IDA with Background Knowledge

Zhuangyan Fang, Yangbo He*
School of Mathematical Sciences
Peking University
Beijing, China

Abstract

In this paper, we consider the problem of estimating all possible causal effects from observational data with two types of background knowledge: direct causal information and non-ancestral information. Following the IDA framework, we first provide locally valid orientation rules for maximal partially directed acyclic graphs (PDAGs), which are widely used to represent background knowledge. Based on the proposed rules, we present a fully local algorithm to estimate all possible causal effects with direct causal information. Furthermore, we consider non-ancestral information and prove that it can be equivalently transformed into direct causal information, meaning that we can also locally estimate all possible causal effects with non-ancestral information. The test results on both synthetic and real-world data sets show that our methods are efficient and stable.

1 INTRODUCTION

Directed acyclic graphs (DAGs) are widely used in causal inference. When the underlying causal DAG is fully specified by background knowledge (Meek, 1995) or experimental data (He & Geng, 2008; Hauser & Bühlmann, 2012), the causal effect of a treatment on a target can be estimated from observational data using the back-door adjustment criterion (Pearl, 2009). However, with observational data, one can only learn a completely partially directed acyclic graph (CPDAG) representing a class of Markov equivalent DAGs (Spirtes et al., 2000), making it difficult to identify all causal effects since equivalent DAGs may entail different causal relations.

*Correspondence to: heyb@pku.edu.cn.

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

To estimate causal effects from observational data without a fully specified DAG, some researchers focus on the identifiability of a causal effect (Perković et al., 2015, 2017; Perković et al., 2018; Jaber et al., 2018a,b, 2019). Since not all causal effects can be uniquely identified, an alternative approach is to learn a CPDAG first, then enumerate all DAGs in the learned Markov equivalence class and estimate the causal effect for each of those equivalent DAGs (Maathuis et al., 2009). For any treatment-target pair, this method returns a *multi-set* of all possible causal effects of the treatment on the target. Since enumerating all DAGs is infeasible when the size of the Markov equivalence class is large (He et al., 2015), Maathuis et al. (2009) further proposed a local algorithm called IDA to estimate the multi-set. Instead of enumerating all DAGs, IDA only enumerates possible parental sets of the treatment, which is shown to be efficient since enumerating possible parental sets only requires the local structure around the treatment (Maathuis et al., 2009).

Incorporating background knowledge into causal inference has drawn more and more attentions in recent years (Perković et al., 2017; Henckel et al., 2019; Perković, 2019). In real applications, practitioners usually have prior knowledge about the causal system. For example, if the causal system is related to time, we may assume that the subsequent events are not the causes of the prior events. In social sciences, it is reasonable to assume that intrinsic attributes, such as gender and race, are not affected by other variables. In medical sciences, previous studies may indicate that some behaviors will definitely cause some diseases, like smoking causes bronchitis or eating betel nuts causes oral cancer. Recently, Perković et al. (2017) extended IDA to deal with the cases where direct causal information is available. They proposed a semi-local algorithm to enumerate all possible causal effects. However, the semi-local IDA needs the entire CPDAG instead of the local structure around the treatment to check the validity of a possible parental set, which limits the application of the semi-local IDA to high dimensional systems.

In this paper, we consider the problem of estimating all possible causal effects from observational data with background knowledge. Our paper extends the work of Maathuis et al. (2009) and Perković et al. (2017), and has the following contributions:

- We provide locally valid orientation rules for maximal partially directed acyclic graphs (PDAGs), which is sufficient and necessary to check whether a set of variables in a maximal PDAG can be the parents of a given target.
- Based on the proposed rules, we give a fully local algorithm to enumerate all possible causal effects with direct causal information.
- We prove that non-ancestral information can be equivalently transformed into direct causal information, making it possible to locally enumerate all possible causal effects with non-ancestral information.

2 PRELIMINARIES

In this section, we introduce the notation, definitions and related work.

2.1 CAUSAL GRAPHICAL MODELS

A graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ is *directed* (*undirected*, or *partially directed*) if all edges in the graph are directed (undirected, or a mixture of directed and undirected ones). The *skeleton* of $\mathcal{G}$ is an undirected graph obtained from removing all arrowheads in $\mathcal{G}$. For any $\mathbf{V}' \subset \mathbf{V}$, the *induced subgraph* of $\mathcal{G}$ over $\mathbf{V}'$ is the graph with vertex set $\mathbf{V}'$ and edge set $\mathbf{E}'$, where $\mathbf{E}' \subset \mathbf{E}$ contains all and only edges between vertices in $\mathbf{V}'$.

Given a graph $\mathcal{G}$, X_i is a *parent* of X_j and X_j is a *child* of X_i if $X_i \to X_j$ in $\mathcal{G}$, and X_i is a *sibling* of X_j if $X_i - X_j$ in $\mathcal{G}$. If there is an edge between X_i and X_j, then they are *adjacent*. We use $pa(X_i, \mathcal{G})$, $ch(X_i, \mathcal{G})$, $sib(X_i, \mathcal{G})$, and $adj(X_i, \mathcal{G})$ to denote the sets of parents, children, siblings, and adjacent vertices of X_i in $\mathcal{G}$, respectively. A graph is called *complete* if every two distinct vertices are adjacent. A *path* is a sequence of distinct vertices $(X_{k_1}, \cdots, X_{k_j})$ such that any two consecutive vertices are adjacent. If every two distinct vertices in a graph are connected by a path, then the graph is called connected. A path is called *partially directed* from X_{k_1} to X_{k_j} if $X_{k_i} \leftarrow X_{k_{i+1}}$ does not occur in $\mathcal{G}$ for any $i = 1, \ldots, j-1$. A partially directed path is *directed* (*undirected*) if all edges on the path are directed (undirected). A (partially directed, directed, or undirected) cycle is a (partially directed, directed, or undirected) path from a vertex to itself. The length of a path (cycle) is the number of edges on the path (cycle). Particularly, a cycle with length three is called a triangle. A vertex X_i is an *ancestor* of X_j and X_j is a *descendant* of X_i if there is a directed path from X_i to X_j or $X_i = X_j$; the sets of all ancestors and all descendants of X_i in a graph $\mathcal{G}$ are denoted by $an(X_i, \mathcal{G})$ and $de(X_i, \mathcal{G})$, respectively. A *chord* of a path (cycle) is any edge joining two nonconsecutive vertices on the path (cycle). A path (cycle) without any chord is called *chordless*[1]. An undirected graph is *chordal* if it has no chordless cycle with length greater than three. A directed graph is *acyclic* (DAG) if there are no directed cycles.

