Beyond the Markov Equivalence Class: Extending Causal Discovery under Latent Confounding

Author names withheld

Editor: Under Review for CLeaR 2023

Abstract

In this work, we show how to combine two popular paradigms for causal discovery from observational data in the presence of latent confounders in order to arrive at a much more informative causal model. Building on the seminal constraint-based causal discovery algorithm, FCI, we exploit the power of direct cause-effect pair identification to uncover new relationships, which can subsequently be propagated to find even more causal links in the rest of the model. This idea has been attempted before, but until now always under the assumption of no latent confounders. Using our new *causal direction criterion* (CDC), we can finally drop this limitation. We derive inference rules for orienting additional cause-effect relations and show how to minimize the number of tests during the CDC search. In our experimental evaluations over a range of simulated data sets, the resulting FCI-CDC algorithm increases recall by between 5%-10% compared to vanilla FCI, without loss in precision.

Keywords: causal discovery, cause-effect inference, additive noise models, latent confounding

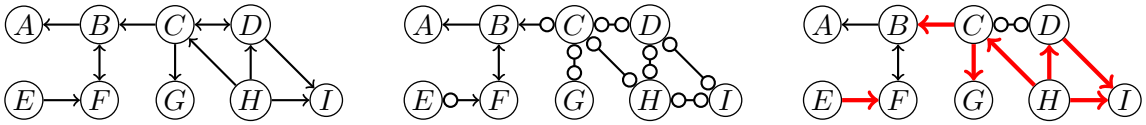
1. Introduction

Causal discovery is a popular topic in a wide range of fields such as epidemiology, economics, and the social sciences, because without understanding of the causal structure, one could wrongly interpret an association as a causal relationship (Pearl, 2009). The nature of an association between variables can be explained by a causal relationship, a hidden confounder, selection bias, a feedback loop, or a constraint relation (Goudet et al., 2019). Using causal discovery algorithms, we can learn from observational data to get a better understanding of the underlying causal structure.

A main class of causal discovery algorithms are constraint-based, in which (conditional) independencies between random variables are used to constrain the set of possible causal structures. In this study, we will build on the seminal constraint-based causal discovery algorithm *fast causal inference* (FCI) designed by Spirtes et al. (2000); Zhang (2008), which outputs a Markov equivalence class and does not rely on the causal sufficiency assumption, i.e., it works under the presence of hidden confounding. Unfortunately, causal graphs in a Markov equivalent class typically contain a significant number of non-invariant edges, that is, edges that can vary in the members in the equivalence class. For instance, consider the causal structure in Figure 1(*a*). Its Markov equivalence class is shown in Figure 1(*b*), in which the variant (or unknown) edge marks are indicated by circle marks. Since every member of the equivalence class contains the same conditional independence information between the corresponding variables, it is impossible to infer more information using only the conditional independencies (or constraints).

One effective way to go beyond the Markov equivalence class is to use cause-effect inference as a complement to the multivariate causal discovery methods based on conditional independencies (Guyon et al., 2019). In contrast to the latter, cause-effect inference methods directly focus on single cause-effect relationships. Cause-effect inference gained a lot of traction at the NeurIPS

© 2023 .



(*a*) The ground truth causal graph.

(*b*) The predicted causal graph by the FCI algorithm assuming oracle data.

(*c*) The predicted causal graph by the oracle FCI-CDC algorithm.

Figure 1: The graph on the left is a ground truth maximal ancestral graph (MAG) that represents a causal model, from which we cannot discover all edges using the FCI algorithm, as shown in the graph in the middle. On the right, the additional orientations indicated by the red thick edges show what can be gained on top of the FCI algorithm using the additional orientation rules defined in this study.

2008 Workshop on Causality, which led to the development of many approaches for inferring the direction of cause-effect pairs such as computing the complexity of marginals and conditionals (Hoyer et al., 2008; Zhang and Hyvärinen, 2009; Mooij and Janzing, 2010; Blöbaum et al., 2018), assuming *independence of causal mechanisms* (ICM) or *algorithmic independence* (Lemeire and Dirkx, 2006; Janzing and Schölkopf, 2010; Schölkopf et al., 2012; Sgouritsa et al., 2015; Goudet et al., 2017; Daniusis et al., 2012; Janzing et al., 2012), and supervised learning (Almeida, 2019; Fonollosa, 2019).

Cause-effect inference methods have already been used successfully to derive multivariate causal discovery algorithms. For instance, Shimizu et al. (2006) define orientation rules for linear non-Gaussian acyclic models, which they use to learn the entire causal structure. Hoyer et al. (2008) and Zhang and Hyvärinen (2009), on the other hand, iteratively search for the causal structure that best fits to the outcome of their cause-effect inference test, which is defined on additive noise models and post-nonlinear causal models, respectively. Tillman et al. (2009) use the PC algorithm (Spirtes et al., 2000), which is analogous to FCI under the assumption of causal sufficiency, together with an additional cause-effect inference orientation rule defined for local additive noise models. While all these methods are sound and make optimal use of the information available to find invariant orientations, they operate under the assumption of causal sufficiency, which severely limits their applicability to real-world problems, where hidden confounding often cannot be excluded.

In a recent paper, Huang et al. (2020) propose CD-NOD, an algorithm wherein they run orientation rules based on a pairwise causal inference test, followed by the same orientation rules used in the PC algorithm. Their pairwise causal inference test relies on the existence of a context variable (such as a time domain) to find independent causal mechanisms by conditioning on a suitable *deconfounding set* for each pair of variables. For the CD-NOD approach, Huang et al. (2020) assume *pseudo causal sufficiency*, i.e., any latent confounder can be described as a function of the context variable. CD-NOD represents an important step towards the goal of going beyond the Markov equivalence class by exploiting nonstationarity in the causal model. However, it is not guaranteed that latent confounding in real-world data can be properly expressed through such a context variable.

In this study, we show that we can improve on the orientation rule of CD-NOD so that we can compute the conditional independencies needed for the pairwise causal inference test directly, without having to use a context variable. This enables us to perform cause-effect inference without

assuming (pseudo) causal sufficiency. To that end, in section 3 we propose the *causal direction criterion* (CDC), a set of novel orientation rules that are sound in the presence of latent confounders and are based on an assumption closely related to the ICM postulate for (local) additive noise models. A key insight behind our approach is that, under very reasonable assumptions, we can actually use (additional) conditional independence tests to recognize when a specific interaction may have unobserved confounding and when it has not. It means that we no longer have to assume causal sufficiency, but can actually test for it when needed. We thus refrain from drawing potentially incorrect conclusions when we cannot rule out the presence of a specific unobserved confounder.

