Practically Applicable Causal Discovery

A dissertation submitted towards the degree
Doctor of Engineering (Dr. Ing.)
of the Faculty of Mathematics and Computer Science
of Saarland University

by Osman Ali Mian

Saarbrücken, 2024

Reviewer number 1
Reviewer number 2
Reviewer number 3

Abstract

This thesis focuses on discovering causal dependencies from observational data, which is one of the most fundamental problems in science. In particular, causal discovery aims to discover directed graphs among a set of observed random variables under specified assumptions. While an active area of research, existing causal discovery approaches are not always applicable to real-world scenarios. This is mainly due to their underlying assumptions, which limit their applicability in practice.

In this dissertation, we aim to develop approaches that can be applied to several real-world scenarios to discover causal dependencies, under mild assumptions. We first focus on a setting where we discover the complete causal DAG and not just the Markov equivalence class from observational data. We do so by using the principle of choosing the simplest explanation, measured in information-theoretic terms, to develop a theoretically sound causal discovery method. Next, we extend causal discovery to data collected across multiple environments, addressing biases from pooling data with different interventional distributions. To this end, we propose an approach that uses a similar information-theoretic score to discover causal networks in distributed settings without requiring prior knowledge of whether the data is observational or interventional. Furthermore, we develop a method for continual causal discovery from episodic data that updates causal hypotheses as new data arrives, without the need to re-learn causal networks from scratch each time. Our proposed approach for this scenario can learn causal networks adaptively over time and distinguish between episodes that do not belong to the same causal mechanism. Lastly, we tackle the important aspect of privacy-preserving federated causal discovery. To do so, we propose a general framework that effectively identifies global causal networks without ever sharing the data or learning parameters, while ensuring differential privacy.

Zusammenfassung

Diese Arbeit befasst sich mit der Entdeckung kausaler Abhängigkeiten aus Beobachtungsdaten, eines der grundlegendsten Problemen der Wissenschaft. Insbesondere zielt die kausale Entdeckung darauf ab, gerichtete Graphen über einer Reihe von beobachteten Zufallsvariablen unter bestimmten Annahmen zu entdecken. Obwohl es sich hierbei um ein aktives Forschungsgebiet handelt, sind bisherige Ansätze nicht in reale Szenarien anwendbar. Dies liegt vor allem an den ihnen zugrunde liegenden Annahmen, die ihre Anwendbarkeit in der Praxis einschränken.

In dieser Dissertation wollen wir Ansätze entwickeln, die unter milden Annahmen auf verschiedene reale Szenarien angewendet werden können, um kausale Abhängigkeiten zu entdecken. Wir konzentrieren uns zunächst auf eine Situation, in der wir den vollständingen kausalen Graphen und nicht nur die Markov-Äquivalenzklasse aus Beobachtungsdaten ermitteln. Dazu verwenden wir das sogenannte Minimum Description Length Prinzip, bei dem die kausale Hypothese nach der einfachsten Erklärung ausgewählt wird, um eine theoretisch fundierte Methode zur Entdeckung von kausalen Abhangigkeiten zu entwickeln. Als Nächstes erweiten wir die kausale Entdeckung auf Daten aus, die in verschiedenen Umgebungen erhoben wurden, und gehen dabei auf Verzerrungen ein, die sich aus der Zusammenführung von Daten mit unterschiedlichen Interventionsverteilungen ergeben. Zu diesem Zweck schlagen wir einen Ansatz vor, der eine ähnliche informationstheoretische Bewertungsmetrik verwendet, um kausale Netzwerke in verteilten Umgebungen zu entdecken, ohne dass vorher bekannt sein muss, ob es sich um Beobachtungs- oder Interventionsdaten handelt. Darüber hinaus entwickeln wir eine Methode zur kontinuierlichen Entdeckung kausaler Zusammenhänge aus episodischen Daten, die kausale Hypothesen beim Eintreffen neuer Daten aktualisiert, ohne kausale Netzwerke jedes Mal von Grund auf neu zu lernen. Der von uns vorgeschlagene Ansatz für dieses Szenario kann kausale Netzwerke über die Zeit hinweg adaptiv lernen und zwischen Episoden unterscheiden, die nicht zum selben kausalen Mechanismus gehören. Schließlich befassen wir uns mit dem wichtigen Aspekt der datenschutzfreundlichen föderierten Kausalerkennung. Zu diesem Zweck schlagen wir einen allgemeinen Rahmen vor, der effektiv globale kausale Netzwerke identifiziert, ohne dass die Daten oder Lernparameter geteilt werden müssen, während gleichzeitig der Schutz der Privatsphäre gewährleistet wird.

And it is God who sends the winds, and then they stir the clouds, and then We drive them to a dead land and then we give life thereby to the earth after its lifelessness. Thus is the resurrection.

[Quran, 35:9]

Acknowledgments

I would like to thank Prof. Vreeken for agreeing to be the advisor for my Ph.D. and helping me bring the best out of me. Over the last five years (and the two before) I have had the honor of learning so much from him. I am glad I decided to do a Ph.D. under his supervision. It has been nothing short of wonderful to work under his guidance.

This acknowledgment section cannot be complete without thanking Alexander Marx. My first go-to person as I started my journey in the land of causal discovery. He continuously supported me and helped me establish the right foundations for this line of work. A large part of my early work would not have been possible without input from Alex in one way or the other.

I am grateful to Michael Kamp, who was very much an unofficial second advisor, a mentor, a collaborator, and a friend, all at the same time.

A special thanks to my coauthors, David Kaltenpoth and Sarah Mameche, for their wonderful collaborations, for bringing the missing pieces of the puzzle to my research ideas, and for giving valuable feedback on this dissertation manuscript.

A big shout out to my wife, the woman who went through this entire doctoral journey with me through the highs and lows with her unwavering belief in my abilities, and for being my support mechanism throughout.

I am eternally grateful to my parents for their support and prayers, and to my sisters, for their motivation and their 'you can do it!' texts.

