Causal Structure Discovery from Incomplete Data

by

Chandler Squires

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2019

Author .....
Department of Electrical Engineering and Computer Science
August 23, 2019

Certified by .....
Caroline Uhler
Associate Professor
Thesis Supervisor

Accepted by .....
Katrina LaCurts
Chair, Master of Engineering Thesis Committee

Causal Structure Discovery from Incomplete Data

by

Chandler Squires

Submitted to the Department of Electrical Engineering and Computer Science
on August 23, 2019, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

Causal structure learning is a fundamental tool for building a scientific understanding of the way a system works. However, in many application areas, such as genomics, the information necessary for current causal structure learning algorithms does not match the information that researchers can actually access, for example when the algorithm requires knowledge of intervention targets but the interventions have off-target effects. In this thesis, we developed, implemented, and tested a novel algorithm for discovering a causal DAG from observational and interventional data, when the intervention targets are either partially or completely unknown. We relate the algorithm to the recently introduced Joint Causal Inference framework. Finally, we evaluate the performance of the algorithm on synthetic datasets and demonstrated its ability to outperform current state-of-the-art causal structure learning algorithms which assume known intervention targets.

Thesis Supervisor: Caroline Uhler
Title: Associate Professor

Acknowledgments

I could not hope to acknowledge all of the people who have made this thesis a possibility, in ways big and small, direct and indirect. In this section, I'll do my best to convey the massive appreciation I feel for at least a few of those people.

Foremost, I would like to thank my advisor Caroline Uhler for exposing me to many fascinating aspects of machine learning, statistics, and mathematics; for perfectly balancing patience and motivation; and for her guidance through the ups and downs of research life.

I would also like to thank Yuhao Wang, who has mentored me over the course of two projects and whose ideas were essential to this project - the original conception of the algorithm and the proof of its consistency are directly thanks to him.

Furthermore, I'd like to thank the many other members of Uhler lab with whom I've had the pleasure of collaborating: Raj Agrawal, Anastasiya Belyaeva, Daniel Bernstein, Basil Saeed, and Karren Yang. Through lonng discussions and late nights with them, I've come to understand causality at a deeper level and managed to enjoy myself at the same time.

Last but not least, I'd like to thank my mother for instilling in me the curiosity, confidence, and integrity that I draw on every day; my grandparents for their listening ears and loving concern, and the friends and loved ones who've made MIT a home for the past five years.

Contents

List of Figures

Chapter 1

Introduction

Causal structure learning is the statistical task of recovering the (unknown) true causal model of a system given some combination of observational and/or experimental data and background knowledge. In some applications, interventional data may be available, but the variables targeted by each intervention may be partially or completely unknown. We propose an algorithm that is capable of learning causal structure from such data.

Chapter two describes the Unknown-Target Interventional Greedy Sparsest Permutation (UT-IGSP) algorithm and its consistency guarantees.

Chapter three empirically demonstrates that UT-IGSP outperforms a variety of other interventional structure-learning methods on synthetic data at both the structure learning and intervention recovery tasks, across a variety of settings.

Chapter four concludes with a discussion of our algorithm and proposes future directions for causal structure learning under different types of uncertainty.

1.1 Motivations for unknown-target causal structure learning

Learning the causal structure of a complex system is an important task for many applications, including genomics, healthcare, and econometrics. For example, the

causal structure may be a tool that allows a biologist to predict the effect of a gene editing experiment on gene expression levels.

However, it is well-known throughout the field of causal inference that causal structure is (in general) not uniquely identifiable from observational data alone. Hence, it is necessary to use interventional data to improve the identifiability of the model.

In genomics, the necessary interventional data can be obtained in large quantities via high-throughput gene-sequencing methods such as Perturb-Seq. However, gene-editing methods such as CRISPR are known to have off-target effects [13], which violates the assumptions of most current interventional structure-learning methods.

1.2 Causal graphs

In the section, we review background on causal DAGs and their observational and interventional Markov equivalence classes. In Appendix A we review the basic graph theory and notation that we use throughout the paper.

1.2.1 Causal DAG models

The basic graph that is used to model causal relationships is a *directed acyclic graph* (DAG), where an edge from one variable to another represents that the former is a direct cause of the latter. Formally, let $D = ([p], E)$ be a DAG with nodes $[p] := \{1, \cdots, p\}$ and edges E. D represents a causal model where each node i is associated with a random variable X_i. Let f denote the density of the data-generating distribution $\mathbb{P}$ over the random vector $X := (X_1, \cdots, X_p)$. We say that $\mathbb{P}$ is *Markov* with respect to the DAG if f factorizes with respect to D, i.e.,

$$f(x) = \prod_{i \in [p]} f_i(x_i | x_{\mathrm{pa}_D(i)}),$$

One foundational result for DAG models [7, Section 3.2.2] is that $\mathbb{P}$ is Markov with respect the DAG D if and only if the set of conditional independence relations

in $\mathbb{P}$ is a subset of the the set of d-separation statements[1] in D. In other words, for any disjoint $A, B, C \subset [p]$, if A and B are d-separated given C in D, then X_A is conditionally independent from X_B given X_C in $\mathbb{P}$. A DAG D is called an IMAP of $\mathbb{P}$ (altenatively, the conditional indepdence statements of $\mathbb{P}$) if $\mathbb{P}$ is Markov to D. The *faithfulness assumption*, which is commonly assumed in existing causal inference algorithms [9], asserts that the converse is also true, i.e., a conditional independence statement is true in $\mathbb{P}$ if and only if the corresponding d-separation statement holds in D.

1.2.2 Markov equivalence

Let $\mathcal{M}(D)$ denote the set of distributions that are Markov with respect to D. Two DAGs D_1 and D_2 are *Markov equivalent*, denoted $D_1 \sim D_2$ if $\mathcal{M}(D_1) = \mathcal{M}(D_2)$. [12] show that $D_1 \sim D_2$ if and only if D_1 and D_2 have the same skeleton and v-structures. Moreover, if D_1 and D_2 are Markov equivalent, then D_1 and D_2 can be transformed to one another by a sequence of *covered edge reversals*, where we call an edge $i \to j$ in a DAG D as a covered edge if $\mathrm{pa}_D(j) = \mathrm{pa}_D(i) \cup \{i\}$. We use $[D]$ to denote the set of DAGs that are Markov equivalent to D, i.e., the Markov equivalence class of D. Since every DAG in $[D]$ explains the same set of conditional independence statements, the true DAG underlying some distribution is not identifiable, instead, we can only identify D up to its Markov equivalence class.

The Markov equivalence class of a DAG D can be concisely represented by the *essential graph* $\mathcal{E}(D)$, which has both undirected and directed edges (i.e., it is a *mixed graph*). $\mathcal{E}(D)$ has the same skeleton as D, with directed edges $i \to j$ for all edges $i \to j \in D$ such that $i \to j \in D'$ for all $D' \in [D]$, and all other edges undirected.

1.3 Interventions

To improve the identifiability of the underlying causal model, we can intervene on the variables. A theoretical framework for modeling interventions was developed in [4].

