, Y\rangle$ that are discovered. This follows the intuition of the scoring in **`LCD`**, and ranks a discovered Y-Structure more confident with the smallest dependence of the



(*a*)

Method	**Unbiased** ($S = 0$)			**Biased** ($S = 1$)		
	#Pred	TP	FP	#Pred	TP	FP
ICP	399	319	80	1987	381	1606
LCD	419	397	22	569	394	175
YSt[4]	396	391	5	393	390	3

(*b*)

Figure 4: **(Fixed Graph)** (a) Graph used for fixed-graph simulations, containing a Y-Structure $\{C, X_4, X_5, X_6\}$ and a failure mode of **LCD** $\{C, X_1, X_2\}$. (b) Results of applying **ICP**, **LCD** and **YSt** to the 200 random causal models of the fixed graph in (a). The total number of predictions (#Pred) of which partition of true positives (TP) and false positives (FP) are given for the unbiased and biased data.

two auxiliary variables (the 'legs' of the Y-Structure) with target variable Y. If there are multiple Y-Structures for a particular ancestral relation of interest $X \in Y$, the largest score is taken.

4.2. Simulations

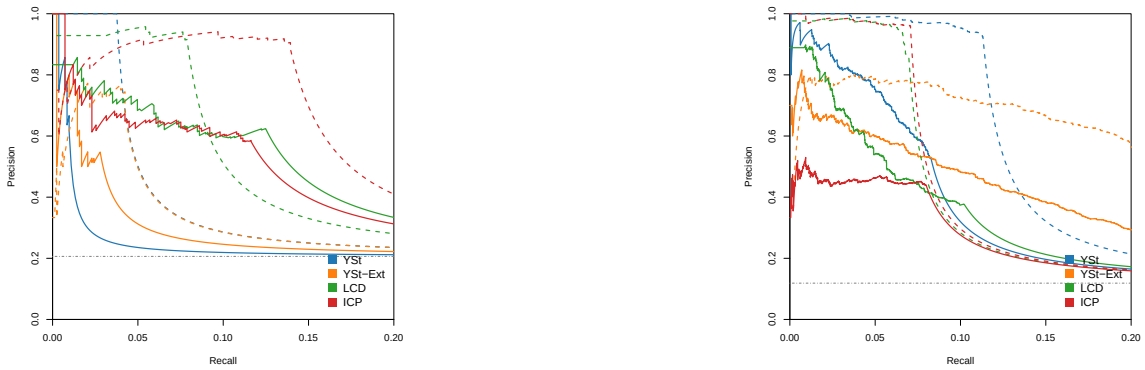
We run several simulation experiments over both randomly sampled and fixed directed graphs, which we describe later in this section. In these we include an explicit selection bias for the preferential sampling and a context node for including interventional data. As our primary aim in these experiments is assessing algorithm performance under the effects of selection bias, we do not include the possibility of latent confounding between variables. Note however that each of the methods included is sound under latent confounding. Both random and fixed graphs are linear-Gaussian models. The linear edge weights between system, context and selection bias variables are sampled uniformly from $[-1.5, -0.5] \cup [0.5, 1.5]$. The exogenous noise variables are drawn independently from a standard-Gaussian distribution for a total of 10000 realizations. In each simulation we report the results of 200 different linear-Gaussian models.

Unbiased and Biased Data For a given graph and edge weights, two data sets are sampled: one where the selection bias mechanism is present and one where it is disabled. In the biased data, $\boldsymbol{S}$ is conditioned on a small interval $[1, 1.25]$ and other samples are repeatedly rejected. In both cases, $\boldsymbol{S}$ is not observed and not included in the training set.

4.2.1. Fixed Graph

As a demonstration of the effect of the selection bias mechanism applied here, we sample 200 models from the fixed ADMG in Fig. 4(*a*). Here a false positive for LCD under selection bias in Fig. 3(*b*) is combined with a graph adhering to the (extended) Y-Structure conditions.

The results in Tab. 4(*b*), where we compare ancestral relation predictions for each method to the true ancestral causal effect as it exists in the graph. The TP levels are similar in the case of unbiased data, where the non-Gaussian estimator of *ICP* as used performs a bit worse in Gaussian conditions. When applying the selection mechanism on $\boldsymbol{S}$, both *ICP* and *LCD* predict a large amount of false positives.