The notion of *d-separation* induces a set of conditional independence relations encoded in a DAG (Pearl, 1988). Two DAGs are Markov *equivalent* if they induce the same set of conditional independence relations. For three distinct vertices X_i, X_j and X_k, if $X_i \to X_j \leftarrow X_k$ and X_i is not adjacent to X_k in $\mathcal{G}$, then the triple (X_i, X_j, X_k) is called a *v-structure* collided on X_j. Pearl et al. (1989) have shown that two DAGs are equivalent if and only if they have the same skeleton and the same v-structures. A *Markov equivalence class* or simply *equivalence class*, denoted by $[\mathcal{G}]$, contains all DAGs equivalent to $\mathcal{G}$. A Markov equivalence class $[\mathcal{G}]$ can be uniquely represented by a partially directed graph called *completely partially directed acyclic graph* (CPDAG) $\mathcal{G}^*$, in which two vertices are adjacent if and only if they are adjacent in $\mathcal{G}$ and a directed edge occurs if and only if it appears in every DAG in $[\mathcal{G}]$ (Pearl et al., 1989). Given a CPDAG $\mathcal{G}^*$, we use $\mathcal{G}^*_u$ and $\mathcal{G}^*_d$ to denote the *undirected subgraph* and *directed subgraph* of $\mathcal{G}^*$, respectively. The former is defined as the undirected graph resulted by removing all directed edges in $\mathcal{G}^*$, and the later is the directed graph obtained by removing undirected edges. Andersson et al. (1997) proved that $\mathcal{G}^*$ is a chain graph, which means, (1) the undirected subgraph $\mathcal{G}^*_u$ of $\mathcal{G}^*$ is the union of disjoint connected chordal graphs, and (2) every partially directed cycle is an undirected cycle in $\mathcal{G}^*$. The isolated connected chordal graphs of $\mathcal{G}^*_u$ are called *chain components* of $\mathcal{G}^*$ (Andersson et al., 1997).

A causal DAG model consists of a DAG $\mathcal{G}$ and a joint distribution P over a common set $\mathbf{V}$ such that P satisfies the *causal Markov assumption* with respect to $\mathcal{G}$, which requires that P can be factorized as,

$$P(x_1, ..., x_n) = \prod_{i=1}^{n} P(x_i | pa(x_i, \mathcal{G})).$$

In this paper, we also assume that there is no hidden variable or selection bias, and a CPDAG representing the Markov equivalence class containing the underlying

[1]The word 'chordless' is mostly used in graph theory (see, e.g. Blair & Peyton, 1993), while in some papers, such paths are called 'unshielded' (see, e.g. Perković et al., 2017)

causal DAG can be recovered from data.[2]

2.2 INTERPRETING BACKGROUND KNOWLEDGE

Background information can be regarded as a set of constraints. In this paper, we consider both direct causal information (Meek, 1995; Perković et al., 2017) and non-ancestral information. A *direct causal claim*, denoted by $X \to Y$, is defined as a constraint which requires X to be a direct cause of Y. Likewise, a *non-ancestral claim*, denoted by $X \nrightarrow Y$, is defined as a constraint which requires X to be a non-ancestor of Y. A direct causal information set is a set of direct causal claims, and a non-ancestral information set is a set of non-ancestral claims. We use $\mathcal{B}_d$, $\mathcal{B}_n$ and $\mathcal{B}$ to denote a direct causal information set, a non-ancestral information set, and an (arbitrary) background knowledge set, respectively.

For a CPDAG $\mathcal{G}^*$, any DAG obtained by orienting the undirected edges in $\mathcal{G}^*$ without creating new v-structures or directed cycles is a member of the equivalence class represented by $\mathcal{G}^*$ (Pearl et al., 1989; Meek, 1995). Let $\mathcal{B}$ denote a background knowledge set related to the true underlying causal DAG. With the constraints in $\mathcal{B}$, we may further reduce the number of possible DAGs including the true one. More formally, a set of constraints $\mathcal{B}$ is *consistent* with a given CPDAG $\mathcal{G}^*$ if there is at least one DAG $\mathcal{G}$ in the Markov equivalence class represented by $\mathcal{G}^*$ such that $\mathcal{G}$ satisfies all constraints in $\mathcal{B}$. If $\mathcal{B}$ is consistent with $\mathcal{G}^*$, then the subset of equivalent DAGs satisfying all constraints in $\mathcal{B}$ is called a *restricted Markov equivalence class* with respect to $\mathcal{G}^*$ and $\mathcal{B}$.

Representing background knowledge graphically can bring a lot of convenience. Clearly, given a CPDAG $\mathcal{G}^*$, a direct causal information set can be equivalently interpreted by orienting corresponding undirected edges in $\mathcal{G}^*$, resulting a partially directed graph $\mathcal{H}$. For simplicity, we say $\mathcal{H}$ (or orientations of some undirected edges in $\mathcal{G}^*$) is *consistent* with $\mathcal{G}^*$ if the corresponding direct causal information is consistent with $\mathcal{G}^*$, and the corresponding restricted Markov equivalence class is represented by $\mathcal{H}$. Meek (1995) proved that, with a series of orientation rules called Meek's criteria, some undirected edges in a consistent $\mathcal{H}$ may be further directed (see Algorithm 6 in Appendix B.1 for details), and the resulting graph is a *maximal partially directed acyclic graph* (maximal PDAG), where two distinct vertices X and Y are adjacent if and only if they are adjacent in $\mathcal{G}^*$, and $X \to Y$ appears if and only if $X \to Y$ appears in every DAG in the restricted Markov equivalence class represented by $\mathcal{H}$. Conversely, if $\mathcal{H}$ is inconsistent, then the resulting graph is not a maximal PDAG.

[2]Note that recovering CPDAG from observational data may need additional assumptions.