In section 4, we describe how we can extend the FCI algorithm with the orientation rules implied by the CDC to arrive at the sound FCI-CDC algorithm. Relying on the strengths of both constraint-based causal discovery and cause-effect inference, we arrive at an algorithm that makes it possible to go beyond the Markov equivalence class and infer a more informative causal structure. For instance, given enough data, we are able to recover the causal structure in Figure 1(*c*), where the red thick edges represent the additionally oriented edges on top of the FCI algorithm output. In section 5, we further show empirically that the FCI-CDC algorithm can recover new causal directions that were impossible to detect when applying the vanilla FCI algorithm.

2. Preliminaries & Notations

In this study, we denote graphs by calligraphic letters, for instance the *maximal ancestral graph* (MAG) $\mathcal{M}$, which is used to describe causal models in the presence of latent confounders, and the *partial ancestral graph* (PAG)[1] $\mathcal{P}$, which contains all invariant edge marks shared by all MAGs in the same Markov equivalence class. See Richardson (1996); Zhang (2008) for detailed definitions. We denote the vertex (or index) set of a graph by I. Standard graphical terms such as adjacency, parent, ancestor and descendant sets are defined as usual, and denoted by $\mathrm{adj}_{\mathcal{G}}(i)$, $\mathrm{pa}_{\mathcal{G}}(i)$, $\mathrm{an}_{\mathcal{G}}(i)$ and $\mathrm{des}_{\mathcal{G}}(i)$, respectively, for the vertex i in graph $\mathcal{G}$. In this study, we use the definition for m-separation and m-active paths in MAGs from Zhang (2008), and denote

$$i \overset{\mathrm{m}}{\underset{\mathcal{G}}{\perp}} j$$

when i and j are m-separated in $\mathcal{G}$. Furthermore, each $i \in I$ corresponds to a random variable X_i and is defined on probability space $\mathcal{X}_i$. For $Z \subseteq I$, the joint random variable X_Z is defined on the joint probability space $\mathcal{X}_Z = \prod_{i \in Z} \mathcal{X}_i$. We denote $X_i \perp\!\!\!\perp X_j \mid X_Z$ for two random variables X_i and X_j conditionally independent given X_Z, where $i, j \in I$ and $Z \subseteq I$.

Let $\mathcal{G}$ be a MAG with vertex set I. Let $\mathcal{D}$ be the corresponding *canonical DAG* (directed acyclic graph) obtained by adding a confounder for each bidirected edge to get the augmented vertex set $J \supseteq I$ (Richardson, 1996). Finally, let X_J be a joint random variable. We suppose that our data can be described as a *structural causal model* (SCM) corresponding to $\mathcal{D}$ and X_J (as defined in Peters et al. (2017)), i.e.,

$$X_i = f_i(X_{\mathrm{pa}_{\mathcal{D}}(i)}, \epsilon_i) \tag{1}$$

for every $i \in J$, where the noise terms ϵ_i are pairwise independent. Moreover, we call this model a *local additive noise model* (lANM) for $Z \subseteq J$ if, for every $i \in Z$, there exists a function h_i so that

$$X_i = h_i(X_{\mathrm{pa}_{\mathcal{D}}(i) \cap Z}) + \eta_i, \tag{2}$$

1. Here, we refer to the maximally informative PAG defined in (Zhang, 2008) simply as PAG.

where the pairwise independent noise terms η_i are additive to $h_i(X_{\mathrm{pa}_{\mathcal{D}}(i)\cap Z})$ (Tillman et al., 2009). For $i \in J$, $Z \subseteq J \setminus \{i\}$, assume that $g_i(X_Z)$ is a function regressing X_i on X_Z. The *residual* w.r.t. $g_i(X_Z)$ is computed as follows

$$\mathrm{R}(X_i|X_Z) := X_i - g_i(X_Z).$$

In this study, we assume that the regression function is close to the conditional expectation $\mathbb{E}[X_i|X_Z]$. If there is an lANM for Z, then for every pair $i, j \in Z$

$$\mathrm{R}(X_i|X_{Z\setminus\{i\}}) \perp\!\!\!\perp X_j. \tag{3}$$

In this case, $\mathrm{R}(X_i|X_{Z\setminus\{i\}})$ equals to the noise η_i from equation (2).

Throughout the paper we assume faithfulness, the Markov condition, acyclicity and no selection bias. Proofs of all theorems can be found in the appendix.

3. Causal Direction Criterion

In this section, we introduce a cause-effect inference orientation rule, the *causal direction criterion* (CDC) to infer pairwise causal relationships without assuming causal sufficiency. The CDC is inspired by the orientation procedure described in section 4 of Hoyer et al. (2008) and by the Causal Direction Determination Rule in Huang et al. (2020), which are similar to the CDC but operate under the assumption of (pseudo) causal sufficiency.

Throughout this section, we let $\mathcal{G}$ be a MAG with an underlying DAG and corresponding SCM with joint random variable X_I defined as in equation (1). As shown in Hoyer et al. (2008); Huang et al. (2020), there is asymmetric statistical information implied by a causal relationship. Thus, when applying the CDC for i and j adjacent in $\mathcal{G}$, we search for the following asymmetry

$$A(j \to i \mid Z) := \mathrm{R}(X_i|X_{\{j\}\cup Z}) \perp\!\!\!\perp X_j \wedge \mathrm{R}(X_j|X_{\{i\}\cup Z}) \not\perp\!\!\!\perp X_i \tag{4}$$

for a set $Z \subset I$, where $A(j \to i \mid Z)$ is a Boolean expressing if the asymmetry in equation (4) holds. If a set Z satisfies $A(j \to i \mid Z)$, we call it a *causal asymmetry set* (CAS) for the ordered pair (i, j).

To decide if j is a parent of i, we search for an independence between the residual $\mathrm{R}(X_i|X_{\{j\}\cup Z})$ and the random variable X_j, and vice versa. In some cases, we will find that both tests output independence, e.g., when the causal relationship is unidentifiable. In other cases, we cannot guarantee to completely remove the dependence, e.g., when we are not able to remove all confounding information due to hidden confounders. However, it is very unlikely to find an asymmetry $A(j \to i \mid Z)$ if j is not a parent of i, as we will see shortly below. This motivates the following assumption:

Assumption 1 *If $j \leftarrow i$ or $j \leftrightarrow i$ in $\mathcal{G}$, then there exists no CAS for the ordered pair (i, j).*

This assumption is violated only if we find the asymmetry in the wrong direction, which is impossible for a large class of models, and extremely unlikely to occur by accident in general.