Table of Contents

Chapter 1

Introduction

The necessity of reasoning about causes and effects is a fundamental part of human nature. From taking action against the imminent danger of a sabertooth tiger lurking around thousands of years ago to learning to take medicine when having a fever, causal reasoning has helped humans survive and thrive on Earth. The study of causes and effects, therefore, is a fundamental problem in all sciences, as having an understanding of the way the world works, i.e. the world model, gives a significant advantage in that it allows humans to reason by *simulating* reality without ever living in it, and hopefully taking better decisions as a result. The curiosity to reason about the world around us motivates researchers to investigate causes and effects in a number of real world settings where our brain does not necessarily know how the world might work. This could happen due to a number of reasons such as novelty of a situation (should I be as afraid for my life if a sabertooth tiger was replaced by a dinosaur?), potential future speculation (will this code also give the same output if we would run it on the computer of my supervisor?), or hypothetical open-ended scenarios (what would happen to the motivation of our employees if we gave a raise to everyone at the company?).

In most situations, the gold standard to infer causality is to meddle with the nature by intervening on the natural process. One such way to intervene is to perform a controlled experiment. Controlled experiments are studies where researchers manipulate one variable while keeping all other variables constant to see the effect of the manipulation. "Participants" are often divided into groups, with one group receiving the treatment and another group not receiving it. This allows researchers to attribute any differences in outcomes between the groups to the manipulated variable, thereby establishing a causal relationship between

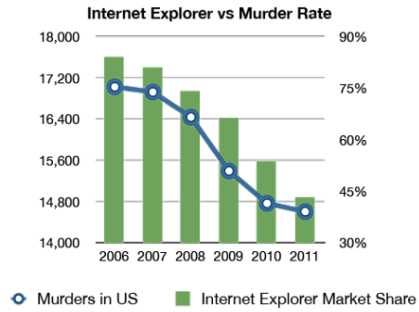


Figure 1.1: [Correlation does not always imply causation]: Just because number of murders predicts the usage of Internet Explorer web browser well, does not mean one causes the other unless we confirm it by performing controlled experiments. Performing a controlled experiment here is impossible as it would require forcing people to commit the illegal act of murder, to check for changes in Internet Explorer usage (Charlatan, 2013).

the manipulated variable and the outcome variable. While controlled experiments are the gold-standard, performing them might be impossible (where do we get dinosaurs from?), unethical (my supervisor would not appreciate me using his computer to run my rookie code) or exorbitant (once we give a raise, we can not reverse it) in many situations.

On the one hand, controlled experiments are not always feasible, on the other hand we can still "observe" the world around us and try to infer causation from it. We can, however, only do so much — it should be easy to see that correlation does not necessarily imply causation. Take Fig. 1.1 for example, just because we observe a correlation between number of murders and Internet Explorer usage, does not mean that one is necessarily the cause of the other. While the latter can not be ruled out, it could be that there is an unobserved third variable that explains both of these phenomena. Without controlled experiments we may not be able know, yet we can not conduct controlled experiments here as it would require forcing people to commit the illegal act of murder, to check for changes in Internet Explorer usage.

This should give reader a glimpse of the difficulty of the task at hand. Pearl (2009), in fact, defines learning problem to be a ladder consisting of three rungs: Associations (seeing), Interventions (doing), and Counterfactuals (imagining). The first rung, Association, involves identifying correlations between variables without implying causation. It focuses on observing patterns and statistical relationships in data, such as noticing that two events often occur together. This level does not address whether one variable causes the other, only that they are linked in some way. Pearl (2009) even shows that it is impossible to derive causal conclusions using only associations unless we make assumptions on how

the data was generated. Extracting causal knowledge from observational data, under a given set of necessary assumptions is the focus of research areas known as causal inference (Pearl et al., 1991; Mohan et al., 2013; Anand et al., 2023) and causal discovery (Spirtes et al., 2000a; Chickering, 2002; Peters et al., 2017). Methods that aim to estimate effects of intervening on the system, given a pre-specified causal structure and observational data are called *causal inference* approaches. On the other hand, methods that aim to learn causal structures from observational data under specified assumptions are known as *causal discovery* approaches. In this thesis, we foucs on causal discovery.

The desiderata for a useful causal discovery approach is three-fold. First, it should be theoretically sound, ensuring that it can accurately identify causal relationships with infinite data. Second, it should make reasonable assumptions that we can expect to be fulfilled with (near) infinite data. Finally, the method must demonstrate reliable performance on existing datasets with known ground truth to indicate that its theoretical guarantees are transferable to real-world scenarios. Methods that meet only some of these criteria may face limitations in their practical use; lacking soundness guarantees undermines the certainty of causal conclusions, strict underlying assumptions may not align with real-world conditions, and poor performance on known data could indicate fundamental flaws in theoretical formulations.

Most of the approaches today only address the first aspect, and only a handful address two of the three requirents. A number of necessary yet restrictive assumptions lie at the core of existing, otherwise theoretically sound, causal discovery approaches. These assumptions range from linearity of relationships between causes and effects, to having single, fully-specified, static, centralized, homogeneous, unbiased data to learn from, none of which is usually the case. This hampers existing methods' performances in real world scenarios. The absence of practically useful causal discovery approaches under reasonable assumptions creates a gap, which motivates the work we present in this thesis.

In this dissertation we develop causal discovery approaches that attempt to achieve this triad of theoretical correctness, reasonability of assumptions and strong practical performance. We investigate and relax a number of existing assumptions such as the need for data centralization, requiring prior knowledge of data heterogeneity, or learning from scratch each time data gets updated. We prove the correctness of our proposed approaches and conduct extensive experiments to show that our proposed methods perform well in practice and beat the state-of-the-art both in terms of causal discovery as well as a variety of practical aspects such as privacy considerations, thereby paving the way towards making causal discovery useful for real world problems. In the following sections we briefly discuss how causal discovery works in general, what are the shortcomings of existing approaches, open problems, and how we attempt to address some of these limitations.