Algorithm 1 The IDA algorithm

Require: A CPDAG $\mathcal{G}^*$, a target variable Y.
Ensure: $\{\Theta_X\}_{X\in\mathbf{V}}$, where Θ_X stores all possible causal effects of X on Y.

```
for each variable X ∈ V do
    set Θ_X = ∅,
    for each S ⊂ sib(X, G*) such that orienting S → X and X → sib(X, G*) \ S does not introduce any v-structure collided on X do
        estimate the causal effect of X on Y by adjusting for S ∪ pa(X, G*), and add the causal effect to Θ_X,
    end for
end for
return {Θ_X}_{X∈V}.
```

2.3 CAUSAL INFERENCE

Given a DAG $\mathcal{G}$ and two distinct variables X and Y, the causal effect of X on Y can be interpreted by the post-intervention distribution of Y intervening on X via *do* operator (Pearl, 1995, 2009). With observational data, if $Y \notin pa(X, \mathcal{G})$, then the post-intervention distribution can be calculated from the pre-intervention distribution by:

$$P(y|do(X = x)) = \int P(y|X = x, pa(x))P\left(pa(x)\right) d\left(pa(x)\right). \quad (1)$$

If $Y \in pa(X, \mathcal{G})$, then $P(y|do(X = x)) = P(y)$. Equation (1) is a special case of *back-door adjustment* (Pearl, 1995, 2009), and $pa(x, \mathcal{G})$ is a special *back-door adjustment set*. However, if we only know a CPDAG $\mathcal{G}^*$, the causal effect of X on Y may not be identifiable from observational data. To address this problem, Maathuis et al. (2009) provided a novel framework called IDA. As shown in Algorithm 1, IDA enumerates all possible causal effects of X on Y by listing all possible parental sets and adjusting for each of them. To decide whether a set of variables is possible to be the parents of X, Maathuis et al. (2009) provided a locally valid orientation rule.

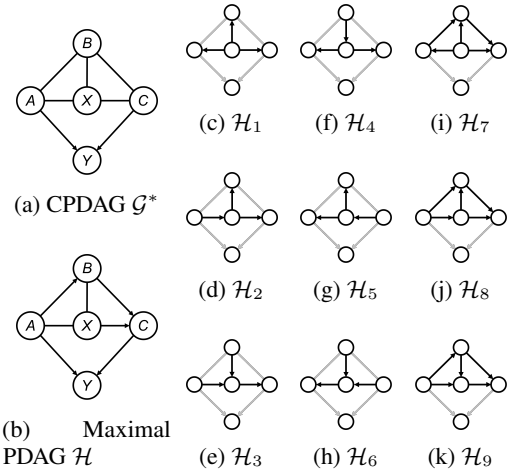
Lemma 1 (Maathuis et al., 2009, Lemma 3.1) *Given a CPDAG $\mathcal{G}^*$, a variable X, and $\boldsymbol{S} \subset sib(X, \mathcal{G}^*)$, orienting $S \to X$ for each $S \in \boldsymbol{S}$ and $X \to C$ for each $C \in sib(X, \mathcal{G}^*) \setminus \boldsymbol{S}$ is consistent with $\mathcal{G}^*$ if and only if new orientations do not introduce v-structures collided on X.*

For simplicity, below we will use $\mathbf{A} \to \mathbf{B}$ for two disjoint sets $\mathbf{A}$ and $\mathbf{B}$ to denote that for any $A \in \mathbf{A}$ and $B \in \mathbf{B}$, $A \to B$. Thanks to Lemma 1, although IDA needs a CPDAG as input, it only needs the local structure around

Algorithm 2 The semi-local IDA algorithm

Require: A CPDAG $\mathcal{G}^*$, a consistent direct causal information set $\mathcal{B}_d$, a target variable Y.
Ensure: $\{\Theta_X\}_{X\in\mathbf{V}}$, where Θ_X stores all possible causal effects of X on Y.

1: Construct the maximal PDAG $\mathcal{H}$ from $\mathcal{G}^*$ and $\mathcal{B}_d$ using Meek's criteria,
2: **for** each variable $X \in \mathbf{V}$ **do**
3: set $\Theta_X = \emptyset$,
4: **for** each $\mathbf{S} \subset sib(X, \mathcal{H})$ **do**
5: orient $\mathbf{S} \to X$ and $X \to sib(X, \mathcal{H}) \setminus \mathbf{S}$ in $\mathcal{H}$, and denote the resulting graph by $\mathcal{H}_{\mathbf{S}\to X}$,
6: using Meek's criteria to check whether $\mathcal{H}_{\mathbf{S}\to X}$ is consistent with $\mathcal{G}^*$,
7: **if** $\mathcal{H}_{\mathbf{S}\to X}$ is consistent with $\mathcal{G}^*$ **then**
8: estimate the causal effect of X on Y by adjusting for $\mathbf{S} \cup pa(X, \mathcal{H})$, and add the causal effect to Θ_X,
9: **end if**
10: **end for**
11: **end for**
12: **return** $\{\Theta_X\}_{X\in\mathbf{V}}$.



(a) CPDAG $\mathcal{G}^*$ (b) Maximal PDAG $\mathcal{H}$ (c) $\mathcal{H}_1$ (d) $\mathcal{H}_2$ (e) $\mathcal{H}_3$ (f) $\mathcal{H}_4$ (g) $\mathcal{H}_5$ (h) $\mathcal{H}_6$ (i) $\mathcal{H}_7$ (j) $\mathcal{H}_8$ (k) $\mathcal{H}_9$

Figure 1: An example to show that the rule in Lemma 1 is no longer valid for maximal PDAGs. Figure 1a shows a CPDAG, and Figure 1b shows the maximal PDAG when adding $A \to B$ to $\mathcal{G}^*$. Figures 1c to 1h enumerate all possible parental sets of X without background knowledge. Figures 1i to 1k enumerate all possible parental sets of X with direct causal information $A \to B$.

the treatment to list all possible parental sets and estimate all possible causal effects. The results are stored in a *multi-set* Θ_X, which can be regarded as an unordered list.