Causal Direction Criterion *Let $i, j \in I$ be adjacent in $\mathcal{G}$ and $Z \subseteq I\setminus\{i, j\}$. We define the following orientation rule:*

$$\exists\, Z \text{ such that } A(j \to i \mid Z) \implies j \to i.$$

Note that then Z is a CAS for (i, j) in $\mathcal{G}$.

The asymmetry in the CDC is not always identifiable, although crucially this will not lead to wrong conclusions. Suppose that we have the lANM defined as in equation (2). An lANM is not identifiable if the function h_i is linear and the noise η_i is Gaussian, or if h_i is constant (Peters et al., 2017). Almost all other lANMs are identifiable (Zhang and Hyvärinen, 2009; Peters and Bühlmann, 2014). Moreover, one can think of examples of pairs that do not have additive noise by construction of the causal mechanism. For example, when we cannot construct an $h_i(X_{\mathrm{pa}_{\mathcal{D}}(i)\cap Z})$ such that it is additive to η_i which can happen if Z is not a causal asymmetry set. In general, the additive noise model is an approximation of the real underlying process, and the method might not pick up on the noise if the estimated additive noise is too weak. Fortunately, this will not lead to any misleading conclusions, since we will find a dependence in both directions, i.e., the CDC does not trigger. This way, we are more conservative with the additional orientations that we can make, but we avoid making any wrong orientations if our data does not follow a (local) additive noise model. See Tillman et al. (2009) for a more detailed discussion. In the following theorem, we prove that the CDC is sound under our assumptions.

Lemma 1 *Suppose that the true causal structure is acyclic and that the random variables measured by the data satisfy Assumption 1. Then, the CDC is sound.*

To explore the possible causal directions between i and j in practice, we test for independence in equation (4) in both directions, $(i \to j)$ and $(j \to i)$, using the kernel conditional independence test (KCIT) (Zhang et al., 2011). We use a multilayer perceptron as the regression method for computing the residuals. The multilayer perceptron is a universal approximator (Hornik et al., 1989) that is capable of obtaining an arbitrary regression function in the large-sample limit.

In order to apply the CDC, we still need to find the CAS Z. For that, we will use the FCI output PAG to restrict our search space to the following sets:

Definition 2 *Let $j \to i$ in $\mathcal{G}$. A subset of vertices $Z \subset I$ is a* deconfounding set *in $\mathcal{G}$ for (i, j) iff $Z \subseteq I \backslash \mathrm{des}_{\mathcal{G}}(i)$ and Z blocks all paths in $\mathcal{G}$ that are into both i and j.*

Deconfounding sets are suitable candidates to potentially remove all confounding information from the residuals, since they block all paths into both i and j. Note that a deconfounding set is not unique, and can be augmented by non-descendants of i which are not in a path going into both i and j. Moreover, an adjustment set is also a deconfounding set, but a deconfounding set is not always an adjustment set (Perković et al., 2015), as it may contain nodes on a directed path from j to i.

Assumption 2 *If Z is a CAS for the pair (i, j), then there exists a deconfounding set $Z' \subseteq Z$ that is also a CAS.*

Assumption 2 excludes the possibility that the independence in the first term of equation (4) is found by conditioning on a descendant that exactly cancels out the dependence via unblocked paths going into both i and j. It also means that for the FCI-CDC algorithm in the next section, whenever we find a CAS it is guaranteed to be a deconfounding set. Essentially, Assumptions 1 and 2 extend standard faithfulness to the residuals in equation (4): it is not impossible, but highly unlikely for them to be violated in practice.

In contrast to Hoyer et al. (2008), we do not assume causal sufficiency by allowing to condition on the deconfounding set in the CDC. However, the principle of the CDC is similar to how Hoyer

et al. (2008) proposed to infer causal directions in ANMs, although the assumptions and implementations details differ. Hoyer et al. (2008) use the Hilbert-Schmidt independence criterion (Gretton et al., 2005) for statistical testing and Gaussian process regression for computing the residuals. Using the KCIT instead of the HSIC allows us to condition on multiple random variables, which is required when dealing with larger deconfounding sets.

The CDC on its own is a reliable cause-effect inference method, but it really shines when combined with a multivariate causal discovery algorithm. In the following section, we will show how we can use the CDC on top of the constraint-based causal inference method FCI to go beyond the Markov equivalence class and arrive at a more informative causal structure in the presence of latent confounders.

4. FCI with CDC

In this section, we describe how we use the CDC on top of the FCI algorithm. First, we run the FCI algorithm to obtain an output PAG $\mathcal{P}$, which represents a Markov equivalence class. We loop over the edges which are (partially) unoriented and test if they can be oriented using the CDC. We describe how we can narrow the search space for selecting possible deconfounding sets when applying the CDC. Once the CDC finds an asymmetry, we orient the respective edge, as well as possible additional edges. The additional orientations are made by using the *Parent Orientation Rule*, which is described below, and the standard FCI orientation rules, as described in Zhang (2008). Combining a constraint-based causal discovery method and a pairwise causal inference method is also proposed by Tillman et al. (2009) under the assumption of causal sufficiency. In this study, we show that with our approach we can relax this assumption.

4.1. Deconfounding Sets in a Markov equivalence class

To test for the CDC between each (partially) oriented pair, we need to select a deconfounding set. However, given a PAG $\mathcal{P}$ for a causal graph $\mathcal{G}$, we cannot always identify deconfounding sets from $\mathcal{P}$ alone. Nevertheless, we will show how to easily identify a set of candidate vertices that are guaranteed to be contained in a deconfounding set, if one exists. We search through possible deconfounding sets by increasing the size of the set with possible deconfounding vertices. We use the principle that, if $j \to i$ in a MAG, then a parent set of the effect i excluding the cause j is a deconfounding set for (i, j) (see Lemma 6 in the appendix). The set of *possible parents* of i in $\mathcal{P}$ is defined as follows

$$\mathrm{popa}_{\mathcal{P}}(i) := \{k \in \mathrm{adj}_{\mathcal{P}}(i) \mid k \to i,\, k \circ\!\!\to i \text{ or } k \circ\!\!-\!\!\circ i\},$$

i.e., vertices k adjacent to i oriented as $k \to i$ in some MAG instance $\mathcal{M}$ in the equivalence class represented by $\mathcal{P}$. From Lemma 6 and Corollary 7 in the appendix, it follows that we can limit our search space to (possible) parents of i or j in the PAG $\mathcal{P}$ that are on an unblocked path going into both i and j that does not contain the direct edge between i and j. In other words, by testing if $z \in \mathrm{popa}_{\mathcal{P}}(i)$ and z is m-connected[2] to j given $Z \backslash \{z\}$ in $\mathcal{P}$ with i removed, we find that z is in

2. Here *m-connected in the PAG* $\mathcal{P}$ means m-connected in a MAG instance $\mathcal{M}$ in the equivalence class represented by $\mathcal{P}$. Likewise, we define m-separation in a PAG by m-separation in all MAG instances in the equivalence class represented by $\mathcal{P}$.

a possible minimal deconfounding set[3] for (i, j). In the special case of a partial orientation $j \circ\!\!\rightarrow i$ in $\mathcal{P}$, we can restrict our search space by selecting subsets of the possible parents of i, as shown in Corollary 8 in the appendix. In addition, we can recognize vertices that are guaranteed to be valid in any deconfounding set from the possible parents of i or j, as illustrated in the following lemma.