[1] d-separation is reviewed in the appendix

1.3.1 Types of Interventions

A *perfect intervention* assumes that all causal dependencies between intervened targets and their causes are removed [4]. For example, consider a perfectly performed gene knockout experiment, where the expression of a gene is set to zero and hence all interactions between gene i and its upstream regulators are eliminated.

In practice, interventions often cannot fully remove the causal dependencies between an intervened target and its causes, but rather *modify* their causal relationship [4]. For example, in genomics, an intervention may only inhibit the expression of a gene [2]. Such interventions are known as *imperfect* of *soft*.

1.3.2 Interventional Markov property

Let $I \subseteq [p]$ denote a perfect or imperfect intervention target and let f^{obs} and f^I denote the densities of the observational and interventional distributions, respectively. A pair $(f^{\emptyset}, f^I)$ is *I-Markov* with respect to a DAG D if $f^{\emptyset}$ and f^I are Markov with respect to D and for any non-intervened variable $j \in [p] \setminus I$, it holds that

$$f^I(x_j \mid x_{\text{pa}_D(j)}) = f^{\text{obs}}(x_j \mid x_{\text{pa}_D(j)}), \tag{1.1}$$

i.e., the conditional distributions of the non-intervened variables are invariant across the observational and interventional distributions. We call such statements across intervention settings *conditional invariance* statements. The interventional Markov property implies that the interventional distribution f^I factors as:

$$f^I(x) = \prod_{i \notin I} f^{\text{obs}}(x_i | x_{\text{pa}_D(i)}) \prod_{i \in I} f^I(x_i | x_{\text{pa}_D(i)}). \tag{1.2}$$

Let $\mathcal{M}_I(D)$ denote the set of distributions that are I-Markov with respect to D. As in the observational setting, two DAGs D_1 and D_2 are in the same *I-Markov equivalence class*, if $\mathcal{M}_I(D_1) = \mathcal{M}_I(D_2)$ [6, 16].

Simarly, given a set of interventions $\mathcal{I} = \{I^1, \ldots, I^K\}$ and a set of densities $f_{\mathcal{I}} = (f^{\text{obs}}, f^{I^1}, \ldots, f^{I^K})$, $f_{\mathcal{I}}$ is $\mathcal{I}$-Markov with respect to a DAG D if each pair $(f^{\text{obs}}, f^{I^k}$

is I^k-Markov with respect to D. Again, $M_{\mathcal{I}}(D)$ represents the set of distributions that are $\mathcal{I}$-Markov with respect to D and two DAGs D_1 and D_2 are in the same $\mathcal{I}$-Markov equivalence class if $\mathcal{M}_{\mathcal{I}}(D_1) = \mathcal{M}_{\mathcal{I}}(D_2)$. The set of DAGs that are $\mathcal{I}$-Markov equivalent to D is denoted $[D]_{\mathcal{I}}$, and can be concisely represented by the $\mathcal{I}$-essential graph, denoted $\mathcal{E}_{\mathcal{I}}(D)$, which again has the same skeleton as D and each edge is directed only if it is directed the same way in all $D' \in [D]_{\mathcal{I}}$.

1.4 Observational causal structure learning algorithms

In the next three sections, we review prior work on causal structure learning.

A variety of algorithms have been developed to learn the Markov equivalence class of a DAG from only observational data. We may divide these algorithms into three basic categories: *constraint-based* methods, *score-based* methods, and *hybrid* methods.

The family of constraint-based methods includes the prominent Peters-Clark (PC) algorithm [11], which first learns the skeleton of the causal graph by removing edges from a complete graph through a series of conditional independence tests, then directs edges associated with v-structures implied by these conditional independence tests.

The family of score-based methods includes the prominent Greedy Equivalence Search (GES) algorithm [1], which assigns a score to each Markov equivalence class (such as the BIC score in the linear Gaussian case) and performs a greedy search over the space of Markov equivalence classes in two phases.

The family of hybrid methods includes the Greedy Sparsest Permutation (GSP) algorithm, upon which this work is based. GSP searches over the smaller space of permutations of the nodes, and assigns to each permutation a score equal to the number of edges in the sparsest DAG that is both consistent with the permutation and is an IMAP of the given conditional independence statements.

1.5 Known-target interventional causal structure learning algorithms

A smaller, though still sizeable number, of algorithms have been developed to learn the interventional Markov equivalence class of a DAG from a combination of observational and interventional data when the intervention targets are known.

An extension of GES, called Greedy Interventional Equivalence Search (GIES) [6], also performs a greedy search over the space of Markov equivalence classes. However, GIES was shown to be inconsistent [14], i.e., there are cases in which the number of observationanl and interventional samples both go to infinity, but the algorithm can get stuck in a local optimum and never discover the true graph.

An extension of GSP, called IGSP [14, 16], adapts the score function of GSP to penalize edge directions which contradict conditional invariance statements inferred from the data, and is proven to consistently learn the true interventional Markov equivalence class.

1.6 Unknown-target interventional causal structure learning algorithms

Relatively little work has addressed the problem of causal structure learning with unknown-target interventional data in a scalable manner. [3] introduced a dynamic programming algorithm for finding exactly the Bayesian posterior over DAG edges and intervention targets. Their algorithm builds on previous work for Bayesian causal structure learning by simply adding nodes to the graph for each interventional setting. However, their framework requires parametric assumptions on the conditional probability of each node, and scales superexponentially with the number of nodes, so it cannot be used for much more than $p = 20$ nodes.

Recently, [8] generalized the idea of adding nodes to the graph to specify the different contexts under which the data was collected. This includes as a special case

a collection of observational and interventional data. Their framework provides a means to adapt any observational structure-learning algorithm to a structure-learning algorithm from multiple contexts. Our algorithm can be understood within the JCI framework, and in 2.5 we will explain the subtle difference between our algorithm and the one which would result from directly applying JCI to GSP.

Chapter 2

Unknown-Target IGSP

In this chapter we introduce the Unknown-Target Interventional Greedy Sparsest Permutation (UT-IGSP) algorithm and establish its consistency guarantees under mild assumptions.

2.1 Mapping Permutations to Graphs and Intervention Targets

Suppose we are given a set $\mathcal{C}$ of conditional independence statements of the form $(i \perp\!\!\!\perp j \mid C)$ coming from a distribution $\mathbb{P}$ that is faithful to a DAG D. Suppose we are also given sets $\{\mathcal{V}^k\}_{k\in[K]}$ of conditional variance statements of the form $(i \mid C)$, where $(i \mid C) \in \mathcal{V}^k$ denotes that the conditional distribution of i given C is invariant between the observational distribution and the k^{th} interventional distribution, i.e., $f^{\text{obs}}(i \mid C) = f^{I_k}(i \mid C)$.

Given a permutation π, [12] established that the graph $D_\pi^{\mathcal{C}} = (V, E_\pi^{\mathcal{C}})$, with

$$E_\pi^{\mathcal{C}} = \{i \to j \mid i <_\pi j, (i \perp\!\!\!\perp j \mid \text{an}_\pi(\{j\}) \setminus \{i\}) \notin \mathcal{C}\}$$

is a *minimal IMAP* of D, i.e., that $D_\pi^{\mathcal{C}}$ is an IMAP of D and that no subgraph of $D_\pi^{\mathcal{C}}$ is an IMAP of D. We will call $D_\pi^{\mathcal{C}}$ the π-consistent minimal IMAP. Intuitively, $D_\pi^{\mathcal{C}}$ is the simplest way to explain a set of conditional independence statements given

that the variables must be ordered according to π.