1.1 Problem Overview

The goal of causal discovery is to learn causal relationships, under specified assumptions, from observational data between a set of defined variables. This could be, for example observational data containing records of patients with a set of symptoms leading to a disease. These causal relationships can be defined by a Structural Causal Model (SCM). An SCM over each variable X in a given variable set $\boldsymbol{X}$ assigns value to X in the form

$$X := f(pa_X, N_X) \ ,$$

where f is a complicated modeling function that only depends on causal parents pa_X of X, and additional uncertainty in the system that we encapsulate as noise N_X associated with variable X. The noise N_X can be considered to be an abstraction of factors that influence the outcome X but are unobserved in data and could, for example, be modeled as Gaussian distributed. As an example, consider a system of 6 covariates, represented using the following SCM,

$$\begin{aligned} U &:= f_U(N_U) & V &:= f_V(N_V) \\ W &:= f_W(U, N_W) & X &:= f_X(V, W, N_X) \\ Y &:= f_Y(W, N_Y) & Z &:= f_Z(X, N_Z) \ . \end{aligned}$$

Together, these equations let us model each covariate as a random variable and define a joint distribution over these covariates. This lets us reason about the system at a modular level where changing the value of any variable, or concretely stated, performing a hypothetical *intervention* only effects the variable in question without altering mechanisms of other variables. If, for example, we were to alter variable X and set it to some fixed constant c, we can represent this changed state by an SCM that differs only in generation of X by replacing $X := f_X(V, W, N_X)$ with $X := c$ while leaving all other variables as they are. This particular intervention is known in literature as a *hard intervention*. Alternatively we could also introduce *soft interventions* or *mechanism changes*, which roughly means that instead of assigning a specific value to X we replace its generating mechanism by a different function $g_X(\hat{pa}_X, N_X)$, that may use only a subset of X's parents. Having access to the underlying SCM for a set of variables, constitutes the third rung (Counterfactuals) of Pearl's causal hierarchy. At this rung, we have a fully defined causal model that can be used to *simulate* potential outcomes.

The causal relationships implied by an SCM can be presented in the form of a causal network, or causal graph. The causal graph for the covariates from the above-mentioned SCM is shown in Fig. 1.2. A causal graph consists of nodes that represent variables from dataset at hand, and edges that represent

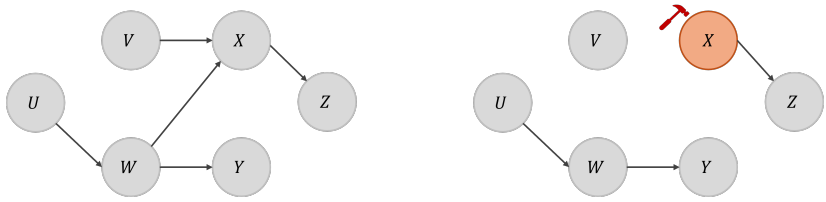


Figure 1.2: [Left] Example of a causal network depicting causes of a given set of variables through directed edges going from the cause to the effect, for an example scenario. Variables without incoming edges, such as U and V, are only influenced by external stochastic factors. **[Right]** Intervening on X by setting it to a fixed value does not change how we compute the values for all other variables.

causal links. A direct edge from variable U to variable W implies that U is the direct cause, or causal parent, of W in that it listens to changes in U in some real world scenario being modeled. Similarly we call U an ancestor of X and Z, because there is a sequence of directed edges that take us from U to X resp. Z. We can update this graph to reflect a hard intervention on X by removing any incoming edges from its parents. Similarly, a soft-intervention may be reflected in the causal graph with X missing some, and not all of its incoming edges. Knowing the causal graph for a set of variables, even in the absence of a Structural Causal Model (SCM), represents the second level (Intervention) of Pearl's causal hierarchy. By applying the do-calculus as outlined by Pearl (2009), it may become possible to infer the potential outcomes of a controlled experiment from observational data alone, without the need to conduct the experiment directly.

The discussion up till this point highlights that the distinction between the first and second levels of the causal hierarchy lies in the availability of a causal network. This causal network is required for advancing to the second rung and is essential for identifying the correct Structural Causal Model (SCM) needed to progress to the third rung. This leads us to the central question of this work: how do we learn this causal network? This, we describe next.

1.1.1 Causal Discovery

Given data sampled from observational distribution induced by an unknown SCM, a causal discovery algorithm aims to *learn* the underlying causal graph. For purely observational data this is impossible unless we are willing to make assumptions (Pearl, 2009). The two most common, and necessary, assumptions include acyclicity, which implies that an effect can not be an ancestor of its own cause, and causal sufficiency meaning that all noise variables are independent of each other and subsequently do not affect more than one observed variables.

Causal sufficiency assumption boils down to assuming that all relevant causes of each variable in a dataset are also included in that dataset. Acyclicity and sufficiency assumptions let us to describe the causal network by a directed acyclic graph (DAG), like the one shown in Fig. 1.2.[1]

Once we can model causal networks using DAGs, we need to make assumptions under which we can reliably test for causality in a given dataset. Two such assumptions to do so are the causal Markov condition (CMC) and the faithfulness assumption. A combination of both these assumptions implies that two variables uncorrelated in the data are not linked to each other in the true causal graph and vice versa. We can leverage this to weed out all of the non-causes for each variable. Once again, consider variables U and V in our described SCM. Given that both these variables take their values independently of each other, by the virtue of faithfulness assumption, we conclude (correctly) that they will not be linked to each other in the underlying DAG. Similarly, looking at Fig. 1.2, we see that W influences Z via a directed path going through X, which is the only causal parent of Z. For this case, by the virtue of the causal Markov condition, we can conclude that once we condition on X, there will no longer be a dependence between Z and W in data. Then there should, at the very least, be a perfect agreement between independences present in data and the separations entailed by the underlying causal network for the same data in a world where sufficiency, CMC, and faithfulness hold.