Recently, Perković et al. (2017) proposed the semi-local IDA which can semi-locally find all possible parental sets of a treatment in a maximal PDAG and then estimate all possible causal effects by adjusting for each of them. Algorithm 2 shows the schema. Different from IDA, Algorithm 2 uses Meek's criteria to check the validity of candidate parents (line 6). However, Meek's criteria are global orientation rules and require an entire $\mathcal{H}$ as input.

3 INCORPORATING DIRECT CAUSAL INFORMATION

In this section, we study the locally valid orientation rules for maximal PDAGs, and present a fully local algorithm for estimating all possible causal effects with direct causal background information.

3.1 LOCALLY VALID ORIENTATION RULES FOR MAXIMAL PDAGS

Let $\mathcal{G}^*$ be a CPDAG learned from data, and $\mathcal{B}_d$ denote a direct causal information set which is consistent with $\mathcal{G}^*$. As discussed earlier, one can use a maximal PDAG $\mathcal{H}$ to interpret $\mathcal{B}_d$. Therefore, the key step for estimating all possible causal effects locally is to develop locally valid orientation rules for maximal PDAGs. The following example demonstrates that the rule in Lemma 1 is no longer valid. That is, the criterion in Lemma 1 may cause directed cycles when applied to a maximal PDAG.

Example 1 *Consider the graphs in Figure 1. Given a CPDAG $\mathcal{G}^*$ in Figure 1a, we would like to estimate all possible causal effects of X on Y using IDA. From $\mathcal{G}^*$ we can see that $sib(X, \mathcal{G}^*) = \{A, B, C\}$. Clearly, there are 8 different subsets of $sib(X, \mathcal{G}^*)$. However, neither $\{A, B, C\}$ nor $\{A, C\}$ can be a parental set of X based on Lemma 1, since $A \to X \leftarrow C$ is a new v-structure. Hence, there are 6 possible parental sets of X, which are listed in Figures 1c to 1h. Now, assume that we know A is a direct cause of B in the underlying DAG. With this background knowledge, we orient $A - B$ in $\mathcal{G}^*$ as $A \to B$. Furthermore, based on Meek's criteria, we can further orient $B \to C$ and $X \to C$, which results the maximal PDAG $\mathcal{H}$. In this case, $sib(X, \mathcal{H}) = \{A, B\}$. Obviously, setting the parents of X to be $\emptyset$, $\{A\}$, $\{B\}$, or $\{A, B\}$ does not introduce a new v-structure, but only three of them are valid, since letting B be the parent and A be the child would cause a directed cycle $A \to B \to X \to A$.*

Example 1 shows that when orienting undirected edges connected to a treatment in a maximal PDAG, it is not only necessary to avoid creating new v-structures, but also important to avoid directed cycles. Given a maximal PDAG $\mathcal{H}$ consistent with a CPDAG $\mathcal{G}^*$, a variable X, and $\mathbf{S} \subset sib(X, \mathcal{H})$, we use $\mathcal{H}_{\mathbf{S}\to X}$ to represent the partially directed graph resulted by orienting $\mathbf{S} \to X$ and $X \to$

$sib(X, \mathcal{H}) \setminus \mathbf{S}$ in $\mathcal{H}$. The next theorem shows the sufficient and necessary conditions for checking whether or not $\mathcal{H}_{\mathbf{S}\to X}$ is consistent with $\mathcal{G}^*$.

Theorem 1 *Let $\mathcal{H}$ be a maximal PDAG consistent with a CPDAG $\mathcal{G}^*$. For any vertex X and $\boldsymbol{S} \subset sib(X, \mathcal{H})$, the following three statements are equivalent.*

(1) There is a DAG $\mathcal{G}$ in the restricted Markov equivalence class represented by $\mathcal{H}$ such that $pa(X, \mathcal{G}) = \boldsymbol{S} \cup pa(X, \mathcal{H})$ and $ch(X, \mathcal{G}) = sib(X, \mathcal{H}) \cup ch(X, \mathcal{H}) \setminus \boldsymbol{S}$.

(2) Compared with $\mathcal{H}$, $\mathcal{H}_{\boldsymbol{S}\to X}$ does not introduce any new V-structure collided on X or any directed triangle containing X.

(3) The induced subgraph of $\mathcal{H}$ over $\boldsymbol{S}$ is complete, and there does not exist an $S \in \boldsymbol{S}$ and a $C \in adj(X, \mathcal{H}) \setminus (\boldsymbol{S} \cup pa(X, \mathcal{H}))$ such that $C \to S$.

The proof of Theorem 1 is provided in Appendix B.1. An important aspect of Theorem 1 is that, it theoretically proves that the only directed cycles we need worry about when orienting undirected edges around a variable X are those triangles containing X, and the only v-structures which might be introduced into the graph are those collided on X. Thus, with Theorem 1, we can locally check whether a set of variables can be the parents of X.

3.2 ESTIMATING CAUSAL EFFECTS

With the help of Theorem 1, we can locally compute all possible causal effects of a treatment on a target. Algorithm 3 shows the framework. Algorithm 3 first constructs the maximal PDAG $\mathcal{H}$ from $\mathcal{G}^*$ and $\mathcal{B}_d$ by using Meek's criteria, then for each treatment variable X, it enumerates all subsets of $sib(X, \mathcal{H})$ and locally checks whether it can be treated as the parental set of X. The correctness of Algorithm 3 is guaranteed by Theorem 1.

Compared with the semi-local IDA, DIDA (Algorithm 3) is a fully local algorithm, which means it only needs the local structure of the treatment when estimating all possible causal effects of the treatment on the target. Furthermore, one can easily see that IDA is an instance of DIDA with no background knowledge, since if $\mathcal{B}_d = \emptyset$, $\mathcal{H}$ is identical to $\mathcal{G}^*$, and orienting undirected edges connected to a given variable in a CPDAG never produces directed cycles.

4 INCORPORATING NON-ANCESTRAL INFORMATION

In practice, we may also have background knowledge about non-ancestral relations among variables. In fact, non-ancestral information is more common than direct causal information, since the later is a special case of the former, with the additional information that two variables are adjacent in the true DAG. However, incorporating non-ancestral information into causal inference is not easy. In this section, we prove that a non-ancestral information set can be equivalently transformed into a direct causal information set. Thus, non-ancestral information, like direct causal information, can be interpreted graphically via maximal PDAGs.