Next, given any candidate graph D', the targets of the k^{th} set of intervention targets must include any variable for which the conditional distribution given its parents has changed, i.e.,

$$I^k(D') = \{i \mid (i \mid \text{pa}_{D'}(i)) \notin \mathcal{V}^k\}$$

Thus, to each permutation π we can also define the following intervention sets, which are based on the π-consistent minimal IMAP:

$$I^k_\pi = I^k(D^{\mathcal{C}}_\pi)$$

Finally, we associate the following cost to each permutation:

$$\text{cost}(\pi) = |E^{\mathcal{C}}_\pi| + \sum_{k \in [K]} |I^k_\pi|$$

2.2 Identifiability of Intervention Targets

Before introducing the algorithm, we will state an assumption that allows us to ensure that we can recover the intervention targets from the interventional data, which will also be a useful assumption for guaranteeing the consistency of our algorithm.

Intuitively, we would like to ensure that any intervened variable changes *all* of its conditional distributions after the intervention (including its marginal distribution). The formal assumption is:

Assumption 1. *[Direct Intervention Faithfulness] We assume that* $f^{I_k}(x_i \mid x_C) \neq f^{obs}(x_i \mid x_c)$*, for all* $k \in [K]$*,* $i \in I_k$*, and* $C \in [p] \setminus i$*.*

This is for instance guaranteed when our interventions are "do-interventions", i.e., they set the value of the variable deterministically, and in the observational distribution, no variable is a determistic function of the others.

Algorithm 1 Learning unknown intervention targets

Input: Distribution f^{obs}, f^{I^k}.
Output: A complete set of estimated intervention targets $\hat{I}^k$.
Set $\hat{I}^k = \emptyset$;
For each node j, if $f^{I^k}(j|S) \neq f^{\emptyset}(j|S)$ for all $S \subseteq [p] \setminus \{j\}$, add j to $\hat{I}^k$.
Return $\hat{I}^k$.

Under Assumption 1, we can identify the targets of an intervention by an exhaustive search over nodes and conditioning sets. Formally,

Theorem 1. *Under Assumption 1, Algorithm 1 returns $\hat{I}^k = I^k$.*

Proof. By Assumption 1, if $i \in I^k$, then $i \in \hat{I}^k$. By the definition of the interventional distribution, if $i \notin I^k$, then $f^{I^k}(i \mid \text{pa}_D(i)) = f^{\text{obs}}(i \mid \text{pa}_D(i))$, so $i \notin \hat{I}^k$. □

We conclude this section by providing an example that demonstrates the need for Assumption 1.

Example 1. *Let $D_a = \{[2], \{1 \to 2\}\}$ and $I_a = \{2\}$, with $f_a^{obs}(x_1) = \mathcal{N}(x_1; 0, 1)$, $f_a^{obs}(x_2 \mid x_1) = \mathcal{N}(x_2; x_1, 1)$, and $f_a^I(x_2 \mid x_1) = \mathcal{N}(x_2; .5x_1, 1.75)$. f_a^I violates Assumption **??** since the marginal distribution of x_2 remains the same across settings, i.e., $f_a^I(x_2) = f_a^{obs}(x_2) = \mathcal{N}(x_2; 0, 2)$. In this case, the DAG $D_b = \{[p], 1 \leftarrow 2\}$ and intervention set I_b, with $f_b^{obs}(x_2) = \mathcal{N}(x_2; 0, 2)$, $f_b^{obs}(x_1 \mid x_2) = \mathcal{N}(x_1; .5x_2, .5)$, and $f_b^{I_b}(x_1 \mid x_2) = \mathcal{N}(.25x_2, .875)$, has the same observational and interventional distributions, and thus D_a cannot be distinguished from D_b.*

2.3 Full Interventional Faithfulness

In this section, we strengthen the notion of direct intervention faithfulness. First, we must introduce the notion of the $\mathcal{I}$-DAG, which was introduced in [16].

Definition 1. *Given a DAG D and intervention sets $\mathcal{I} = \{I^k\}_{k \in [K]}$, the $\mathcal{I}$-DAG, denoted $D_{\mathcal{I}}$ is formed by adding an intervention node ζ^k for $k \in K$, with $\zeta_k \to i$ for $i \in I^k$.*

Assumption 2 (Full intervention faithfulness)**.** *Let $\mathcal{I} = \{I_1, \ldots, I_K\}$. A set of distributions $f^{obs}, f^{I^1}, \ldots, f^{I^K}$ is* fully intervention faithful *to the $\mathcal{I}$-DAG $D_{\mathcal{I}}$ if f^{obs} is faithful to D and $f^{I^k}(x_A \mid x_B) = f^{obs}(x_A \mid x_B)$ if and only if A is d-separated from ζ_k by $B \cup \zeta_{[K]\setminus\{k\}}$.*

Clearly, full intervention faithfulness subsumes the notion of direct intervention faithfulness since the latter is achieved when the condition holds only for sets A of size one.

2.4 Joint Causal Inference

In this section, we review the Joint Causal Inference framework and compare the $\mathcal{I}$-DAG to the DAG derived in the JCI framework.

2.4.1 Basic framework and assumptions

Given data from a DAG in K different contexts, in which some conditional probabilities might change between the contexts, the Joint Causal Inference adds K *context variables* $C_1, \ldots, C_K$ to the structure learning problem. These contexts could include experimental conditions (e.g. time of day, lab temperature), population differences (e.g. cell type, demographic), or in our case, external interventions, where $C_k = 1$ only in the k^{th} interventional setting. The true DAG with the augmenting causal variables is denoted D_C, which we call the *JCI-DAG*. We will distinguish the original variables of D by referring them as "system variables".

The learning problem of JCI can be simplified by making a number of assumptions, including:

Assumption (JCI Assumption 1, Exogeneity)**.** *No system variables cause any context variables, i.e. $x_i \to C_K \notin D_C$ for any $i \in [p]$, $k \in [K]$.*

Assumption (JCI Assumption 2, Randomized contexts)**.** *$X \cup C$ remains causally sufficient, i.e., there are no latent confounders between the system and context variables.*

Assumption (JCI Assumption 3, Generic contexts)**.** *The induced graph over the context varialbes, $D_C[C]$, is complete and known.*

For the sake of learning intervention targets, we use all of the above assumptions. The first two assumptions reflect reasonable experimental setups, i.e., since the interventions are picked before the data is generated, the interventions are not caused by the observed values (exogeneity), and that we can assign the interventions at random. The third assumption also holds: since only one interventional setting is "active" at a time, if $C_k = 1$ then $C_\ell = 0$ for $\ell \neq k$, a strong form of dependence that one can check results in a complete graph, and the order of the intervention variables does not matter.

2.4.2 Comparison to $\mathcal{I}$-DAGs

Note that the only difference between the JCI-DAG and the $\mathcal{I}$-DAG for a set of interventions is the complete graph over the intervention variables. Given any observational causal structure learning algorithm, JCI inherits the same guarantees as the original algorithm once the assumptions are extended to the JCI-DAG.