Lemma 3 *Assume $j \rightarrow i$ in $\mathcal{G}$, and that there exists a deconfounding set $Z \subseteq \mathrm{pa}_{\mathcal{G}}(i)$ for (i, j). Let $z \in \mathrm{popa}_{\mathcal{P}}(i)$ in PAG $\mathcal{P}$. Then if either*

(i) $i \leftarrow\!\circ z \rightarrow j$, $i \leftarrow z \circ\!\!\rightarrow j$, or $i \leftarrow z \rightarrow j$ in $\mathcal{P}$, or

(ii) there exist $z_1, \ldots, z_k \in I$ with $k \geq 1$ such that $\langle i, z, z_1, \ldots, z_k, j \rangle$ is an unshielded non-collider path[4] into both i and j in $\mathcal{P}$,

then $Z \cup \{z\}$ is also a deconfounding set for (i, j).

Note that, if Lemma 3(i) is satisfied for a vertex z, then z is necessarily a part of every deconfounding set $Z \subseteq \mathrm{popa}_{\mathcal{P}}(i)$. However, this is not the case if Lemma 3(ii) is satisfied, as other parents of i may already block the unshielded non-collider path. However, in both cases z is contained in a deconfounding set $Z \subseteq \mathrm{popa}_{\mathcal{P}}(i)$.

Now, we define the steps to search for a deconfounding set given an incompletely oriented edge $j \circ\!\!\rightarrow i$ or $j \circ\!\!-\!\!\circ i$ in $\mathcal{P}$. Suppose that $\mathcal{M}$ is a MAG instance in the equivalence class represented by $\mathcal{P}$. Note that by definition of the Markov equivalence class, all MAGs in the same equivalence class have the same m-separations. We will remove all edges adjacent to i in the MAG $\mathcal{M}$ and denote the resulting MAG by $\mathcal{M}_{-i}$.

D1 Let $Q_i = \mathrm{popa}_{\mathcal{P}}(i)$.

D2 While there exists a $z \in Q_i$ such that $z \overset{\mathrm{m}}{\underset{\mathcal{M}_{-i}}{\perp}} j \mid Q_i \backslash \{z\}$: update $Q_i = Q_i \backslash \{z\}$.

D3 Let $P_i \subset Q_i$ be the subset of vertices $z \in Q_i$ that satisfies Lemma 3(i).

D4 For each $Z \subset Q_i \backslash P_i$:

(i) if the CDC triggers for $j \rightarrow i$ with deconfounding set $P_i \cup Z$, orient $j \rightarrow i$ in $\mathcal{P}$ and go to next edge, or

(ii) if the CDC triggers for $j \leftarrow i$ with deconfounding set $P_i \cup Z$ and $j \circ\!\!-\!\!\circ i$ in $\mathcal{P}$[5], orient $i \rightarrow j$ in $\mathcal{P}$ and go to next edge

Otherwise, if $j \circ\!\!-\!\!\circ i$ in $\mathcal{P}$, then repeat D1-D4 for Q_j.

If there exists a deconfounding set, we can find one by repeatedly applying the CDC on the pair given the currently selected possible deconfounding set. We will find a CAS[6] once an asymmetry has been found by the CDC.

3. Huang et al. (2020) define the *potential* deconfounding set as a set containing all possible but not definite deconfounding vertices. However, in this study, a *possible* deconfounding set contains definite *and* possible deconfounding vertices.
4. See Zhang (2008) for the definition of an unshielded non-collider path.
5. We do not overrule the edge marks oriented by the FCI algorithm.
6. A CAS is not necessarily the smallest possible deconfounding set.

4.2. FCI-CDC Orientation Rules

Here we describe the FCI-CDC procedure. Besides the orientation of $j \to i$ found by applying CDC given deconfounding set Z for (i, j), we may orient even more edges with the following rule.

Parent Orientation Rule *Let $\mathcal{P}$ be a PAG. Suppose that i and j are adjacent in $\mathcal{P}$ and share a (partial) unoriented edge. If we orient $j \to i$ after running CDC on the pair (i, j) given a CAS Z, then for all $z \in \mathrm{adj}_{\mathcal{P}}(i) \cap Z$ we orient $z \to i$.*

Note that the parent orientation rule differs from the corresponding orientation rule in Algorithm 3 of Huang et al. (2020), because here we only orient the possible parents towards the effect i.

Lemma 4 *The parent orientation rule is sound under Assumption 2.*

After running CDC and the parent orientation rule, we will again run rules R1-R4' and R8-R10 of the FCI algorithm, as described in Zhang (2008), until no new orientations are made. Here, R4' implies that if there is a *discriminating path* between i and j for k, and $k \circ\!\!-\!\!\circ j$ or $k \circ\!\!\to j$ in the graph, then we orient $k \to j$. This is a simpler version of R4 from the standard FCI orientation procedure and is sufficient since the orientations triggered by the second part of R4 are invariant in all MAGs in the same equivalence class by Proposition 2 (Zhang, 2008). In conclusion, given a PAG $\mathcal{P}$ provided by (oracle) FCI, supposing that for an incompletely oriented edge we have $j \circ\!\!\to i$ or $j \circ\!\!-\!\!\circ i$ and a CAS Z, we find that the CDC triggers with implication $j \to i$. Then, the following orientation rules apply:

RC1 (CDC) Orient the edge between i and j in $\mathcal{P}$ as $j \to i$.

RC2 (parent orientation rule) For all $z \in \mathrm{adj}_{\mathcal{P}}(i) \cap Z$, orient $z \to i$.

RC3 (FCI) Propagate standard FCI orientation rules R1-R4',R8-R10 until exhaustion.

We call this procedure, in which we first apply the FCI algorithm and then run the rules RC1-RC3, the FCI-CDC algorithm. The FCI-CDC algorithm outputs an *overcomplete PAG* and is sound, as shown below. In the overcomplete PAG, each invariant edge mark is also present in the MAG of the underlying causal graph, and each variant edge mark is also present in the corresponding PAG.