(*a*) Small graphs ($p = 6$)

(*b*) Large graphs ($p = 16$)

Figure 5: **(Random Graphs)** PR curves for experiments with random graphs.

4.2.2. RANDOM GRAPHS

We sample random graphs in the following way. For a given observed number of observed variables p, directed edges are sampled independently with a fixed probability between each node and this is repeated for a given seed until an acyclic graph is found. In this procedure, graphs that do not meet a predetermined minimum number of collider patterns dependent on p are discarded, to increase the prevalence of spurious correlations due to selection bias. Finally, we randomly pick with uniform weight one of the source nodes as a context variable, and one of the descendant leaf nodes of the set of colliders as the single selection bias variable.

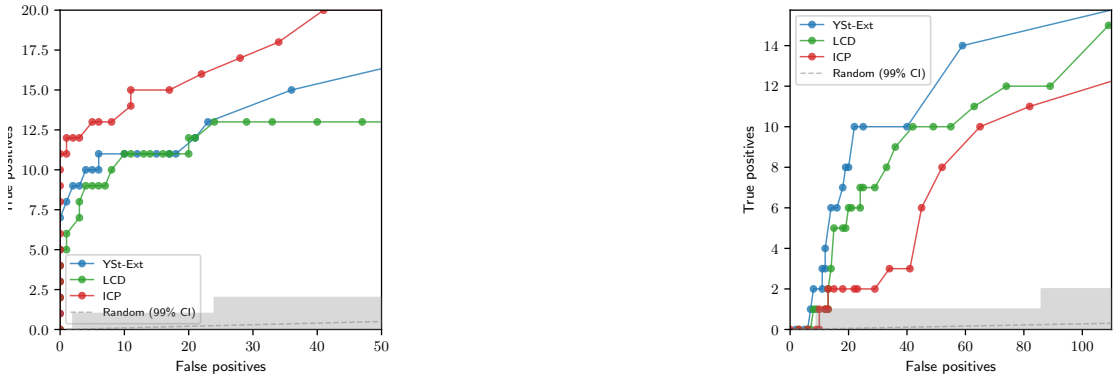
The results for two vertex set sizes are given in Fig 4.2.2. For the small graphs, we find a larger drop in Precision when comparing predictions computed on $\boldsymbol{S} = 1$ data to those from $\boldsymbol{S} = 0$ data, indicating a relatively stronger effect of the selection mechanism in small graphs. We find in this regime, **`LCD`** and **`ICP`** are outperforming Y-Structure methods in both precision and recall in the $\boldsymbol{S} = 1$ setting. The low recall indicates that not enough Y-Structures are randomly created in such a small variable set. In the larger graph we see almost the reverse situation in $\boldsymbol{S} = 1$, where **`YSt`** is outperforming all other methods, which might indicate a good finite-sample score (8), and **`YSt-Ext`** showing a large recall. Additionally, the large recall of **`LCD`** might indicate additional predictions due to selection bias adjacencies, see Fig. 3(*c*) and 3(*d*).

4.3. Real-World Data

We use real-world data from a gene expression setting (Kemmeren et al., 2014). Expression levels of genes, indicating regulation activity affecting other genes, are captured for each of $p = 6179$ genes under $n = 262$ observational and $m = 1479$ experimental (interventional) conditions. In each of these experiments, one gene has been knocked-out by a chemical and expression levels are measured once for all variables.

The availability of a large set of single-variable interventional data is appealing as it allows for both learning from (part of) the interventional data and validating the predicted causal effects from (a disjoint part of) test set.

High-Dimensionality We operate here in a high-dimensional setting ($p < m \ll n$), where statistical, let alone causal, procedures often require a form of regularization.



(*a*) Internal validation (*b*) Domain validation

Figure 6: **(Real-World data)** ROC curves for experiments with real-world gene expression data.

We use L_2-boosting regression for each target variable Y as a preselecting method for `LCD` (Versteeg and Mooij, 2019). Essentially this reduces the search for each X in the LCD triple $\langle C, X, Y\rangle \in V_{\text{con}} \times V_{\text{sys}}^2$ to a potential parent set $X \in \boldsymbol{V}_Y^{\text{sel}} \subsetneq \boldsymbol{V}$, where $\boldsymbol{V}_Y^{\text{sel}}$ are the variables selected by a L_2-boosting regression for Y. The **`ICP`** estimator in `R` uses a similar preselection technique to limit the set of plausible parents (ancestors) when $p > 8$. To improve stability of the predictions, we bootstrap each method for 100 random subsamples of the data (Meinshausen and Bühlmann, 2010).