With assumptions specified, the next step is to develop a test for causality under these proposed assumptions. To do so, we need some way to quantify the agreement between independences in data and the separations in a given candidate graph. We can define a *score* to do exactly that. Without loss of generality, if we can devise a theoretically sound score that is maximized when there is a perfect agreement between independences and separations in data resp. graph, we can perform a search over plausible causal structures to find the true causal network as the one that maximizes this score for given data. As exhaustive search over all possible structures has been shown to be NP-hard (Chickering et al., 2004), our aim is to repeatedly, and systematically, propose plausible causal graphs and score them on given data until we find the one that has the best score. One such consistent score is the Bayesian Information Criterion (Schwarz, 1978, BIC), and one such algorithm that optimizes over BIC to search for a causal network is the Greedy Equivalence Search (Chickering, 2002, GES).

Greedy Equivalence Search (GES) (Chickering, 2002) is a score-based causal discovery approach that learns a causal network G from an observational dataset, using the BIC score. Starting from an empty network, GES iteratively builds

[1]The acyclicity assumption is required when we work with tabular data, but can be relaxed for time-series data. For the scope of this dissertation we work with tabular data and all existing methods we look at, require this assumption

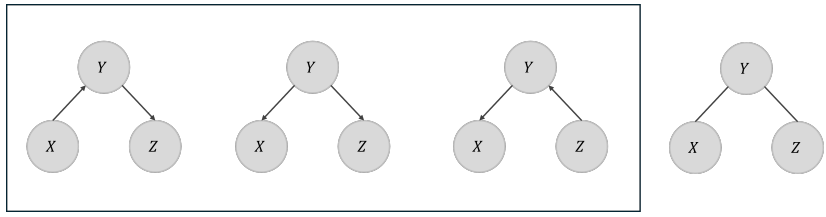


Figure 1.3: [Graphs in the box entail the same set of conditional independences]. All three networks shown within the box above satisfy the independence constraint that X and Z are independent conditioned on Y. Such a set of graphs that entail the same set of conditional independences are said to belong to the same Markov equivalence class. Since none of the edge directions are invariant across all three networks we can not conclude the true edge directions based on independences alone. Hence we can only predict the undirected network as shown on the right.

a causal network through repeated forward respectively backward-search. In each step of the forward search, GES chooses a single edge addition to the current best network such that the edge improves score the most and uses the new network as the best network for the next step. Similarly, in each step of the backward search, single edge deletions that improve score the most are chosen. Each phase ends when no modifications of the current network improve score anymore. This final network is then reported as the causal network.

This sounds good, with the exception of one problem — the mapping between independence constraints and separations in a graph is not one to one. The same set of (conditional) independences can be satisfied by more than one graph. One such example is shown in Fig. 1.3 where X is independent of Z given Y in all three of the graphs shown inside the box. The set of all graphs that satisfy the same set of conditional independences is known as the Markov equivalence class (MEC). This means that a score such as BIC, that only exploits independences, will give the same score to all the graphs within a Markov equivalence class and can not differentiate between members within this MEC. While this property can be used to speed up search within GES, by directly searching over MECs instead of DAGs, it is a downside in terms of the output that GES produces. A different class of approaches known as constraint-based methods, of which the Peter Clark (PC) algorithm (Spirtes et al., 2000a) is an example, suffer from similar limitations. The causal networks that these approaches learn, therefore, contain a number of undirected edges, which is one of many shortcomings of existing approaches.

1.1.2 Existing Shortcomings

Under the assumptions of causal sufficiency, causal Markov condition, and faithfulness specified so far, a partially directed causal network is the best we can get (Pearl, 2009). This means that within a Markov equivalence class we can only be sure of those causal links in the network whose directions are invariant across all graphs, else we can only infer that a causal link exists but can not decide on its direction. Since latter if often the case, we can only obtain causal graphs that are partially oriented as we show in Fig. 1.3. We can not go further unless we place more assumptions on the generating mechanism, such as assuming the causal relationships to be linear functions with additive non-Gaussian noise (Shimizu et al., 2006). Assumptions like these can be confining and would limit the applicability of a causal discovery approach.

The lack of full orientation is only the beginning. Most existing algorithms are limited to finding causal networks over a single dataset where samples are independent and identically distributed (i.i.d.). Such methods, therefore, can only be used to learn networks from individual datasets. This is a problem especially because in real world setting, data is collected across different environments in multiple batches. The only way to make such data compatible with such approaches is to stack it together. This, however, fails because the stacked dataset almost always violates the i.i.d. assumption. This subsequently limits the applicability of causal discovery approaches to real-world scenarios.

Suppose we could, for the sake of only obtaining theoretical guarantees, introduce additional assumptions to accommodate learning from data collected over different sources. All existing algorithms can only work with single, static data and would be infeasible for the setting where data may arrive over time, in chunks. This further reduces the real-world applicability of such approaches.

Alongside the above mentioned limitations, causal discovery approaches are not built for privacy-sensitive applications. These methods lack the mechanism to learn from data distributed across multiple environments, let alone privacy guarantees. Consequently, reliable, scalable causal discovery with the need to protect individual privacy remains a significant challenge. This consequently limits the use of causal discovery approaches in areas where data confidentiality is critical.

The power of causal discovery, while significant, is not fully harnessed by existing methods. This leaves a gap that needs to be filled for these causal discovery approaches to contribute meaningfully to other branches of science. In this dissertation, we make an attempt to do exactly this.

1.2 Research Questions

Following from the discussion in the previous section, the overarching goal of this thesis, in principle, can be summarized in its title:

Practically Applicable Causal Discovery

This aim consists of developing causal discovery approaches that can be applied to real-world scenarios e.g. healthcare, weather forecast, such that the results from these methods can be used by experts to identify causes for phenomena of interest while fulfilling the three-fold desiderata of theoretical correctness, reasonability of assumptions and reliable practical performance.