Theorem 5 *Under Assumptions 1 and 2, and assuming oracle independence test results, the FCI-CDC algorithm is sound.*

4.3. Conservative FCI-CDC

The results above assume oracle information, i.e., a perfect independence test with infinite sample sizes. In practice, with finite data, mistakes are likely to be made, both during the traditional FCI stage as well as in the subsequent CDC stage, possibly leading to incorrect orientations and propagation thereof. One way of trying to guard against incorrect orientations in the CDC stage is to exploit redundant available information and to avoid making orientations when possible inconsistencies are encountered, similar to the ideas behind conservative PC/FCI. One straightforward approach is to continue the CDC search among the possible candidates after an asymmetry is found by the CDC to check for conflicting asymmetries, i.e., asymmetries that imply the orientation in the other direction, and to only orient if no inconsistencies are found.

5. Results

In this section, we showcase the performance of the CDC and of the FCI-CDC algorithm. We do not compare against methods like CD-NOD because they are not designed to handle (the same kind of) hidden confounders, and that is why we focus on the added benefits over vanilla FCI. We try to asses how much we can gain from the CDC and how reliable the extra information is.

In our simulation studies, we generate data of the form

$$X_i = f_i(X_{\mathrm{pa}(i)}) + \epsilon_i \tag{5}$$

where the operator f_i is randomly generated by a Gaussian Process (GP) with Gaussian kernel including a linear term. The noise ϵ_i is randomly generated from the uniform distribution $\mathcal{U}([-1, -\epsilon] \cup [\epsilon, 1])$ ($\epsilon = 0.01$), or from a heteroskedastic Gaussian distribution with variance sampled from the uniform distribution $\mathcal{U}([0.01, 2])$ for each data point and causal relationship. We standardized all data through

$$X_i' = \frac{X_i - \mathrm{mean}(X_i)}{\mathrm{std}(X_i)}$$

before running the algorithms.

5.1. Improvements of FCI-CDC over FCI

We study what can be gained in performance using the CDC on top of the FCI algorithm. To this end, we generate random MAGs with latent confounders and no cycles. The average node degree over all MAGs is 3.5, and around 20% of all edges are bidirected. Again, the data is generated as in equation (5), and we generate an additional random variable (or hidden confounder) for each bidirected edge. We compare all predicted edge marks with the edge marks in the ground truth MAGs. Recall and precision are computed by aggregating the predictions of all edge marks in all graphs, see appendix B for the details. We observe that if the data size increases for MAGs with 10 vertices, the recall of the FCI-CDC algorithm significantly increases, as shown in Figure 2. Moreover, the precision remains essentially the same for both the FCI algorithm and the FCI-CDC algorithm.

In order to illustrate what can be gained from the CDC, we ran the FCI algorithm with oracle conditional independence information. When adding the CDC on top, we considered the CDC with oracle causal relationship information, the vanilla CDC using finite-sample data, and the conservative CDC, respectively (Figure 3(*a*)). Our experiments are done on 100 generated MAGs, in a similar setting as before, with 10 vertices and 400 data points per data set. Note that the oracle CDC cannot predict bidirected edges, so some orientations remain unknown. For the vanilla CDC and the conservative CDC, we split the predicted edge marks per orientation rule RC1, RC2 and RC3, see Figure 3(*b*). Interestingly enough, there is only a small difference between the conservative CDC and the vanilla CDC. We observe that rules RC1 and RC3 are responsible for many correct orientations, whereas the rule RC2 (the parent orientation rule) does not get triggered that often. This happens because deconfounding sets for vertices sharing a partial unoriented edge rarely appear in PAGs, and because some deconfounding sets contain already oriented parents.

5.2. Deconfounding improves the performance of the CDC

We show how effective the CDC is by exploring the effects of conditioning on observed confounders. We generated 400 data sets with a causal relationship ($i \to j$) and 400 data sets with

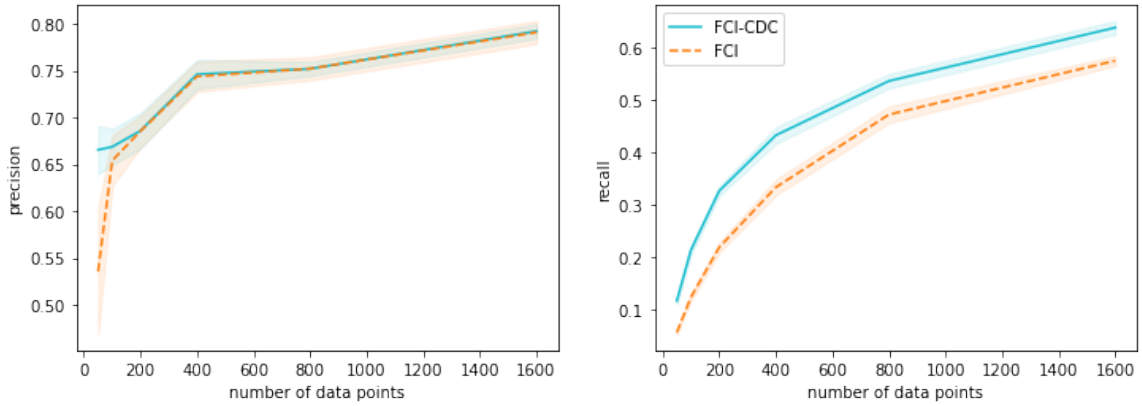
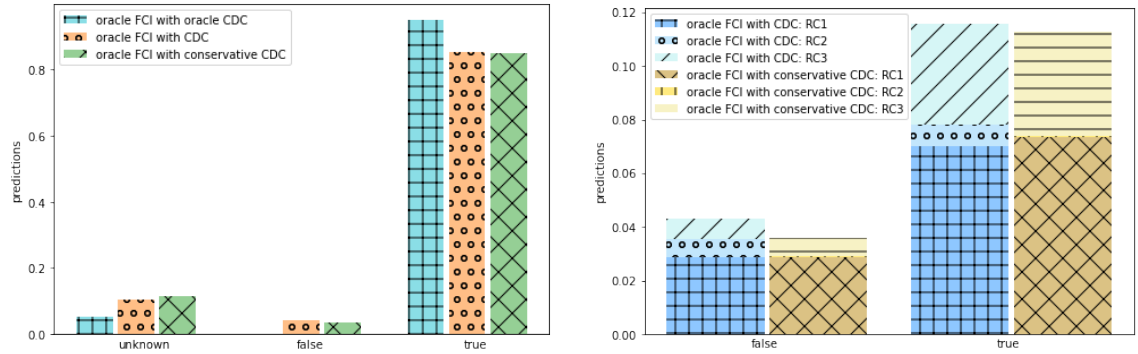


Figure 2: Average precision and recall of all edge marks across 100 generated graphs, each with 10 vertices, as a function of the number of data points for the FCI algorithm and the FCI-CDC algorithm. The bands correspond to 95% confidence intervals, obtained through bootstrapping.