Algorithm 3 DIDA: A fully local method for estimating possible causal effects with direct causal information.

Require: A CPDAG $\mathcal{G}^*$, a consistent direct causal information set $\mathcal{B}_d$, a target variable Y.
Ensure: $\{\Theta_X\}_{X\in V}$, where Θ_X is the multi-set of possible causal effects of X on Y.

```
Construct the maximal PDAG H from G* and B_d using Meek's criteria,
for each variable X ∈ V do
    set Θ_X = ∅,
    for each S ⊂ sib(X, H) such that orienting S → X and X → sib(X, H) \ S does not introduce any V-structure collided on X or any directed triangle containing X do
        estimate the causal effect of X on Y by adjusting for S ∪ pa(X, H), and add the causal effect to Θ_X,
    end for
end for
return {Θ_X}_{X∈V}.
```

4.1 EQUIVALENT BACKGROUND KNOWLEDGE

In this part, we give theoretical foundations as well as an algorithm for transforming non-ancestral information. We begin our discussion with a new concept called *equivalent background knowledge*.

Definition 1 (Equivalent Background Knowledge) *Given a CPDAG $\mathcal{G}^*$, two background knowledge sets $\mathcal{B}_1$ and $\mathcal{B}_2$ are equivalent with respect to $\mathcal{G}^*$, if the restricted Markov equivalence class with respect to $\mathcal{G}^*$ and $\mathcal{B}_1$ is identical to the restricted Markov equivalence class with respect to $\mathcal{G}^*$ and $\mathcal{B}_2$.*

Definition 1 means that two background knowledge sets are equivalent if and only if they put the same constraints on an equivalence class. Note that, the equivalence of background knowledge depends on $\mathcal{G}^*$. Generally, two equivalent background knowledge sets with respect to one CPDAG may not be equivalent anymore with respect to

Algorithm 4 Construct equivalent direct causal information

Require: A CPDAG $\mathcal{G}^*$, a consistent non-ancestral information set $\mathcal{B}_n$.
Ensure: An equivalent direct causal information set $\mathcal{B}_d$.
1: Set $\mathcal{B}_d = \emptyset$,
2: **for** each constraint $X \nrightarrow Y$ in $\mathcal{B}_n$ **do**
3: find the critical set $\mathbf{C}$ of X with respect to Y in $\mathcal{G}^*$, and add $C \to X$ to $\mathcal{B}_d$ for each $C \in \mathbf{C}$,
4: **end for**
5: **return** $\mathcal{B}_d$.

another CPDAG.

In the following, we will prove that a non-ancestral information set is equivalent to a certain direct causal information set with respect to a given CPDAG. Another new concept is needed here.

Definition 2 (Critical Set) *Let $\mathcal{G}^*$ be a CPDAG. X and Y are two distinct vertices in $\mathcal{G}^*$. The critical set of X with respect to Y in $\mathcal{G}^*$ consists of all adjacent vertices of X lying on at least one chordless partially directed path from X to Y.*

Note that Y itself may be in the critical set. Critical sets are important in transforming non-ancestral information to direct causal information, as stated in the following lemma.

Lemma 2 *Let $\mathcal{G}^*$ be a CPDAG. For any two distinct vertices X and Y in $\mathcal{G}^*$, X is not an ancestor of Y in the underlying DAG if and only if every vertex in the critical set of X with respect to Y in $\mathcal{G}^*$ is a direct cause of X in the underlying DAG.*

The proof of Lemma 2 is in Appendix B.2. With Lemma 2, we can construct an equivalent direct causal information set from a given non-ancestral information set. Algorithm 4 shows the procedure. Notice that, the main step of Algorithm 4 is to find the critical set, which can be done by using width-first-search (Perković et al., 2017).

The correctness of Algorithm 4 is guaranteed by Theorem 2, where the proof is given in Appendix B.3. It is worth noting that Algorithm 4 (and Theorem 2) is not only for consistent non-ancestral information, but also for the mixture of consistent non-ancestral and direct causal information, as the later is a special case of the former. Thus, if the input background knowledge of Algorithm 4 is a consistent direct causal information set, then the output is identical to the input minus a collection of information that does not introduce any constraint, e.g., the information $X \nrightarrow Y$ while $Y \to X$ is already present in the CPDAG.

Algorithm 5 NIDA: A fully local method for estimating possible causal effects with non-ancestral information

Require: A CPDAG $\mathcal{G}^*$, a consistent non-ancestral information set $\mathcal{B}_n$, a target variable Y.
Ensure: $\{\Theta_X\}_{X \in V}$, where Θ_X is the multi-set of possible causal effects of X on Y.
1: Construct the equivalent direct causal information $\mathcal{B}_d$ by calling Algorithm 4, with input $\mathcal{G}^*$ and $\mathcal{B}_n$,
2: compute $\{\Theta_X\}_{X \in V}$ by calling DIDA (Algorithm 3), with input $\mathcal{G}^*$, $\mathcal{B}_d$, and Y,
3: **return** $\{\Theta_X\}_{X \in V}$.

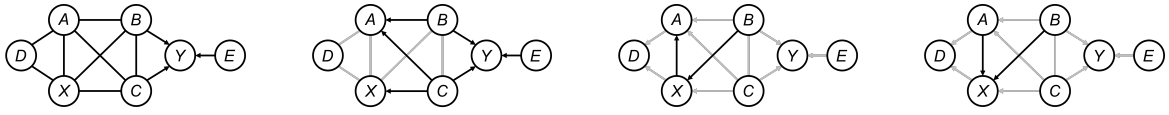
Theorem 2 *Let $\mathcal{G}^*$ be a CPDAG. For any consistent non-ancestral information set $\mathcal{B}_n$, the direct causal information set $\mathcal{B}_d$ constructed according to Algorithm 4 is equivalent to $\mathcal{B}_n$.*

4.2 TRANSFORMING NON-ANCESTRAL INFORMATION AND ESTIMATING CAUSAL EFFECTS

Section 4.1 shows that a consistent non-ancestral information set can be equivalently transformed into a direct causal information set. Therefore, we can graphically interpret non-ancestral information via maximal PDAGs. Once we obtain a maximal PDAG, the possible causal effects of a treatment on a target can be estimated locally based on DIDA (Algorithm 3). The above procedure is summarized in Algorithm 5.