To define the assumptions under which JCI-GSP is consistent, we need to collect the observational and interventional distributions into a single distribution, which we call the *JCI-distribution*. Formally,

$$f_{\text{JCI}}(X, C) = f_{\text{JCI}}(C) f^{\text{obs}}(X)^{\mathbb{1}_{C=0}} \prod_{k \in [K]} f^{I^k}(X)^{\mathbb{1}_{C_k=1}}$$

where $f_{\text{JCI}}(C)$ dictates the experimental design and satisfies $f_{\text{JCI}}(C = 0) > 0$ and $f_{\text{JCI}}(C) = 0$ for $|C| > 2$, i.e., there is positive probability of observational data and zero probability of more than one intervention variable being active.

Since GSP is consistent under the faithfulness assumption, JCI-GSP is consistent under the faithfulness assumption on f_{JCI}. We briefly show here that if f_{JCI} is faithful to D_C, then $f^{\text{obs}}, f^{I^1}, \ldots, f^{I^k}$ are fully intervention faithful to $D_{\mathcal{I}}$.

Recall that by the definition of faithfulness, f_{JCI} is faithful to D_C when $f^{\text{JCI}}(y_\alpha \mid$

$y_\beta, y_\gamma) = f^{\text{JCI}}(y_\alpha \mid y_\gamma)$ if and only α is d-separated from β given γ, where now y can be picked as either system or context variables and α, β, γ can be indices for system or context variables as well. Picking $y_\alpha = x_A$, $y_\beta = C_k$, and $y_\gamma = C_{[K]\setminus\{k\}} \cup x_B$ from Assumption 1, we see that the faithfulness assumption on D_C implies that $f_{\text{JCI}}(x_A \mid C_k, C_{[K]\setminus\{k\}}, x_B) = f_{\text{JCI}}(x_A \mid C_{[K]\setminus\{k\}}, x_B)$ if and only if A is separated from C_k by $B \cup C_{[K]\setminus\{k\}}$. By the definition of conditional independence, we do not need to check equality when conditioning on events of probability zero, so we only need to consider the cases where $C_{[K]\setminus\{k\}}$ has zero or one entries nonzero.

If $C_\ell = 1$ for $\ell \neq k$, then C_k is deterministically zero and the equality trivially holds. Thus, the only non-trivial case is when $C_{[K]\setminus\{k\}} = 0$. If $C_k = 1$, we have

$$f_{\text{JCI}}(x_A \mid C_k = 1, C_{[K]\setminus\{k\}} = 0, x_B) = f^{I^k}(x_A \mid x_B)$$

If $C_k = 0$, we have

$$f_{\text{JCI}}(x_A \mid C_{[K]} = 0, x_B) = f^{\text{obs}}(x_A \mid x_B)$$

Since both must equal the same right hand side, we recover the condition that $f^{I^k}(x_A \mid x_B) = f^{\text{obs}}(x_A \mid x_B)$. Finally, since we are conditioning on all of the context variables, which are not colliders, the d-separation statement in D_C is equivalent to the d-separation statement in $D_{\mathcal{I}}$.

Thus, we have established that the assumptions needed for consistency of JCI-GSP are at least as strong as the full intervention faithfuless assumption. We leave to future work the question of whether these faithfulness assumptions are equivalent.

2.5 Unknown Target Interventional Greedy Sparsest Permutation Algorithm

In this section, we introduce UT-IGSP. First, we define a more restricted version of covered edges based on any intervention targets which may already be known. This

Algorithm 2 Unknown-target IGSP

Input: Interventional distributions $f^{\emptyset}, f^1, \cdots, f^K$ and their corresponding partially known intervention sets $\mathcal{I}^{\text{Kn}} := \{I_1^{\text{Kn}}, I_2^{\text{Kn}} \cdots, I_K^{\text{Kn}}\}$, a starting permutation π_0.

Output: A permutation π and associated minimal I-MAP D_π, a complete set of estimated intervention targets $\mathcal{I} := \{I^1, \cdots, I^K\}$.

Set $\pi := \pi_0$;

Using a depth-first search with root π, search for a permutation τ such that $\text{cost}(\tau) < \text{cost}(\pi)$ and that the corresponding minimal I-MAP D_τ is connected to D_π by a list of $\mathcal{I}$-covered arrow reversals. If such τ exists, set π as τ and continue this step; otherwise, for each k, set $I^k := \{j : f^{I^k}(j \mid \text{pa}_{D_\pi}(j)) \neq f^{\text{obs}}(j \mid \text{pa}_{D_\pi}(j)\}$ and return π, D_π and $\mathcal{I} := \{I^1, \cdots, I^K\}$.

restriction will improve the efficiency of our algorithm by reducing the number of search directions at each step.

In each interventional setting $k \in [K]$, we denote the set of known intervention targets as I_{Kn}^k. In the setting where intervention targets are completely unknown, $I_k^{\text{Kn}} = \emptyset$, and in the setting where they are completely known, $I_{\text{Kn}}^k = I^k$. Many setting, such as gene knockout experiments, may fall in between, e.g., I_{Kn}^k may include the intended intervention targets but $I^k \setminus I_{\text{Kn}}^k$ may contain off-target effects.

Definition 2. *An edge $i \to j$ is $\mathcal{I}$-covered in D_π if it is a covered arrow in D_π, and for all k s.t. $i \in I_{Kn}^k$, $(j \mid \text{pa}_{D_\pi}(j)) \notin \mathcal{V}^k$.*

This definition is easiest to read in terms of which covered arrows it excludes from being $\mathcal{I}$-covered: we no longer consider flipping an edge $i \to j$ if i is intervened on but the conditional distribution of j remains the same in some interventional setting, since this implies that we have the correct set of parents for j.

We now state our main result, which guarantees the consistency of UT-IGSP:

Theorem 2. *Under Assumption 1, Algorithm 2 is consistent, i.e., given infinite data, it discovers the true DAG D and the correct sets of intervention targets $I^1, \ldots, I^K$.*

Since the proof that Algorithm 2 is consitent under the faithfulness assumption on the JCI-DAG (henceforth, JCI-faithfulness) is much simpler, and the relationship of that assumption to Assumption 1 is not established, we will describe here the proof

of consistency of JCI-faithfulness, though we note that the above theorem has an alternative proof that will appear in future versions of this work if necessary.

Briefly, if $i \to j$ is covered in $(D_C)_\pi$ (the JCI DAG associated with permutation π), then $i \to j$ is $\mathcal{I}$-covered in D_π. To see this, note that the covered edges of D_π are a subset of the covered edges of $(D_C)_\pi$, since only the parents which are system variables must match in the former case whereas all parents must match in the latter case. Furthermore, if $(j \mid \mathrm{pa}_{D_\pi}(j)) \in \mathcal{V}^k$, but $i \in I^k_{\mathrm{Kn}}$, then $i \to j$ is also not covered in $(D_C)_\pi$ since i is a child of the context variable C_k, but j is not. Thus, UT-IGSP has some superset of the search directions of JCI-GSP at each step, and thus if there is a weakly decreasing path from $(D_C)_\pi$ to the true D_C, there is also a weakly decreasing path from D_π to the true D.