(*a*) Comparison of the oracle, vanilla and conservative in proportions of predicted edge marks.

(*b*) Proportions of the predictions done by the orientation rules RC1, RC2 and RC3 for the vanilla CDC and the conservative CDC on top of the FCI algorithm.

Figure 3: The outputs of the FCI-CDC algorithm on 100 sets of 400 data points generated from graphs with 10 vertices. We ran the FCI algorithm with oracle conditional independence information, and the CDC with oracle causal relationship information, the vanilla version, and the conservative version.

a hidden confounder and no causal relationship ($i \leftrightarrow j$). In each data set, we also include two observed confounders, which the CDC can use as a possible deconfounding set. Each data set contains 400 data points.

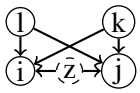
In Table 1, we compare the results of the CDC with and without regressing on the deconfounding set. On data sets with a causal relationship, regressing on the deconfounding set helps to increase the number of true predictions (correct direction). Even without deconfounding, the CDC picks

Table 1: For each of the two graphs above the tables, 400 data sets have been generated of 400 data points each. The graph on the left-hand side contains a causal relationship, the one on the right-hand side a hidden confounder. All graphs contain a deconfounding set of size two, that the CDC can either ignore (left columns) or make use of (right columns). The rows correspond to the predictions of the CDC. Conditioning on a deconfounding set helps the CDC to make more accurate predictions.

(*a*)

	without deconfounding	with deconfounding
→	81	119
←	52	56
?	267	225

(*b*)



	without deconfounding	with deconfounding
→	75	90
←	81	91
?	244	219

up on a potential asymmetry in the data and often guesses the right direction, where in theory for infinite data it should only give back unknowns. The CDC without deconfounding has a recall of 0.20 and a precision of 0.52, and the CDC with deconfounding has a recall of 0.30 and a precision of 0.57. This suggest that, with deconfounding, the CDC manages to make much fewer mistakes when choosing a direction. Because of sampling error, the CDC with deconfounding makes slightly more mistakes on the data sets with a hidden confounder, essentially because after deconfounding the effect of the hidden confounding is weaker than before.

In order to test the performance of the CDC under non-ideal settings, we did a sensitivity analysis. We generated pairs in the same setting as before, but only with linear interactions between the random variables and all noise terms have a heteroskedastic Gaussian distribution. Without deconfounding, we have a recall of 0.11 and a precision of 0.46. If we run the CDC with deconfounding, we obtain a recall of 0.30 and a precision of 0.66. Hence, the over all performance of the CDC with deconfounding is better if we have linear Gaussian data. In comparison to the other experiments, the CDC with deconfounding has a higher precision for linear Gaussian data, this can be explained by that the linear Gaussian relationship is not as pure anymore in the residuals or after normalizing the data.

6. Conclusion

In this study, we introduced the causal direction criterion (CDC) for extending causal discovery algorithms (such as FCI) beyond the Markov equivalence class in the presence of latent confounders. We saw that conditioning in CDC enabled us to distinguish between the effect of observed and latent confounders in a graphical structure, as shown in Table 1. The FCI-CDC algorithm, the extension to the FCI algorithm proposed in this paper, outputs an overcomplete PAG in which significantly more orientations are unraveled relative to the FCI algorithm (see Figure 3). We illustrated how we can apply the CDC using the notion of deconfounding sets, and how we can select a deconfounding set using information available in a PAG. The FCI-CDC algorithm achieves significantly better recall than the FCI algorithm for larger numbers of data points, as shown in Figure 2. In conclusion, the FCI-CDC algorithm has the potential to unravel more causal relationships than the FCI algorithm,

which makes it an important advancement in the field of causal inference where we have to handle the potential presence of hidden confounders.

For future research, it would be interesting to extend CDC to capture residuals of non-additive noise models, possibly motivated by the post-nonlinear causal models (Zhang and Hyvärinen, 2009). This would be a relevant extension, considering that cause-effect pairs in reality may have many hidden mediators in between, but then cannot be properly represented by additive noise models because these are not closed under marginalization (Peters et al., 2017, p. 138). Another interesting future direction would be an extension to a causal model with feedback loops.

References

Diogo Moitinho de Almeida. Pattern-based causal feature extraction. In *Cause Effect Pairs in Machine Learning*, pages 321–329. Springer, 2019.

Patrick Blöbaum, Dominik Janzing, Takashi Washio, Shohei Shimizu, and Bernhard Schölkopf. Cause-effect inference by comparing regression errors. In *International Conference on Artificial Intelligence and Statistics*, pages 900–909. PMLR, 2018.

Povilas Daniusis, Dominik Janzing, Joris Mooij, Jakob Zscheischler, Bastian Steudel, Kun Zhang, and Bernhard Schölkopf. Inferring deterministic causal relations. In *UAI'10: Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, pages 143–150, 2012.

José Fonollosa. Conditional distribution variability measures for causality detection. In *Cause Effect Pairs in Machine Learning*, pages 339–347. Springer, 2019.

Olivier Goudet, Diviyan Kalainathan, Philippe Caillou, Isabelle Guyon, David Lopez-Paz, and Michèle Sebag. Causal generative neural networks. *arXiv preprint arXiv:1711.08936*, 2017.

Olivier Goudet, Diviyan Kalainathan, Michèle Sebag, and Isabelle Guyon. Learning bivariate functional causal models. In *Cause Effect Pairs in Machine Learning*, pages 101–153. Springer, 2019.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *International Conference on Algorithmic Learning Theory*, pages 63–77. Springer, 2005.

Isabelle Guyon, Olivier Goudet, and Diviyan Kalainathan. Evaluation methods of cause-effect pairs. In *Cause Effect Pairs in Machine Learning*, pages 27–99. Springer, 2019.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

Patrik Hoyer, Dominik Janzing, Joris Mooij, Jonas Peters, Bernhard Schölkopf, et al. Nonlinear causal discovery with additive noise models. In *NIPS*, volume 21, pages 689–696, 2008.

Biwei Huang, Kun Zhang, Jiji Zhang, Joseph Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, 21(89):1–53, 2020.

Dominik Janzing and Bernhard Schölkopf. Causal inference using the algorithmic Markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010.