While the first wave of causal discovery approaches gave us algorithms like Ges (Chickering, 2002) and PC (Spirtes et al., 2000a), they could only discover causal networks from observational data up to Markov equivalence class, rendering them incapable of identifying *all* causal directions for a given variable set. To go beyond partially directed causal networks, approaches like Lingam (Shimizu et al., 2006) and later Resit (Peters et al., 2014) introduced additional assumptions such as linearity of causal relationship resp. independence of residuals to achieve a fully oriented causal network. While sound, these algorithms either work with too strict assumptions or suffer from limitations entailed by independence testing in high dimensions.

Deriving motivation from the above gap, there is a need for developing methods that discover fully oriented causal networks without requiring independence tests involving conditioning on a large number of variables (Peters et al., 2014), or placing assumptions on parametric form of causal relationships (Shimizu et al., 2006). In essence we investigate **how can we discover fully oriented causal networks from observational data?** In Chapter 2 we attempt to fill this gap. We take inspiration from the postulate of Algorithmic Markov Condition (AMC) by Janzing and Schölkopf (2010b). This comes down to discovering fully oriented causal network from observational data by choosing the network that results in the simplest description, measured in bits, of given data. We propose an AMC-based score instantiated using Minimum Description Length (MDL) principle (Grünwald, 2007), with theoretical guarantees. We then present Globe, which is a practical algorithm to discover fully oriented causal networks using our proposed score. Unlike the algorithms that precede it, we show that Globe is hyper-parameter free and is able to beat the state-of-the-art by a clear margin while scaling up to 500 variables.

The next question that follows naturally is **how do we perform causal discovery when data is collected over multiple environments?** For example across different hospitals, each with their own diagnostic device. If, for example, one hospital has a diagnostic device with an (undiscovered) internal anomaly, the data collected there will be from a different, interventional distribution. Pooling data together in such cases, to use with Globe or any other

single-dataset discovery method, can introduce bias in estimation (Lee and Tsui, 1982; Tillman, 2009). Proposals to discover causal networks over multiple such environments are of limited applicability as a number of them focus on a single target variable at a time (Peters et al., 2016; Yu et al., 2019a) and can not trivially be extended to discover causal networks. For existing methods that have been developed, almost all assume the unlikely scenario where we already know which variables (Hauser and Bühlmann, 2012; Triantafillou and Tsamardinos, 2015; Yang et al., 2018), or environments devices(Squires et al., 2020; Brouillard et al., 2020) behave anomalously.

The latter, being yet another challenge, creates the need for a more versatile approach to discover causal networks from datasets collected across multiple environments. In Chapter 3, we turn our focus on doing this. We again define a theoretically sound MDL score for jointly discovering the causal model *and* local interventions, and provide a practical, highly parallelizable, algorithm, ORION, to optimize our proposed score. We explicitly do not assume any prior knowledge of which datasets are observational or interventional and neither assume anything about the functional form of causal relationships. Our extensive evaluation shows that ORION is able to reliably discover causal networks better in such a distributed setting where GLOBE could fail. Keeping in mind the practical aspects, we implement ORION as a highly parallel algorithm.

The versatility of ORION makes it attractive for a number of settings, but limitations still exist in setting where we obtain observations over time. Not only does this mean that we need to learn and update our causal hypothesis over time, but each episode likely contains samples from a specific time period, resulting in a biased distribution. It may not be straightforward to update our knowledge systematically once new data becomes available. Existing algorithms for causal discovery would need to relearn the causal model from scratch whenever a new episode arrives, making them computationally impractical. To make the situation harder, collective data distribution over all episodes may often not be i.i.d since the causal interactions could change over time, leading to the research question: **how do we efficiently discover causal networks from episodic data that arrives over time?** We focus on this in Chapter 4, where we show how we can avoid learning the causal model from scratch upon the arrival of each episode, and can instead learn it in an online fashion using the first-of-its-kind CONTINENT algorithm. Our proposed consistent strategy updates the causal hypothesis as new episodes arrive, using distribution matching and an information-theoretic perspective of causality. CONTINENT not only discovers causal networks reliably from data with episodic selection bias, under interventions, it is the only method, to the best of our knowledge, to learn the causal model adaptively from data arriving over time. This can address a novel experimental setting where different causal networks underlie episodic data and we predict, for a new incoming episode, which causal network it is

generated from without explicitly having to learn a network over the incoming episode.

The inapplicability of causal discovery approaches in privacy preserving settings, such as medical domain, is also one of the fundamental limitations of modern day approaches. In such settings we can neither pool data, nor expect it to arrive over time, in episodes. This introduces its own set of additional challenges and constraints. For privacy sensitive scenarios, we usually have multiple sites each with their own private data. Learning causal networks over such data presents a challenging setting where we do not just want to discover the underlying causal network in a federated manner, we can also not compromise on privacy. This in turn raises the questions as to **how do we perform privacy-preserving federated causal discovery?** In Chapter 5, consequently, we focuses on how we can discover the global causal network without ever sharing any data, model parameters, or even local causal networks— using regrets. Intuitively, the regret measures how much worse a given causal network is, compared to the best causal network for a given dataset. Using regret, we first introduce the RFCD algorithm that can be used to find the underlying causal structure for distributed, private datasets by minimizing over worst-case regret, without theoretical guarantees. We then show that the worst-case regret over these distributed datasets allows us to define a scoring criterion that, under mild assumptions, is guaranteed to be consistent. Using this result we propose a consistent, theoretically sound, and scalable PERI framework. Crucially, we show that using the Laplace mechanism on the shared regrets guarantees ϵ-differential privacy. Keeping true to our goal of practical causal discovery, we show that PERI discovers causal networks of higher quality than the baseline on both synthetic and real-world data, and scales up to 100 distributed environments while requiring orders of magnitude less communication.

1.3 Contributions of this Thesis

This thesis is in large part based on the research articles listed in Table 1.1. To keep this cumulative dissertation coherent, we have modified some content from these research articles and included additional discussions and experiments to connect the dots across different, evolving settings. For the most part, nevertheless, the main content of these articles has been included verbatim. We have removed abstracts, rewritten parts of introduction sections, changed the notation to be consistent across chapters, removed redundancies in related work, and performed additional experiments to reflect on the work in hindsight. For ease of reading, we present our main theorems in each chapter and postpone the proofs to the appendix. For a high level picture note that Chapter 2 acts

Table 1.1: Publications that build up this thesis. Authors with equal contributions are specified using *.