2.6 Hypothesis Tests

In this section, we explain the hypothesis tests for conditional independence and conditional invariance in the case of linear Gaussian data. We assume that there are n_{obs} samples of observational data and n_k samples of data from the k^{th} interventional setting. The sample covariance matrices $\hat{\Sigma}^{\mathrm{obs}}$ and $\hat{\Sigma}^k$, $k \in [K]$, are sufficient statistics, i.e., all of the following statistics can be computed from these alone. Throughout, we assume that we test the null hypothesis at significance level α.

2.6.1 Conditional Independence Hypothesis Test

If x has a multivariate Gaussian distribution, the conditional independence statement $x_i \perp\!\!\!\perp x_j \mid x_C$ is equivalent the partial correlation $\rho_{ij|C} = 0$. Moreover, given n samples of x, it is well-known that if $\rho_{ij|C} = 0$, then the Fisher z-transform $\hat{z}_{ij|C} = \frac{1}{2} \ln \frac{1+\hat{\rho}_{ij|C}}{1-\hat{\rho}_{ij|C}}$ of the sample partial correlation $\hat{\rho}_{ij|C}$ is approximately normally distributed with mean 0 and standard deviation $\frac{1}{\sqrt{n-|C|-3}}$ (see for example [10]).

Thus, if f^{obs} and all f^{I^k} are multivariate Gaussian, we can test the null hypothesis $x_i \perp\!\!\!\perp x_j \mid x_C$ by computing $\hat{z}_{ij|C}$ from each dataset and rejecting if any of the p-values are less than $\frac{\alpha}{K+1}$, where the denominator comes from the Bonferroni correction for

multiple hypothesis tests.

2.6.2 Conditional Invariance Hypothesis Test

In a multivariate Gaussian distribution $f(x)$, the conditional distribution $f(x_i \mid x_C)$ can be expressed as a normal distribution with a mean that is an affine function of x_C, i.e., $f(x_i \mid x_C) = \mathcal{N}(x_i; \beta_{i|C}[x_C^\top, 1], s_{i|C})$.

By linear regression of x_i on x_C, we can compute the regression coefficients $\hat{\beta}^{\text{obs}}_{i|C}$ and $\hat{\beta}^{k}_{i|C}$, respectively, and unbiased residual variance estimators $\hat{s}^{\text{obs}}_{i|C}$ and $\hat{s}^{k}_{i|C}$, respectively. Suppose we have equal variances, i.e., $s^{k}_{i|C} = s^{\text{obs}}_{i|C}$. Then, if the null hypothesis $\beta^{k}_{i|C} = \beta^{\text{obs}}_{i|C}$ is true, it is well-known that the test statistic

$$\hat{T} = \frac{1}{|C|}(\hat{\beta}^{\text{obs}}_{i|C} - \hat{\beta}^{k}_{i|C})^\top(\hat{s}^{\text{obs}}_{i|C}(n_{\text{obs}}\Sigma^{\text{obs}}_{CC})^{-1} + \hat{s}^{k}_{i|C}(n_k\Sigma^{k}_{CC})^{-1})^{-1}(\hat{\beta}^{\text{obs}}_{i|C} - \hat{\beta}^{k}_{i|C})$$

follows a $F(|C|, n_{\text{obs}} + n_k - |C|)$ distribution (see for example [5]).

Next, we can use an F-test to compute equality of the estimated residual variances, and test both statistics at significance level $\frac{\alpha}{2}$.

Technically, we should simultaneously test for equality of the regression coefficients and residual variances, but that test (found in [15]) is more computationally expensive and we find that the suggested approach performs well empirically.

Chapter 3

Experimental Results

In this section, we compare UT-IGSP and JCI-GSP to the known-target interventional structure learning methods IGSP and GIES on the task of recovering the true $\mathcal{I}$-MEC from interventional data. We generate data from a ground truth DAG D on p nodes and set of intervention targets. The distribution over the variables $X = x_1, \ldots, x_p$ will follow a linear structural equation model with additive Gaussian noise, i.e.,

$$X = WX + \epsilon$$

where W an upper-triangular matrix with the same sparsity pattern as D and $\epsilon \sim \mathcal{N}(0, I_p)$. In each setting, we average the results over 100 random DAGs from an Erdös-Rényi model with 1.5 expected neighbors. The non-zero entries of W are sampled indepdendent from a uniform distribution over $[-1, -.25] \cup [.25, 1]$ (the bounding away from zero ensures the presence of an edge is discernible in our sample regime).

For each DAG, we generate intervention targets from $K = 5$ settings. Since IGSP and GIES require at least one known intervention target, we first sample 5 nodes from $[p]$ without replacement and assign one to each of $I^1_{\text{Kn}}, \ldots, I^5_{\text{Kn}}$. We vary the number of additional, unknown intervention targets from $\ell = 0$ to $\ell = 3$, picked independently for each I^k from $[p] \setminus I^k_{\text{Kn}}$. In all significance tests, we use a significance level of $\alpha = 10^{-5}$, which we have found to perform well in experiments. All code is implemented as part of the Python package `causaldag`.

3.1 Metrics

In this section, we discuss the metrics we use to evaluate the performance of the algorithms for different tasks.

3.1.1 Graph Recovery Metrics

We measure the accuracy of the recovered graph in two ways.

The *structural Hamming distance* between two mixed graphs G and G' is the minimum number of edge additions, deletions, and type changes (i.e., changing direction or changing between undirected/directed) required to transform G into G' (or vice versa), and is denoted $\mathrm{SHD}(G, G')$. Given the output DAG $\hat{D}$ of an algorithm and the output intervention targets $\hat{\mathcal{I}}$, we compute the structural Hamming distance between the estimated interventional essential graph $\mathcal{E}_{\hat{\mathcal{I}}}(\hat{D})$ and the true one, $\mathcal{E}_{\mathcal{I}}(D)$.

Second, we measure exact recovery of the true interventional essential graph, i.e., the proportion of runs for which $\mathcal{E}_{\hat{\mathcal{I}}}(\hat{D}) = \mathcal{E}_{\mathcal{I}}(D)$.

3.1.2 Intervention Recovery Metrics

To measure the accuracy of intervention recovery, which is only possible in UT-IGSP and JCI-GSP, we calculate the number of false positives (i.e., $|\hat{\mathcal{I}} - \mathcal{I}|$) and the number of false negatives (i.e., $|\mathcal{I} - \hat{\mathcal{I}}|$).

3.2 Accuracy Comparison: Perfect Interventions

In this section, we explored the performance of the algorithms under *perfect* interventions, in which the intervened nodes become independent of their parents. Specifically, in Figures 3-1, 3-2, 3-3, 3-4, 3-5, and 3-6, each intervened node $i \in I^k$ had its conditional distribution changed to $f^{I^k}(x_i \mid x_{\mathrm{pa}_D(i)}) = \mathcal{N}(x_i; 0, .1)$, and in Figures 3-7, 3-8, and 3-9, each intervened node $i \in I^k$ had its conditional distribution changed to $f^{I^k}(x_i \mid x_{\mathrm{pa}_D(i)}) = \mathcal{N}(x_i; 1, .1)$. False negative intervention targets are not shown since neither UT-IGSP nor JCI-GSP had any false negatives in these settings.