Dominik Janzing, Joris Mooij, Kun Zhang, Jan Lemeire, Jakob Zscheischler, Povilas Daniušis, Bastian Steudel, and Bernhard Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182:1–31, 2012.

Jan Lemeire and Erik Dirkx. Causal models as minimal descriptions of multivariate systems, 2006.

Joris Mooij and Dominik Janzing. Distinguishing between cause and effect. In *Causality: Objectives and Assessment*, pages 147–156. PMLR, 2010.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Emilija Perković, Johannes Textor, Markus Kalisch, Marloes H Maathuis, Marina Meila, Tom Heskes, et al. A complete generalized adjustment criterion. In *Uncertainty in Artificial Intelligence*, pages 682–691. AUAI Press, 2015.

Jonas Peters and Peter Bühlmann. Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014.

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

Thomas Richardson. *Models of Feedback: Interpretation and Discovery*. PhD thesis, Carnegie-Mellon University, 1996.

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. In *ICML'12: Proceedings of the 29th International Conference on Machine Learning*, pages 459–466, 2012.

Eleni Sgouritsa, Dominik Janzing, Philipp Hennig, and Bernhard Schölkopf. Inference of cause and effect with unsupervised inverse regression. In *Artificial Intelligence and Statistics*, pages 847–855. PMLR, 2015.

Shohei Shimizu, Patrik Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10): 2003–2030, 2006.

Peter Spirtes, Clark Glymour, Richard Scheines, and David Heckerman. *Causation, Prediction, and Search*. MIT Press, 2000.

Robert Tillman, Arthur Gretton, and Peter Spirtes. Nonlinear directed acyclic structure learning with weakly additive noise models. In *NIPS*, pages 1847–1855, 2009.

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

Kun Zhang and Aapo Hyvärinen. On the identifiability of the post-nonlinear causal model. In *25th Conference on Uncertainty in Artificial Intelligence (UAI 2009)*, pages 647–655. AUAI Press, 2009.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 804–813, 2011.

Appendix A. Theorems

We assume faithfulness, the Markov condition, acyclicity and no selection bias.

Lemma 1 *Suppose that the true causal structure is acyclic and that the random variables measured by the data satisfy Assumptions 1. Then, the CDC is sound.*

Proof Suppose that there exists an $Z \subseteq I \backslash \{i, j\}$ such that $A(j \to i \mid Z)$. Then, by Assumption 1, we find that $i \to j$ and $i \leftrightarrow j$ are not in $\mathcal{G}$. By assumption, i and j are adjacent, hence $j \to i$ in $\mathcal{G}$. ■

For the following lemma, let $\mathcal{G}$ be the ground truth MAG with corresponding PAG $\mathcal{P}$.

Lemma 3 *Assume $j \to i$ in $\mathcal{G}$ and that there exists a deconfounding set $Z \subseteq \mathrm{pa}_{\mathcal{G}}(i)$ for (i, j). Let $z \in \mathrm{popa}_{\mathcal{P}}(i)$ in PAG $\mathcal{P}$. Then, if either*

(i) $i \leftarrow\!\circ z \to j$, $i \leftarrow z \circ\!\!\to j$, or $i \leftarrow z \to j$ in $\mathcal{P}$, or

(ii) there exist $z_1, \ldots, z_k \in I$ with $k \geq 1$ such that $\langle i, z, z_1, \ldots, z_k, j \rangle$ is an unshielded non-collider path into both i and j in $\mathcal{P}$,

then $Z \cup \{z\}$ is also a deconfounding set for (i, j).

Proof Note that for both (i) and (ii), in the MAG $\mathcal{G}$ it holds that $z \subset (\mathrm{an}_{\mathcal{G}}(i) \cup \mathrm{an}_{\mathcal{G}}(j)) \setminus \{i, j\}$, which means that conditioning on z cannot unblock any paths from i to j. Hence, given that Z blocks all paths into both i and j, then $Z \cup \{z\}$ also blocks all paths into both i and j. ■

Lemma 4 *The parent orientation rule is sound under Assumption 2.*

Proof Suppose that we find $j \to i$ with deconfounding set Z by applying CDC. For each $z \in Z$ adjacent to i, if there is an edge $i \to z$ or $i \leftrightarrow z$ in the true causal graph, then either z is no ancestor of i, or the true causal graph will contain a cycle. This is a contradiction with the definition of the deconfounding set and with the acyclicity assumption, respectively. Hence, we should orient the edges as $z \to i$ for each $z \in Z$. ■

Theorem 5 *Under Assumptions 1 and 2, and assuming oracle independence test results, the FCI-CDC algorithm is sound.*

Proof Lemmas 1 and 4 show that RC1 and RC2 are sound. Furthermore, FCI is sound, and the orientation rules R1-R4 and R8-10 are sound. Note that in all MAG instances $\mathcal{M}$ in the equivalence class represented by the output PAG $\mathcal{P}$ of the FCI algorithm, the colliders on discriminating paths in all graphs are invariant by Proposition 2 in (Zhang, 2008). Therefore, orientation rule R4' is sound when applying it only on (overcomplete) PAGs. Hence, the orientation rule RC3 is sound. This shows that the complete procedure FCI-CDC is sound. ■

In Proposition 6.41, Peters et al. (2017) show that if $j \to i$ in a directed acyclic graph $\mathcal{G}$, and there exists an adjustment set Z for (i, j), then $\mathrm{pa}_{\mathcal{G}}(i) \setminus \{j\}$, is also an adjustment set. We can show similar results for deconfounding sets in MAGs.

Lemma 6 *If $j \to i$ in a MAG $\mathcal{G}$, and there exists a deconfounding set Z for (i, j), then $\mathrm{pa}_{\mathcal{G}}(i) \setminus \{j\}$, is also a deconfounding set.*

Proof Note that $\mathrm{pa}_{\mathcal{G}}(i) \setminus \{j\} \subseteq I \setminus \mathrm{des}_{\mathcal{G}}(i)$. Suppose that there is a path into both i and j. Then this path is of the form

$$j \leftarrow\!\!* k *\!\!-\!\!* \ldots *\!\!-\!\!* l *\!\!\rightarrow i.$$

When $l \to i$, then $l \in \mathrm{pa}_{\mathcal{G}}(i) \setminus \{j\}$ and, by including it in the deconfounding set, we block the path between i and j. Otherwise, if $l \leftrightarrow i$, l can be either be a collider or non-collider on the path. If l is a collider on the path, then the path is blocked since no parent of i can be a descendant of l (ancestrality violation). If l is a non-collider on the path, then either l is an ancestor of j, or there is a collider on the subpath between l and j. In the first case, we find that l is also ancestor of i because $j \to i$, which is again an ancestrality violation. In the second case, we have to consider the closest collider to l, let us denote this collider by c. Note that, since c is the closest collider, then l has to be ancestor of c, otherwise there must be another collider closer to l. We now show that c blocks the path between i and j. If either c or one of its descendants were in the parent set of j, then that would form an almost directed cycle containing l and i and going through c, which is yet another ancestrality violation.