For the (Extended) Y-Structures, we fix V as the context variable C. We also reduce the search space of X and W in each $\langle C, W, X, Y\rangle$ pattern to the preselected variable sets $X \in \boldsymbol{V}_Y^{\text{sel}}$ and $W \in \boldsymbol{V}_X^{\text{sel}}$ respectively.

4.3.1. Internal cross-validation

Here we use the gene expression data for a causal cross-validation setting, where all observational and interventional samples are partitioned in 5 equal parts. In a sequence of 5 experiments, one of each partition is used as a test set and the remaining 4 parts are merged as training data. The training data is merged to one JCI dataset with one binary context variable C, where $C = 0$ ($C = 1$) corresponds to all observational (interventional) samples.

The ground truth set of 'true' causal relations is computed from the test set in the following way. Let $X_{j;i}$ be the single sample of the expression of $X_j \in \boldsymbol{V}$ under the intervention of $X_i \in \boldsymbol{V}$, where i and j take values in $[p] = \{1, \ldots, p\}$). Following Versteeg and Mooij (2019), we compute a score S_{ij} representing the size of the absolute causal effects $S_{ij} = \frac{|X_{j;i} - \mu_j|}{\sigma_j}$, where μ_j and σ_j are the empirical mean and empirical standard deviation of the observational test data for X_j. The ground truth set for ancestral causal relations between X_i and X_j is then simply $\{(i,j) \in ([p] \times [p]) \mid S_{ij} > t \wedge i \neq j\}$, for some minimal threshold value t and with self effects excluded.

In Fig. 6(*a*), the resulting ROC curve is shown, where the threshold t is chosen in way that 1 percent of the causal relations is contained in the ground truth. We find that all causal methods **`ICP`**, **`LCD`** and **`YSt-Ext`** outperform both random guessing (in gray). We see that **`ICP`** here is the most successful. **`YSt-Ext`** is seen to slightly outperform **`LCD`**.

4.3.2. DOMAIN VALIDATION

We query the known genetic relations from an online compendium of expression data (Cherry et al., 2012) to compile a ground truth of domain knowledge, where we interpret the causal direction in a similar manner as Meinshausen et al. (2016). Compared to Sec. 4.3.1, we use all gene expression data here in one training set.

The results are shown in Fig. 6(*b*), where we find that again all methods outperform the random baseline, but **`YSt-Ext`** is here superior to **`ICP`** and **`LCD`** in the complete FP range.

5. Discussion

In this work we investigated the performance of local constraint-based algorithms when selection bias is present in addition to latent confounding. LCD was found to not be sound, might produce anti-causal predictions, but can also increase recall when the data is selected. The (Extended) Y-Structures methods were shown to be successful in predicting ancestral relations in this setup. Empirically, we show that Y-Structures in particular work well in some simulation settings, when a finite-sample score is used as a measure of confidence. Using two real-world high-dimensional datasets and the same score, we showed an increase of performance of Y-Structures on an external dataset compared to the interventions available in the data. This indicates that a selection bias mechanism might underlay the data.

References

Elias Bareinboim and Judea Pearl. Controlling selection bias in causal inference. In *Artificial Intelligence and Statistics*, pages 100–108, 2012.

Elias Bareinboim and Jin Tian. Recovering causal effects from selection bias. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 3475–3481, 2015.

Elias Bareinboim, Jin Tian, and Judea Pearl. Recovering from selection bias in causal and statistical inference. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pages 2410–2416, 2014.

Stephan Bongers, Patrick Forré, Jonas Peters, and Joris M Mooij. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5):2885–2915, 2021.

J Michael Cherry, Eurie L Hong, Craig Amundsen, Rama Balakrishnan, Gail Binkley, Esther T Chan, Karen R Christie, Maria C Costanzo, Selina S Dwight, Stacia R Engel, et al. Saccharomyces genome database: the genomics resource of budding yeast. *Nucleic acids research*, 40 (D1):D700–D705, 2012.

Tom Claassen and Tom Heskes. A logical characterization of constraint-based causal discovery. In *UAI 2011, Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 135–144. AUAI Press, 2011.