Publication	Used In
Mian, O, Marx, A and Vreeken, J *Discovering Fully Oriented Causal Networks*. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), AAAI, 2021.	Chapter 2
Mian, O, Kamp, M and Vreeken, J *Information-Theoretic Causal Discovery and Intervention Detection over Multiple Environments*. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), AAAI, 2023.	Chapter 3
Mian, O and Mameche, S *An Information Theoretic Framework for Continual Learning of Causal Networks*. In: Proceedings of The Second AAAI Bridge Program on Continual Causality at AAAI Conference on Artificial Intelligence, PMLR, 2024.	Chapter 4
Mian, O*, Mameche, S* and Vreeken, J *Learning Causal Networks from Episodic Data*. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), ACM, 2024.	Chapter 4
Mian, O, Kaltenpoth, D and Kamp, M *Regret-based Federated Causal Discovery*. In: Proceedings of the ACM SIGKDD Workshop on Causal Discovery, PMLR, 2022.	Chapter 5
Mian, O, Kaltenpoth, D, Kamp, M and Vreeken, J *Nothing but Regrets — Privacy-Preserving Federated Causal Discovery*. In: Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS), PMLR, 2023.	Chapter 5

as a preliminary for Chapter 3, whereas Chapters 4 and 5 can be considered two branches stemming from the outcomes we saw in Chapters 2 and 3.

For each of the six research articles listed in Table 1.1, the author of this thesis was the first author. For each first author paper, the author defined the idea, the theory and assumptions to implement the idea, as well as came up with practical instantiaion and conducted experiments for the implemented approach. For Mian et al. (2024), the contributions were split equally with Sarah Mameche who helped extensively in devising the conditions and proofs for the consistency of our proposed approach. For Chapter 2, Alexander Marx helped derive the consistency proof, whereas David Kaltenpoth and Michael Kamp contributed to formalizing the proofs and theoretical guarantees for differential-privacy of the proposed approach for Chapter 5.

In addition to the main chapters included in this work, the author investigated bivariate causal inference for heterogeneous causal relationships titled: *Inferring cause and effect in the presence of heteroscedastic noise* (Xu et al.,

2022), which is yet another practical issue that a number of existing causal inference approaches can not trivially handle. This work, for which the author was the second author, was accepted and published in proceedings of the International Conference on Machine Learning (ICML) in 2022.

Chapter 2

Discovering Fully Oriented Causal Networks

In this chapter, we consider the problem of recovering the causal network over a set of continuous-valued random variables based on an i.i.d. sample from their joint distribution. The state-of-the-art does so by first recovering an undirected causal skeleton—which identifies the variables that have a direct causal relation—and then uses conditional independence tests to orient as many edges as possible. By the nature of these tests this can only be done up to Markov equivalence classes, which means that these methods in practice return networks where a large number of edges are left unoriented. In contrast, we develop an approach that discovers fully directed causal graphs.

We base our approach on the algorithmic Markov condition (AMC), a postulate that states that the factorization of the joint distribution according to true causal network coincides with the one that achieves the lowest Kolmogorov complexity (Janzing and Schölkopf, 2010a). As an example, consider the case where X causes Y. Whereas the traditional *statistical* Markov condition cannot differentiate between $P(X)P(Y|X)$ and $P(Y)P(X|Y)$, as both are valid factorizations of joint distribution $P(X, Y)$, the *algorithmic* Markov condition additionally takes the complexities of these distributions into account: in this case, the simplest factorization of $P(X, Y)$ is $K(P(X)) + K(P(Y|X))$ as only this factorization upholds the true independence between the marginal and conditional distribution—any competing factorization will be more complex because of inherent redundancy between the terms. As Kolmogorov complexity can

This chapter is based on Mian, Marx, and Vreeken (2021).

capture any physical process (Li and Vitányi, 2009) the AMC is a very general model for causality. However, Kolmogorov complexity is not computable, and hence we need a practical score to instantiate it. Throughout this thesis, we do so through the Minimum Description Length principle (Grünwald, 2007), which provides a statistically well-founded approach to approximate Kolmogorov complexity from above.

We develop an MDL-based score for directed acyclic graphs (DAGs), where we model the dependencies between variables through non-parametric multivariate regression. Simply put, the lower the regression error of the discovered model, the lower its cost, while more parameters mean higher complexity. We show this score is consistent: given sufficiently many samples from the joint distribution, we can uniquely identify the true causal graph if the causal relations are either non-invertible or nearly deterministic. To efficiently discover causal networks directly from data we introduce the Globe algorithm, which much like the well-known Ges (Chickering, 2002) algorithm greedily adds and removes edges to optimize the score. Unlike Ges, however, Globe traverses the space of DAGs rather than Markov equivalence classes—orienting edges during its search based on the AMC—and hence is guaranteed to result in a fully directed network.

Through extensive empirical evaluation we show that Globe performs well in practice and outperforms the state-of-the-art conditional independence and score based causal discovery algorithms. On synthetic data we confirm Globe does not discover spurious edges between independent variables, and overall achieves the best scores on both the structural as well as causal similarity metrics. Last, but not least, on real-world data we show that Globe even works well when it is unlikely that our modelling assumptions are met.

This chapter is organized as follows: we first introduce the essential notation in Section 2.1, define the theory behind our approach as well provide identifiability results in Section 2.2, and describe the Globe algorithm in Section 2.3. We discuss the existing approaches in Section 2.4 before providing empirical results in Section 2.5. We conclude after a short discussion in Section 2.6.

2.1 Preliminaries

First, we introduce the notation for causal graphs and the main information theoretic concepts that we need later on.