Similar to Algorithm 4, NIDA is also valid when the input is a direct causal information set. From this point of view, DIDA is a special case of NIDA. However, if one is certain that the type of background knowledge is direct causal information, we suggest to use DIDA directly, since calling Algorithm 4 in NIDA may bring unnecessary costs.

Example 2 *We use an example to show how NIDA works. Consider the graphs in Figure 2 as well as the treatment X. Figure 2a shows the CPDAG $\mathcal{G}^*$ learned from data. Suppose we also have the background knowledge which states that A is not an ancestor of Y and X is not an ancestor of C. Notice that, X is not an ancestor of C is also a piece of direct causal information, i.e., C is a direct cause of X, since X and C are adjacent in $\mathcal{G}^*$. The background knowledge is marked on Figure 2b. Figure 2c shows the partially directed graph $\mathcal{H}_1$ resulted by converting the output of Algorithm 4 to a PDAG, that is, for any $X \to Y$ in $\mathcal{B}_d$, if $X - Y$ in $\mathcal{G}^*$, then we orient $X - Y$ as $X \to Y$. Besides $C \to X$, $C - A$ and $B - A$ are oriented as $C \to A$ and $B \to A$ respectively since $A - B \to Y$ and $A - C \to Y$ are chordless par-*



(a) CPDAG $\mathcal{G}^*$ (c) PDAG $\mathcal{H}_1$ (e) $\mathcal{H}_3$ (g) $\mathcal{H}_5$

(b) Background Knowledge (d) Maximal PDAG $\mathcal{H}_2$ (f) $\mathcal{H}_4$ (h) $\mathcal{H}_6$

Figure 2: An example to illustrate how NIDA (Algorithm 5) works.

tially directed paths. Figure 2d further gives the maximal PDAG extending $\mathcal{H}_1$ based on Meek's criteria. Since $sib(X, \mathcal{H}_2) = \{A, B\}$, there are four candidate parental sets of X, namely, $\{A, B\}$, $\{A\}$, $\{B\}$, and $\emptyset$. However, setting $\{A\}$ to be X's parental set will introduce a directed triangle, see Figure 2h. Thus, the only three possible parental sets are illustrated in Figure 2e-2g.

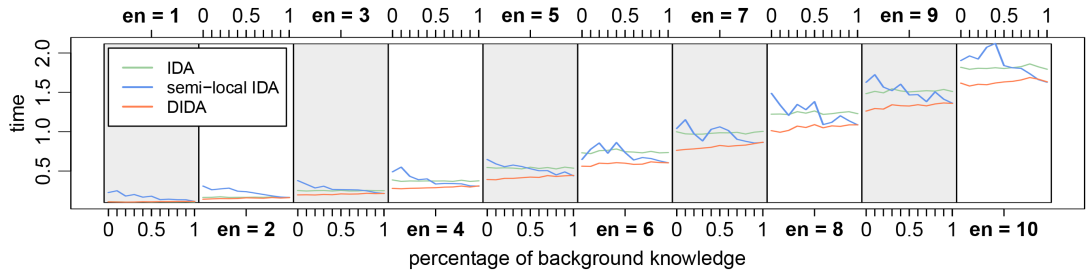
5 EXPERIMENTS

The algorithms proposed in this paper enable us to fully locally estimate possible causal effects with two different types of background knowledge. In this section, with both synthetic and real-world data, we empirically show that the local nature of our algorithms can indeed reduce the computational costs. In Section 5.1, we compare DIDA and NIDA to IDA and the semi-local IDA, with direct causal information and non-ancestral information, respectively. Note that, the semi-local IDA is not directly applicable to non-ancestral information, thus we combined it with Algorithm 4. In Section 5.2, we apply our methods to the *Arabidopsis thaliana* data set. Since DIDA is a special case of NIDA, we only use NIDA in this part.

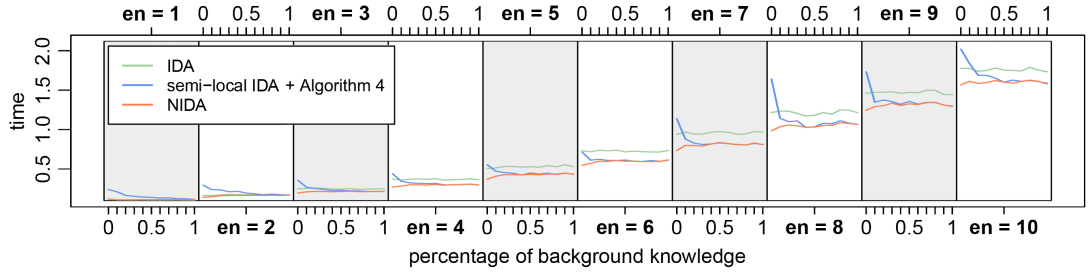
5.1 SIMULATIONS

Our simulations were conducted as follows. In the first scenario, we first sampled a random DAG $\mathcal{G}$ with $N = 100$ vertices and expected neighborhood size $e \in \{1, 2, ..., 10\}$, then randomly picked a treatment X and a target Y, and generated a consistent direct causal information set $\mathcal{B}_d(\mathcal{G})$ by randomly choosing $p \in \{0, 10, ..., 100\}$ percent of directed edges in $\mathcal{G}$ as background knowledge. Notice that in our simulations, a chosen direct causal claim may put no constraint on the Markov equivalence class. This procedure was repeated 100 times, resulting 100 $(\mathcal{G}, \mathcal{B}_d(\mathcal{G}))$ pairs for each setting. (There are totally 10×11 settings.) Next, for each $(\mathcal{G}, \mathcal{B}_d(\mathcal{G}))$ pair, we randomly generated a multivariate Gaussian distribution with edge weights uniformly sampled from $[0.5, 2]$ independently and independent standard normal noises (Maathuis et al., 2009), and sampled 1000 observations from this distribution. Finally, we transformed each sampled DAG to the corresponding CPDAG, added background knowledge to the CPDAG, and estimated possible causal effects of the chosen treatment on the chosen target. In the second scenario, we consider non-ancestral background knowledge. The non-ancestral background information set with respect to a given DAG $\mathcal{G}$ was generated by randomly choosing $p \in \{0, 10, ..., 100\}$ percent of non-ancestral relations according to $\mathcal{G}$, i.e., variable pairs like (X, Y) where Y is not an ancestor of X. Except for sampling background knowledge, other procedures were similar to those in the first scenario. We note that, following Perković et al. (2017), the input CPDAG for each setting in both scenarios is the true CPDAG rather than the estimated one, since we do not want to bring any estimation bias caused by learning graphs to the evaluation of different methods. Besides, it is difficult to incorporate background knowledge to a incorrect CPDAG since they may conflict to each other.