Tom Claassen, Joris M. Mooij, and Tom Heskes. Learning sparse causal models is not np-hard. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence, UAI 2013*. AUAI Press, 2013.

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

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

Juan D. Correa and Elias Bareinboim. Causal effect identification by adjustment under confounding and selection biases. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI 2017*, pages 3740–3746. AAAI Press, 2017.

Juan D. Correa, Jin Tian, and Elias Bareinboim. Generalized adjustment under confounding and selection biases. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, AAAI 2018*, pages 6335–6342. AAAI Press, 2018.

Juan D. Correa, Jin Tian, and Elias Bareinboim. Adjustment criteria for generalizing experimental findings. In *Proceedings of the Thirty-Sixth International Conference on Machine Learning, ICML 2019*, volume 97 of *Proceedings of Machine Learning Research*, pages 1361–1369. PMLR, 2019a.

Juan D. Correa, Jin Tian, and Elias Bareinboim. Identification of causal effects in the presence of selection bias. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019*, pages 2744–2751. AAAI Press, 2019b.

Patrick Forré and Joris M Mooij. Markov properties for graphical models with cycles and latent variables. *arXiv preprint arXiv:1710.08775*, 2017.

Patrick Forré and Joris M Mooij. Causal calculus in the presence of cycles, latent confounders and selection bias. In *Uncertainty in Artificial Intelligence*, pages 71–80. PMLR, 2020.

Patrick Kemmeren, Katrin Sameith, Loes AL Van De Pasch, Joris J Benschop, Tineke L Lenstra, Thanasis Margaritis, Eoghan O'Duibhir, Eva Apweiler, Sake van Wageningen, Cheuk W Ko, et al. Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell*, 157(3):740–752, 2014.

Subramani Mani, Peter Spirtes, and Gregory F Cooper. A theoretical study of y structures for causal discovery. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 314–323. AUAI Press, 2006.

Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

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

Karthika Mohan and Judea Pearl. Graphical models for processing missing data. *Journal of the American Statistical Association*, pages 1–16, 2021.

Joris M. Mooij and Tom Claassen. Constraint-based causal discovery using partial ancestral graphs in the presence of cycles. In *Proceedings of the Thirty-Sixth Conference on Uncertainty in Artificial Intelligence, UAI 2020*, volume 124 of *Proceedings of Machine Learning Research*, pages 1159–1168. AUAI Press, 2020.

Joris M Mooij and Jerome Cremers. An empirical study of one of the simplest causal prediction algorithms. In *Proceedings of the UAI 2015 Conference on Advances in Causal Inference-Volume 1504*, pages 30–39, 2015.

Joris M Mooij, Sara Magliacane, and Tom Claassen. Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, 21:1–108, 2020.

Judea Pearl. *Causality*. Cambridge university press, 2009.

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

Thomas Richardson and Peter Spirtes. Ancestral graph markov models. *The Annals of Statistics*, 30(4):962–1030, 2002.

Anna Roumpelaki, Giorgos Borboudakis, Sofia Triantafillou, and Ioannis Tsamardinos. Marginal causal consistency in constraint-based causal learning. In *Proceedings of the UAI 2016 Workshop on Causation: Foundation to Application co-located with the 32nd Conference on Uncertainty in Artificial Intelligence (UAI 2016)*, volume 1792 of *CEUR Workshop Proceedings*, pages 39–47. CEUR-WS.org, 2016.

Peter Spirtes. An anytime algorithm for causal inference. In *International Workshop on Artificial Intelligence and Statistics*, pages 278–285. PMLR, 2001.

Peter Spirtes, Christopher Meek, and Thomas Richardson. Causal inference in the presence of latent variables and selection bias. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 499–506. Morgan Kaufmann Publishers Inc., 1995.

Peter Spirtes, Christopher Meek, and Thomas Richardson. An algorithm for causal inference in the presence of latent variables and selection bias. *Computation, causation, and discovery*, 21: 1–252, 1999.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search, Second Edition*. Adaptive computation and machine learning. MIT Press, 2000.

Philip Versteeg and Joris M. Mooij. Boosting local causal discovery in high-dimensional expression data. In *2019 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2019*, pages 2599–2604. IEEE, 2019.

Jiji Zhang. *Causal Inference and Reasoning in Causally Insufficient Systems*. PhD thesis, Citeseer, 2006.

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

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