Causal Graph We consider data over the joint distribution of m continuous valued random variables $\boldsymbol{X} = \{X_1, \ldots, X_m\}$ with $X_i \in \mathbb{R}$. As is common, we assume *causal sufficiency*. That is, we assume that $\boldsymbol{X}$ contains all random variables that are relevant to the system, or in other words, that there exist no

latent confounders. Under the assumptions of causal sufficiency and acyclicity, we can model causal relationships over $\boldsymbol{X}$ using a *directed acyclic graph* (DAG). A causal DAG G over $\boldsymbol{X}$ is a graph in which the random variables are the nodes and edges identify the causal relationship between a pair of nodes. In particular, a directed edge between two nodes $X_i \to X_j$ indicates that X_i is a *direct cause* or *parent* of X_j, and that X_j is a *child* of X_i. We denote the set of all parents and children of X_j with pa_j resp. ch_j.

When working on causal DAGs, we assume the common assumptions, the *causal Markov condition* and the *faithfulness condition*, to hold. Simply put, the combination of both assumptions implies that each separation present in the true graph G implies an independence in the joint distribution P over the random variables $\boldsymbol{X}$ and vice versa (Pearl, 2009).

Identifiability of Causality A causal relationship is said to be *identifiable* if it is possible to unambiguously recover it from observational data alone. In general, causal dependencies are not identifiable without assumptions on the causal model. The common assumptions for discovering causal DAGs allow identification up to the Markov equivalence class (Pearl, 2009). Given additional assumptions, such as that the relation between cause and effect is a non-linear function with additive Gaussian noise (Hoyer et al., 2009b), it is possible to identify causal directions within a Markov equivalence class (Glymour et al., 2019). This is the causal model we investigate.

Kolmogorov Complexity The Kolmogorov complexity of a finite binary string x is the length of the shortest binary program p^* for a universal Turing machine $\mathcal{U}$ that outputs x and then halts (Kolmogorov, 1965; Li and Vitányi, 2009). Formally,

$$K(x) = \min\{|p| \mid p \in \{0,1\}^*, \mathcal{U}(p) = x\} \; .$$

Simply put, p^* is the most succinct *algorithmic* description of x, and therewith Kolmogorov complexity of x is the length of its ultimate lossless compression. Conditional Kolmogorov complexity, $K(x \mid y) \leq K(x)$, is then the length of the shortest binary program p^* that generates x, and halts, given y as input.

The Kolmogorov complexity of a probability distribution P, $K(P)$, is the length of the shortest program that outputs $P(x)$ to precision q on input $\langle x, q \rangle$ (Li and Vitányi, 2009). More formally, we have

$$K(P) = \min \left\{ |p| : p \in \{0,1\}^*, |\mathcal{U}(p, x, q) - P(x)| \leq \frac{1}{q} \right\} .$$

The conditional, $K(P \mid Q)$, is defined similarly except that the universal Turing machine $\mathcal{U}$ now gets the additional information Q. For more details on

Kolmogorov complexity see Li and Vitányi (2009).

Minimum Description Length Principle Although Kolmogorov complexity is not computable, we can approximate it from above through lossless compression (Li and Vitányi, 2009). The Minimum Description Length (MDL) principle (Rissanen, 1978; Grünwald, 2007) provides a statistically well-founded and computable framework to do so. Conceptually, instead of all programs, *ideal MDL* considers only those programs for which we know that they output x and halt, i.e., lossless compressors. Formally, given a model class $\mathcal{M}$, MDL identifies the best model $M \in \mathcal{M}$ for data D as the one minimizing

$$L(D, M) = L(M) + L(D \mid M),$$

where $L(M)$ is the length in bits of the description of M, and $L(D \mid M)$ is the length in bits of the description of data D given M. This is known as two-part, or *crude* MDL. There also exists one-part, or *refined* MDL. Although refined MDL has theoretically appealing properties, it is efficiently computable for a small number of model classes. Asymptotically however, there is no difference between the two (Grünwald, 2007).

To use MDL in practice we need to define a model class, and how to encode a model, resp. the data given a model, into bits. Note that we are only concerned with optimal code *lengths*, not actual codes—our goal is to measure the *complexity* of a dataset under a model class, after all (Grünwald, 2007). Hence, all logarithms are to base 2, and we use the common convention that $0 \log 0 = 0$.

2.2 Theory

In this section, we will first introduce the algorithmic model of causality which is based on Kolmogorov complexity. To put it into practice, we need to introduce a set of modelling assumptions that allow us to approximate it using MDL. We conclude this section by providing consistency guarantees.

2.2.1 Algorithmic Model of Causality

Here we introduce the main concepts of algorithmic causal inference as introduced by Janzing and Schölkopf (2010a), starting with the causal model.

Postulate 2.1 (Algorithmic Model of Causality) *Let G be a DAG formalizing the causal structure among the strings $x_1, \ldots, x_m$. Then, every x_j*

is computed by a program q_j with constant length from its parents pa_j and an additional input n_j. That is

$$x_j = q_j(pa_j, n_j) \ ,$$

where the inputs n_j are jointly independent.

As any mathematical object x can be described as a binary string, and a program q_j can model any physical process (Deutsch, 1985) or possible function h_j (Li and Vitányi, 2009), this is a particularly general model of causality. Equivalent to the statistical model, we can derive that the algorithmic model of causality fulfils the *algorithmic* Markov property (Janzing and Schölkopf, 2010a), that is

$$K(x_1, \ldots, x_m) \stackrel{+}{=} \sum_{j=1}^{m} K(x_j \mid \mathrm{pa}_j^*) \ ,$$

where $\stackrel{+}{=}$ denotes equality up to an additive constant. Meaning, to most succinctly describe all strings, it suffices to know what are the parents and additional inputs n_j for each string x_j. Unlike its statistical counterpart which can only identify the causal network up to Markov equivalence, the *algorithmic* Markov property can identify a single DAG as the most succinct description of all strings. As any mathematical object, including distributions, can be described by a binary string, Janzing and Schölkopf (2010a) define the following postulate.