Figure 3 shows the average CPU time of IDA, the semi-local IDA and DIDA (NIDA), with direct causal information (Figure 3a) and non-ancestral information (Figure 3b). As expected, it takes more time to estimate the multiset of possible effects when the graph is dense. Since IDA is directly applied to the CPDAGs without considering background knowledge, the average CPU time of IDA is stable when the percentage of background knowledge varies. Similar to IDA, the average CPU time of DIDA (NIDA) is also stable, as DIDA (NIDA) is fully local and adding background knowledge does not change the neighborhood size. Although the figure suggests that DIDA (NIDA) is slightly faster than IDA, we find this difference is insignificant, as shown in Figure 4. On the other hand, the average CPU time of the semi-local IDA decreases when the percentage of background knowledge increases. When no background knowledge is given, the semi-local



(a) With Direct Causal Information



(b) With Non-Ancestral Information

Figure 3: The average CPU time (secs.) of IDA, the semi-local IDA and DIDA (NIDA), with direct causal information and non-ancestral information. IDA is directly applied to the CPDAGs without adding any background knowledge. **en** is an abbreviation for 'expected neighborhood size'.

IDA is usually slower than both IDA and DIDA (NIDA), but as the percentage of background knowledge increases, the number of undirected edges in the maximal PDAG decreases, which makes the CPU time of the semi-local IDA converge to that of DIDA (NIDA).

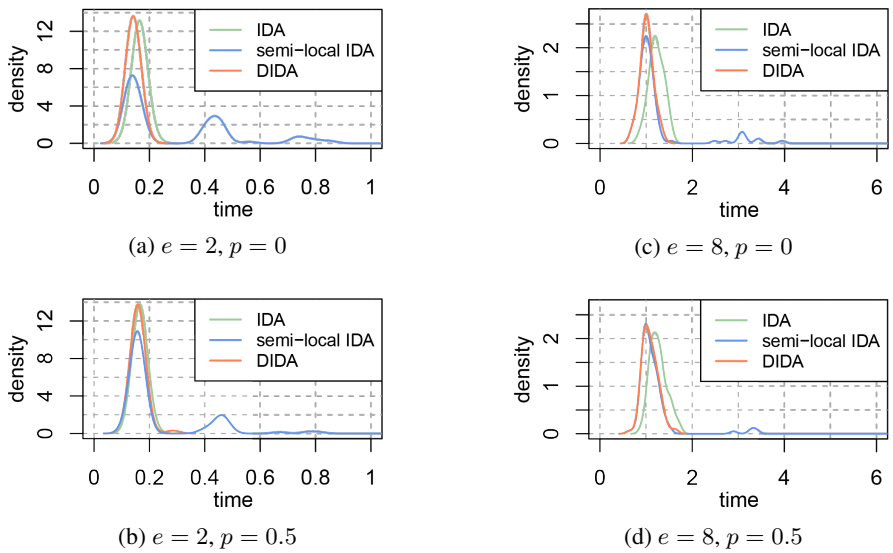
Another important feature indicated by Figure 3 is that, non-ancestral information is more informative than direct causal information. Fix an expected neighborhood size, one can see that the average time of the semi-local IDA given non-ancestral information decreases faster than that given direct causal information. This means that with the same percentage of background knowledge, there are less undirected edges in the maximal PDAG resulted from adding non-ancestral information. Figure 5 in Appendix A also supports this claim, where we report the average number of possible effects of one treatment on one target. This interesting feature is supported by Lemma 2 and Theorem 2. A direct causal claim can at most orient one edge, while a non-ancestral claim can potentially orient more than one undirected edge.

We also analyze the distribution of the CPU time. As an example, Figure 4 shows the estimated densities with the expected neighborhood size $e \in \{2, 8\}$ and the percentage of direct causal information $p \in \{0, 0.5\}$. From the figures we know that the CPU time distributions of IDA and DIDA are unimodal, while that of the semi-local IDA is usually multimodal. Another important result is that, the CPU time distributions of all three methods have one common peak near zero. When the graph becomes dense or more background knowledge is given, the other peaks of the semi-local IDA become flat. Finally, the CPU time distribution of the semi-local IDA becomes unimodal.

5.2 REAL-WORLD DATA

We now apply NIDA to the *Arabidopsis thaliana* data set (Opgen-Rhein & Strimmer, 2007). The *Arabidopsis thaliana* data set can be directly loaded from R package `GeneNet` (Schäfer et al., 2006). The data set consists of 11 samples of 800 genes, and each variable approximately follows a Gaussian distribution. We used a hybrid method to learn a CPDAG from the *Arabidopsis thaliana* data set (see Appendix A for more details). The final CPDAG contains 32 undirected edges and 266 directed edges, and there are 185 genes in the network after removing all singletons. The background knowledge was obtained from the *ARTH150* network.[3] The *ARTH150* network is a DAG with 107 nodes and 150 directed edges, which describes the causal relations among a subset of 800 genes in the *Arabidopsis thaliana* data set. We constructed $\mathcal{B}_n$ by adding all $Y \nrightarrow X$ such that X is an ancestor of Y in the

[3]The network can be found at `http://www.bnlearn.com/bnrepository/`.



(a) $e = 2, p = 0$ (c) $e = 8, p = 0$

(b) $e = 2, p = 0.5$ (d) $e = 8, p = 0.5$

Figure 4: The estimated densities of CPU time (secs.) of different approaches, with the expected neighborhood size $e \in \{2, 8\}$ and the percentage of direct causal information $p \in \{0, 0.5\}$.