Hence, paths into both i and j are blocked by the parents of i excluding j. This proves that $\mathrm{pa}_{\mathcal{G}}(i) \setminus \{j\}$ is a deconfounding set. ■

Corollary 7 *A deconfounding set Z for (i, j) is* minimal *if and only if every $z \in Z$ blocks a path into both i and j that is not blocked by any $k \in Z \setminus \{z\}$.*

Proof ($\Rightarrow$) Suppose that Z is a minimal deconfounding set for (i, j), i.e., there is no strict subset of Z that is a deconfounding set. Hence, $Z \setminus \{z\}$ is no deconfounding set for (i, j). This implies that there exists a path into both i and j that is blocked by z and is not blocked by any $k \in Z \setminus \{z\}$. ($\Leftarrow$) Suppose that Z is a deconfounding set for (i, j), and every $z \in Z$ blocks a path into both i and j that is not blocked by any $k \in Z \setminus \{z\}$. Let Z' be a strict subset of Z. Then, each $z \in Z \setminus Z'$ blocks a path p into both i and j which is not blocked by any $k \in Z'$. Therefore, Z' cannot be a deconfounding set. Hence, Z is a minimal deconfounding set. ■

Corollary 8 *If $j \circ\!\!\rightarrow i$ in PAG $\mathcal{P}$, and there is a deconfounding set for (i, j), then there exists a subset $Z \subseteq \mathrm{popa}_{\mathcal{P}}(i)$ in $\mathcal{P}$ that is a deconfounding set for (i, j).*

Proof Note that, if there exists a deconfounding set for (i, j) and $j \circ\!\!\rightarrow i$ in $\mathcal{P}$, then $j \to i$ in the ground truth MAG $\mathcal{G}$. Note that $\mathrm{pa}_{\mathcal{G}}(i)$ is a deconfounding set for (i, j) and $\mathrm{pa}_{\mathcal{G}}(i) \subseteq \mathrm{popa}_{\mathcal{P}}(i)$. Hence, there exists a subset $Z \subseteq \mathrm{popa}_{\mathcal{P}}(i)$ in $\mathcal{P}$ that is a deconfounding set for (i, j). ■

Appendix B. Metrics

In this appendix, we illustrate how we defined the metrics to measure the performance for the CDC and the FCI-CDC algorithm.

B.1. CDC

For the CDC, we need to compare the predicted pairwise relationships to the ground truth pairwise relationships. Note that there are three classes we can predict, namely a causal relationship in either direction or an unknown direction. Therefore, we consider an alternate way to define the correct and incorrect predicted classes.

We consider the situations wherein we have a causal relationship or a hidden confounder (and no causal relationship), i.e., the ground truth classes are causal relationships (right arrows[7] ($i \to j$)) and hidden confounders ($i \leftrightarrow j$). The predicted outcome classes are causal relationships in both directions (right arrows ($i \to j$), left arrows ($i \leftarrow j$)) and hidden confounders ($i \leftrightarrow j$). In Figure 4, we illustrate how the true (T) and false (F), right arrows (R), left arrows (L) and hidden confounders (C) outcomes are defined. For instance, TR is a true predicted right arrow, and FR_C is the false predicted right arrow when the ground truth is a hidden confounder (or bidirected edge). The elements in the inner circle in Figure 4 are called *retrieved* (TR, FR_C), and the elements in the left slice are called *relevant* (FC_R, FL_R, TR). We define the *recall* of the directions $i \to j$ as the number of 'true positive' divided by the relevant elements:

$$\frac{\text{TR}}{\text{TR} + \text{FC}_\text{R} + \text{FL}_\text{R}}. \tag{6}$$

and the *precision* of the directions $i \to j$ as the number of 'true positive' divided by the retrieved elements:

$$\frac{\text{TR}}{\text{TR} + \text{FR}_\text{C}}. \tag{7}$$

B.2. FCI-CDC

To measure the performance of the FCI algorithm and the FCI-CDC algorithm, we compare the predicted graphs to the ground truth MAGs. Note that the predicted graphs for both algorithms have the same skeleton. For all edges occurring in both the predicted graphs and ground truth MAGs, we define the following three classes:

(i) the *true positives* correspond to the correctly predicted edge marks,

(ii) the *false positives* correspond to the incorrectly predicted edge marks that are not predicted to be unknown, and

(iii) the *false negatives* correspond to the edge marks that are predicted incorrectly or as unknown.

Using these classes we can compute the precision and recall as usual.

7. By convention, we say that the ground truth causal relationship is $i \to j$, so the correct orientations are the arrows towards the right and the incorrect orientations are the arrows towards the left.

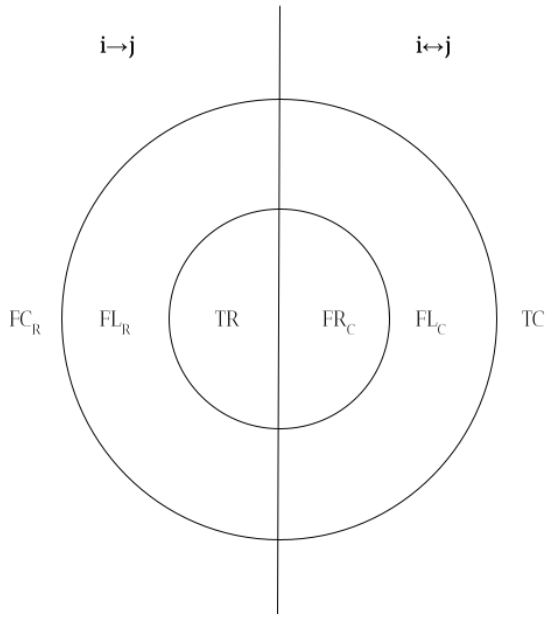


Figure 4: In this diagram we defined true (T) and false (F), right arrow ($i \rightarrow j$) (R), left arrow ($i \leftarrow j$) (L) and hidden confounder ($i \leftrightarrow j$) (C). The diagram is divided into two parts, of which the left slice has $i \rightarrow j$ as ground truth, and the right slice has $j \leftrightarrow i$ as ground truth. For instance, we write TR for the true predicted right arrows $i \rightarrow j$, and we write FR_C for the false predicted right arrows that are hidden confounders.