When $p = 10$, in the case where all intervention targets are known, GIES outperforms all other algorithms in both interventional settings and both metrics. However, the performance of GIES quickly deteriorates as the number of off-target effects increases, and we even see in Fig. 3-7 that for large ℓ, the performance deteriorates as the number of samples increases. Surprisingly, IGSP, which is designed only for cases in which all intervention targets are known, still performs as well or better than JCI-GSP on the structure learning task.

Indeed, as the number of off-target effects increases, we see that in both interventional settings and both structure-learning metrics, UT-IGSP outperforms all other methods. However, JCI-GSP and UT-IGSP are roughly equal in performance on intervention target recovery, which might be attributed to the fact that their only difference is through their differing notion of covered edges, which directly affects the causal structure and only indirectly affects the intervention targets.

Figures 3-4, 3-5, and 3-6 study the (nearly) high-dimensional case, i.e., when the number of variables ($p = 30$) is close to the number of samples ($n = 40, 50, 60, 70, 80$). In this case, there are not enough to learn the true interventional Markov equivalence class of any of the DAGs, but the same trends remain true for the structural Hamming distance as were present in the low-dimensional case.

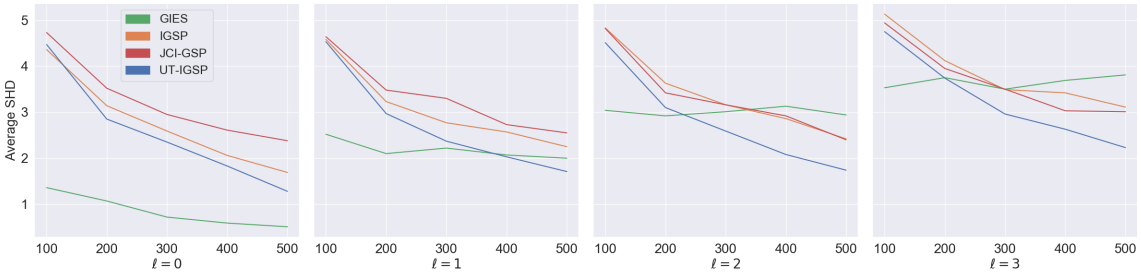

Figure 3-1: **Zero-mean perfect interventions,** $p = 10$**.** Average Structural Hamming distance between true $\mathcal{I}$-essential graph and estimated $\mathcal{I}$-essential graph over 100 DAGs with 1 known target and ℓ unknown targets per intervention setting.

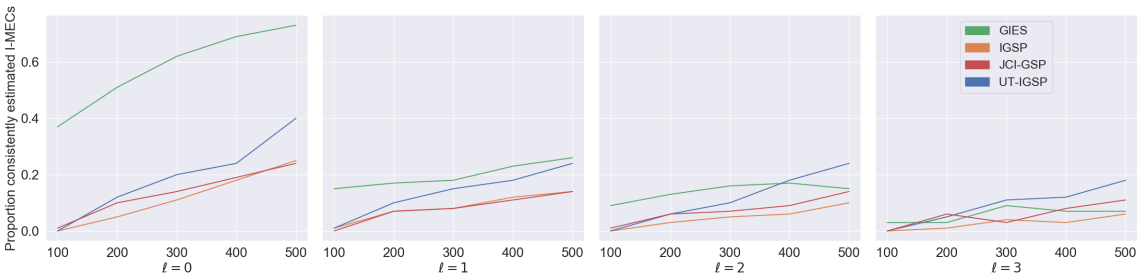

Figure 3-2: **Zero-mean perfect interventions,** $p = 10$**.** Proportion of $\mathcal{I}$-essential graph correctly estimated over 100 DAGs with 1 known target and ℓ unknown targets per intervention setting.

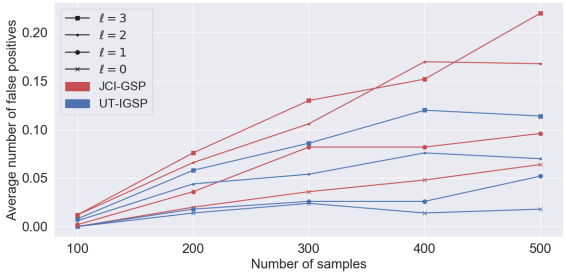

Figure 3-3: **Zero-mean perfect interventions,** $p = 10$**.** Average number of false positive intervention targets over 100 DAGs with 1 known target target and ℓ unknown target per intervention setting.

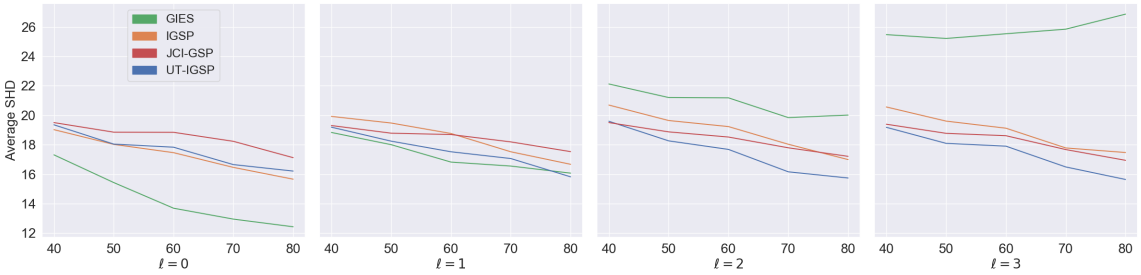

Figure 3-4: **Zero-mean perfect interventions,** $p = 30$**.** Average Structural Hamming distance between true $\mathcal{I}$-essential graph and estimated $\mathcal{I}$-essential graph over 100 DAGs with 1 known target and ℓ unknown targets per intervention setting.

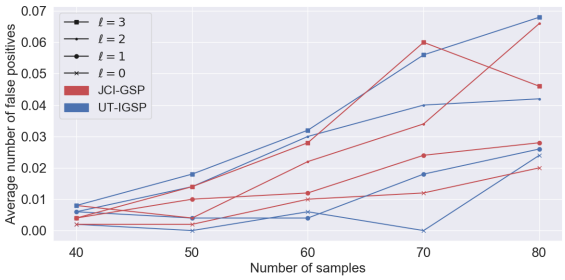

Figure 3-5: **Zero-mean perfect interventions,** $p = 30$**.** Average number of false positive intervention targets over 100 DAGs with 1 known target target and ℓ unknown target per intervention setting.

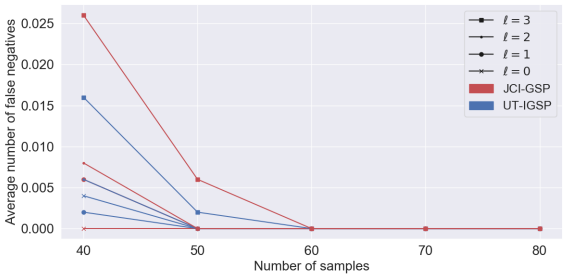

Figure 3-6: **Zero-mean perfect interventions,** $p = 30$**.** Average number of false positive intervention targets over 100 DAGs with 1 known target target and ℓ unknown target per intervention setting.