ARTH150 network. Clearly, $\mathcal{B}_n$ is a non-ancestral information set. The total number of non-ancestral relations in $\mathcal{B}_n$ is 525. After adding the background knowledge to the CPDAG, the maximal PDAG contains 16 undirected edges and 282 directed edges.

All methods were applied to all 185×184 pairs of distinct variables (X, Y) in the learned maximal PDAG to estimate the possible effects of each X on each Y. Similar to the simulation results, the CPU time distributions of IDA and NIDA are unimodal while the CPU time distribution of the semi-local IDA is multimodal. In fact, the maximal time of the semi-local IDA is 65.48 seconds, while the maximal time of IDA and NIDA is 3.68 and 3.02 seconds respectively. Since the maximal PDAG only contains 16 undirected edges, the semi-local IDA and NIDA perform similarly on average. However, NIDA is more stable across all situations, no matter how large the set of possible causal effects is.

6 CONCLUDING REMARKS

Estimating causal effects from observational data has been widely studied. However, in practice, one may also have prior knowledge about the causal system. This additional information may have great influence on causal inference. In this paper, we consider the problem of estimating all possible causal effects from observational data with direct causal information and non-ancestral information. We provide locally valid orientation rules for maximal PDAGs, which extend Maathuis et al. (2009, Lemma 3.1). Based on the rules, we propose a fully local algorithm to estimate all possible causal effects of a treatment on a target. We further consider non-ancestral information and prove that a non-ancestral information set can be equivalently transformed into a direct causal information set, making it possible to estimate possible causal effects with non-ancestral information locally. Experiments show that our algorithms are efficient and stable.

There are some interesting future directions. First, how to represent incoherent background knowledge with maximal PDAGs is an important problem in real applications. To solve the problem, we may need additional information such as the confidence level of each claim, and perhaps use the Answer Set Programming (ASP) to find a maximal PDAG that minimizes the confidence level of the input claims which the maximal PDAG does not satisfy (Zhalama et al., 2019). Moreover, it is worth considering the causal system containing hidden variables and selection biases (Richardson & Spirtes, 2002; Zhang, 2008). However, as discussed in Perković et al. (2017), interpreting background knowledge graphically in this case is still challenging. Another possible extension is to consider other forms of background knowledge, such as ancestral relations or structural priors.

Acknowledgements

We appreciate the anonymous reviewers for valuable comments and suggestions. We would also like to thank Prof. Zhi Geng from Peking University and Dr. Yue Liu from Huawei Noah's Ark Lab for helpful discussions. This research was supported by National Key R&D Program of China (2018YFB1004300) and NSFC (11671020).

References

Andersson, S. A., Madigan, D., and Perlman, M. D. A characterization of Markov equivalence classes for acyclic digraphs. *The Annals of Statistics*, 25(2):505–541, 04 1997.

Blair, J. R. S. and Peyton, B. An introduction to chordal graphs and clique trees. In *Graph Theory and Sparse Matrix Computation*, pp. 1–29, New York, NY, 1993. Springer New York.

Hauser, A. and Bühlmann, P. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13(Aug):2409–2464, 2012.

He, Y. and Geng, Z. Active learning of causal networks with intervention experiments and optimal designs. *Journal of Machine Learning Research*, 9(Nov):2523–2547, 2008.

He, Y., Jia, J., and Yu, B. Counting and exploring sizes of Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 16:2589–2609, 2015.

Henckel, L., Perković, E., and Maathuis, M. H. Graphical criteria for efficient total effect estimation via adjustment in causal linear models. *arXiv e-prints*, art. arXiv:1907.02435, Jul 2019.

Jaber, A., Zhang, J., and Bareinboim, E. A graphical criterion for effect identification in equivalence classes of causal diagrams. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pp. 5024–5030. International Joint Conferences on Artificial Intelligence Organization, 7 2018a.

Jaber, A., Zhang, J., and Bareinboim, E. Causal identification under Markov equivalence. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*. AUAI press, 2018b.

Jaber, A., Zhang, J., and Bareinboim, E. Causal identification under Markov equivalence: Completeness results. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2981–2989. PMLR, 09–15 Jun 2019.

Maathuis, M. H., Kalisch, M., and Bühlmann, P. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A): 3133–3164, 12 2009.

Meek, C. Causal inference and causal explanation with background knowledge. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pp. 403–410. Morgan Kaufmann Publishers Inc., 1995.

Opgen-Rhein, R. and Strimmer, K. From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC Systems Biology*, 1(37): 1–10, Aug 2007.

Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

Pearl, J. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 12 1995.

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

Pearl, J., Geiger, D., and Verma, T. Conditional independence and its representations. *Kybernetika*, 25(7): 33–44, 1989.

Perković, E. Identifying causal effects in maximally oriented partially directed acyclic graphs. *arXiv e-prints*, art. arXiv:1910.02997, Oct 2019.

Perković, E., Textor, J., Kalisch, M., and Maathuis, M. H. A complete generalized adjustment criterion. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pp. 682–691. AUAI Press, 2015.

Perković, E., Kalisch, M., and Maathuis, M. H. Interpreting and using CPDAGs with background knowledge. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence*. AUAI press, 2017.

Perković, E., Textor, J., Kalisch, M., and Maathuis, M. H. Complete graphical characterization and construction of adjustment sets in Markov equivalence classes of ancestral graphs. *Journal of Machine Learning Research*, 18(220):1–62, 2018.

Richardson, T. and Spirtes, P. Ancestral graph Markov models. *The Annals of Statistics*, 30(4):962–1030, 08 2002.

Schäfer, J., Opgen-Rhein, R., and Strimmer, K. Reverse engineering genetic networks using the genenet package. *The Newsletter of the R Project*, 6(9):50–53, 2006.

Spirtes, P., Glymour, C. N., and Scheines, R. *Causation, Prediction, and Search*. MIT Press, second edition, 2000.

Zhalama, Z., Zhang, J., Eberhardt, F., Mayer, W., and Li, J. ASP-based discovery of semi-Markovian causal models under weaker assumptions. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pp. 1488–1494. International Joint Conferences on Artificial Intelligence Organization, Jul 2019.

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