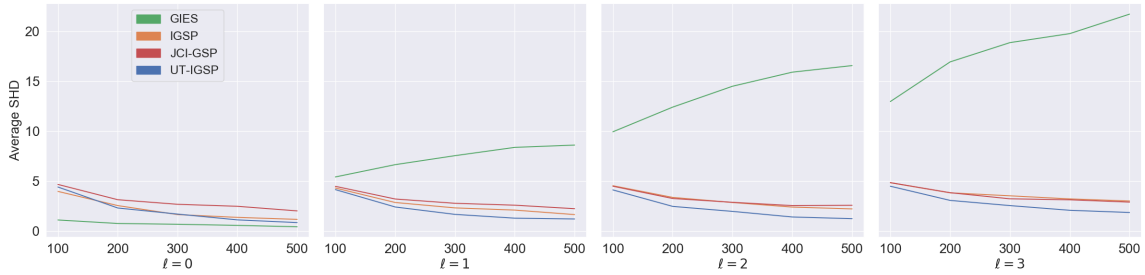

Figure 3-7: **Mean-1 perfect interventions,** $p = 10$**.** Average Structural Hamming distance between true $\mathcal{I}$-essential graph and estimated $\mathcal{I}$-essential graph over 100 DAGs with 1 known target and ℓ unknown targets per intervention setting.

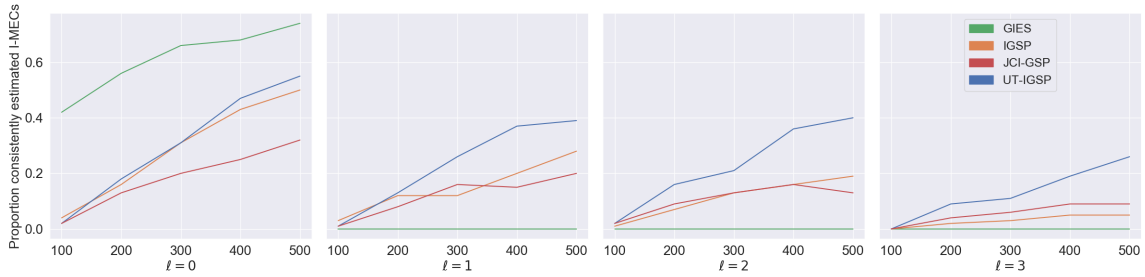

Figure 3-8: **Mean-1 perfect interventions,** $p = 10$**.** Proportion of $\mathcal{I}$-essential graph correctly estimated over 100 DAGs with 1 known target and ℓ unknown targets per intervention setting.

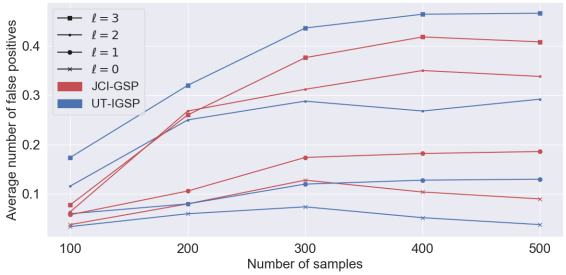

Figure 3-9: **Mean-1 perfect interventions,** $p = 10$**.** Average number of false positive intervention targets over 100 DAGs with 1 known target target and ℓ unknown target per intervention setting.

3.3 Accuracy Comparison: Soft Interventions

In this section, we explored the performance of the algorithms under *soft* interventions, in which the intervened nodes still have a dependence on their parents, but the exact conditional probability is altered. Specifically, in Figures 3-10, 3-11, 3-12, and 3-13, each intervened node $i \in I^k$ had its conditional distribution changed from $f^{\text{obs}}(x_i \mid x_{\text{pa}_D(i)}) = \mathcal{N}(x_i; \beta x_{\text{pa}_D(i)}, 1)$ to $f^{I^k}(x_i \mid x_{\text{pa}_D(i)}) = \mathcal{N}(x_i; .5\beta x_{\text{pa}_D(i)}, .5)$.

Now, UT-IGSP outperforms all other methods in all metrics and all settings, achieving an average structural Hamming distance from the true interventional essential graph of only one edge. GIES always performs poorly, likely due to the strong dependence remaining between the intervened node and its parents.

In this setup, we also find that UT-IGSP outperforms JCI-GSP in terms of intervention target recovery by more than a factor of 2 in false positives, although JCI-GSP performs slightly better in terms of false negative rate when the number of samples is low.

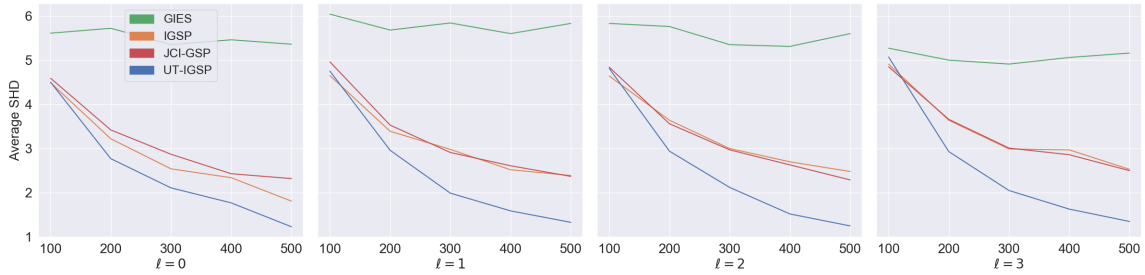

Figure 3-10: **Soft interventions,** $p = 10$**.** Average Structural Hamming distance between true $\mathcal{I}$-essential graph and estimated $\mathcal{I}$-essential graph over 100 DAGs with 1 known target and ℓ unknown targets per intervention setting.

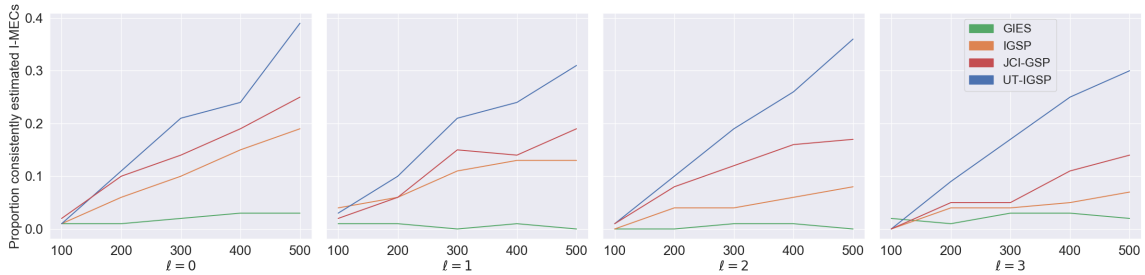

Figure 3-11: **Soft interventions,** $p = 10$**.** Proportion of $\mathcal{I}$-essential graph correctly estimated over 100 DAGs with 1 known target and ℓ unknown targets per intervention setting.

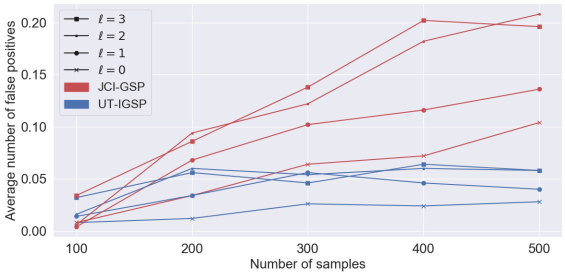

Figure 3-12: **Soft interventions,** $p = 10$**.** Average number of false positive intervention targets over 100 DAGs with 1 known target target and ℓ unknown target per intervention setting.

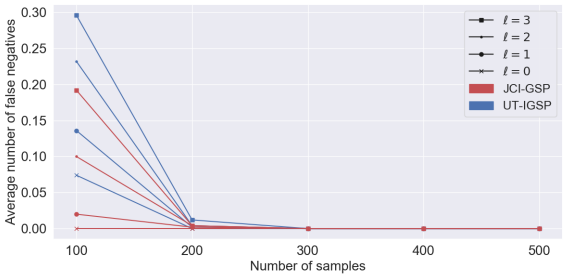

Figure 3-13: **Soft interventions,** $p = 10$**.** Average number of false negative intervention targets over 100 DAGs with 1 known target target and ℓ unknown target per intervention setting.

Chapter 4

Discussion and Future Work

In this thesis, we introduced the Unknown Target Intervention Greedy Sparsest Permutation (UT-IGSP) algorithm for causal structure learning when intervention targets are unknown, and showed that it performs favorably in experiments to JCI-GSP and the known-target interventional structure learning algorithms IGSP and GIES.

Learning causal structure from data that comes from multiple contexts, such as the interventional contexts that are explored in this thesis, is of fundamental scientific and decision-making interest. Heterogeneity of contexts helps with the identification of causal structure, which in turn helps in downstream tasks such as treatment assignment, prediction after domain shift, and treatment discovery.

Future work along this direction may continue to combine elements of ordering-based causal inference and joint causal inference, for example by removing the assumption of causal sufficiency, i.e., allowing for latent confounders.

The experimental results in this work raise the question of what makes some algorithms, such as IGSP, more robust to modeling mispecifications than others, such as GIES. Understanding the reasons for the differences between these results could open the way to designing structure-learning algorithms that are robust against violations of their assumptions, which in practice are never likely to be completely true.

Appendix A

Graph Terminology and Notation

A *directed graph* $D = (V, E)$ consists of nodes V and edges E between ordered pairs of nodes, with only a single edge between two vertices. We denote D has an edge from i to j by $i \to j \in D$.

We describe the neighborhood of a vertex i with the sets $\mathrm{pa}_G(i) = \{j \in V \mid j \to_G i\}$ (the *parents* of i) and $\mathrm{ch}_G(i) = \{j \in V \mid i \to j\}$ (the *children* of i).

A *chain* between vertices i and j is a sequence γ of distinct adjacent vertices starting at i and ending at j. A *path* between i and j is a chain such that all edges point to later nodes in the sequence, i.e., $\gamma_i \to \gamma_{i+1}$. A directed cycle is a directed path from i to j together with an edge $j \to i$. A *directed acyclic graph* (DAG) is a directed graph with no cycles. We say that j is a collider on a chain γ if $j = \gamma_i$ and $\gamma_{i-1} \to \gamma_i \leftarrow \gamma_{i+1}$. Otherwise, we say that j is a non-collider on γ. We denote the set of colliders on γ by $\mathrm{coll}(\gamma)$ and the set of noncolliders by $\mathrm{ncoll}(\gamma)$.

The *ancestors* of a node i, denoted by $\mathrm{an}_D(i)$, are the nodes with a path to i, including i itself. Given any function on single nodes, we overload it to take sets as arguments by returning the union, e.g. $\mathrm{an}_D(A) = \cup_{i \in A} \mathrm{an}_D(i)$.

The *induced subgraph* of G on a vertex set $V' \subseteq V$, denoted $G[V']$, is the graph with vertices V' and edges from G for which both endpoints are in V'. A *v-structure* (also called an *immorality* or *unshielded colliders*) is an induced subgraph of the form $i \to j \leftarrow k$.

Given a set C, we say that two nodes i and j are *d-connected by a chain* γ if

$\mathrm{ncoll}(\gamma) \cap C = \emptyset$ and $\mathrm{coll}(\gamma) \subseteq \mathrm{an}_D(C)$. We say that i and j are *d-connected* given C if they are d-connected by any chain. Otherwise, we say i and j are d-separated.

Bibliography

[1] David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

[2] Antonia A Dominguez, Wendell A Lim, and Lei S Qi. Beyond editing: repurposing crispr–cas9 for precision genome regulation and interrogation. *Nature reviews Molecular cell biology*, 17(1):5, 2016.

[3] Daniel Eaton and Kevin Murphy. Exact bayesian structure learning from uncertain interventions. In *Artificial Intelligence and Statistics*, pages 107–114, 2007.

[4] Frederick Eberhardt and Richard Scheines. Interventions and causal inference. *Philosophy of Science*, 74(5):981–995, 2007.

[5] AmirEmad Ghassami, Saber Salehkaleybar, Negar Kiyavash, and Kun Zhang. Learning causal structures using regression invariance. In *Advances in Neural Information Processing Systems*, pages 3011–3021, 2017.

[6] Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13(Aug):2409–2464, 2012.

[7] Steffen L Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

[8] Joris M Mooij, Sara Magliacane, and Tom Claassen. Joint causal inference from multiple contexts. *arXiv preprint arXiv:1611.10351*, 2016.

[9] Judea Pearl. *Causality: models, reasoning and inference*, volume 29. Springer, 2000.

[10] Liam Solus, Yuhao Wang, Lenka Matejovicova, and Caroline Uhler. Consistency guarantees for permutation-based causal inference algorithms. *arXiv preprint arXiv:1702.03530*, 2017.

[11] Peter Spirtes, Clark N Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper, and Thomas Richardson. *Causation, prediction, and search*. MIT press, 2000.

[12] Thomas Verma and Judea Pearl. *Equivalence and synthesis of causal models*. UCLA, Computer Science Department, 1991.

[13] Xiaoling Wang, Yebo Wang, Xiwei Wu, Jinhui Wang, Yingjia Wang, Zhaojun Qiu, Tammy Chang, He Huang, Ren-Jang Lin, and Jiing-Kuan Yee. Unbiased detection of off-target cleavage by crispr-cas9 and talens using integrase-defective lentiviral vectors. *Nature biotechnology*, 33(2):175, 2015.

[14] Yuhao Wang, Liam Solus, Karren Yang, and Caroline Uhler. Permutation-based causal inference algorithms with interventions. In *Advances in Neural Information Processing Systems*, pages 5822–5831, 2017.

[15] Samaradasa Weerahandi. Testing regression equality with unequal variances. *Econometrica: Journal of the Econometric Society*, pages 1211–1215, 1987.

[16] Karren D Yang, Abigail Katcoff, and Caroline Uhler. Characterizing and learning equivalence classes of causal dags under interventions. *arXiv preprint arXiv:1802.06310*, 2018.