Postulate 2.2 (Algorithmic Markov Condition) *A causal DAG G over random variables $\boldsymbol{X}$ with joint density P is only acceptable if the shortest description of P factorizes as*

$$K(P(X_1, \ldots, X_m)) \stackrel{+}{=} \sum_{j=1}^{m} K(P(X_j \mid pa_j)) \ . \tag{2.1}$$

Hence, under the assumption that the true causal graph can be modelled by a DAG, it has to be the one minimizing Eq. (2.1). As K is not computable we cannot directly compute this score. What we can however, restrict our model class from allowing all possible functions to a subset of these and then approximate K from above using MDL.

2.2.2 Causal Model

As causal model we consider a rich class of structural equation models (Pearl, 2009) (SEMs) where the value of each node is determined by a linear combina-

tion of functions over all possible subsets of parents and additional independent noise. Formally, for all $X_i \in \boldsymbol{X}$ we have

$$X_i := \sum_{\mathcal{S}_j \in \mathcal{P}(\mathrm{pa}_i)} \beta_j \cdot h_j(\mathcal{S}_j) + N_i \ , \tag{2.2}$$

where h_j is a non-linear function of the j-th subset over the power set, $\mathcal{P}(\mathrm{pa}_i)$, of parents of X_i, and N_i is an independent noise term. Without loss of generality, any function defined over parents of an effect can be expressed this way. We assume that all noise variables are jointly independent, Gaussian distributed and that $N_i \perp\!\!\!\perp \mathrm{pa}_i$. Naturally, we do not expect all subsets over parents to be part of SEM, which would simply evaluate to $\beta_j = 0$ for that subset. An advantage of defining the causal model this way is that we can also model interactions between parents. Note that if $\beta_j \neq 0$ only for parent subsets of size 1, Eq. (2.2) simplifies to an additive model over individual parents.

2.2.3 MDL Encoding of the Causal Model

Next, we specify our MDL score for DAGs. Given an iid sample $\boldsymbol{X}^n$, containing n rows, drawn from the joint distribution P over $\boldsymbol{X}$, our goal is to approximate Eq. (2.1) using two-part MDL, which means we need to define a model class $\mathcal{M}$ for which we can compute the optimal code length. Here, we define $\mathcal{M}$ to include all possible DAGs over $\boldsymbol{X}$ and their corresponding parametrization according to our causal model. That is, for each node X_i a model $M \in \mathcal{M}$ contains an index indicating the parents of X_i (which is equivalent to storing the DAG structure), and the corresponding functional dependencies.

Building upon Eq. (2.1), we want to find that model $M^* \in \mathcal{M}$ such that

$$\begin{aligned} M^* &= \operatorname*{argmin}_{M \in \mathcal{M}} L(\boldsymbol{X}^n, M) \\ &= \operatorname*{argmin}_{M \in \mathcal{M}} \left(L(M) + \sum_{i=1}^{m} L(X_i^n \mid \mathrm{pa}_i, M) \right) \\ &= \operatorname*{argmin}_{M \in \mathcal{M}} \left(L(M) + \sum_{i=1}^{m} L(\epsilon_i) \right) \end{aligned}$$

where pa_i are the parents of X_i according to the model M. In the last line, we replace $L(X_i^n \mid \mathrm{pa}_i, M)$ with $L(\epsilon_i)$ to clarify that for continuous-valued data, encoding a node given M and its parents comes down to encoding the residuals ϵ_i.

Encoding the Model

The model complexity $L(M)$ for a model $M \in \mathcal{M}$, comprises of the parameters of the functional dependencies and the graph structure. The total cost is simply the sum of the code lengths of the individual nodes

$$L(M) = \sum_{i=1}^{m} L(M_i) \ .$$

To encode the individual nodes X_i, we need to transmit its parents, the form of the functional dependency, and the bias or mean shift μ_i. We encode the model M_i for a node X_i as

$$L(M_i) = L_{\mathbb{N}}(k) + k \log m + L_F(f_i) \ ,$$

where we first encode the number of parents using $L_{\mathbb{N}}$, the MDL-optimal encoding for integers $z \geq 0$ (Rissanen, 1983). It is defined as $L_{\mathbb{N}}(z) = \log^* z + \log c_0$, where $\log^* z = \log z + \log\log z + \ldots$ and we consider only the positive terms, and c_0 is a normalization constant to ensure the Krafft-inequality holds (Krafft, 1949). Next, we identify which out of the m random variables these are, and then proceed to encode the function f_i over these parents, where f_i represents the summation term on the right hand side of Eq. (2.2).

Encoding the Functions We will instantiate the framework using non-parametric functions h_i that also allow for non-linear transformations of the parent variables. To this end, we fit non-parametric Multivariate Adaptive Regression Splines (Friedman, 1991). In essence, we estimate X_i as

$$\hat{X}_i := \sum_{j=1}^{|H|} h_j(\mathcal{S}_j) \ ,$$

where h_j is called a hinge function that is applied to a subset of the parents, $\mathcal{S}_j$, with size $|\mathcal{S}_j|$, that is associated with the j-th hinge. A hinge takes the form

$$h(\mathcal{S}) = \prod_{i=1}^{T} a_i \cdot \max(0, g_i(s_i) - b_i) \ ,$$

where T denotes the number of multiplicative terms in h, $s_i \in \mathcal{S}$ is the parent associated with the i-th term, g_i is a non-linear transformation applied to s_i where g_i belongs to the function class $\mathcal{F}$, e.g. the class of all polynomials up to a certain degree, but the encoding can be very general and can include any regression function as long as we can describe the parameters and $|\mathcal{F}| < \infty$. If

$T = 1$ for all hinges, the above definition simplifies to an additive model over individual parents. We encode a hinge function as follows

$$L_F(h) = L_{\mathbb{N}}(|H|) + \sum_{h_j \in H} \left[L_{\mathbb{N}}(T_j) + \log \binom{|\mathcal{S}| + T_j - 1}{T_j} + T_j \log(|\mathcal{F}|) + L_p(\theta(h_j)) \right]$$

First, we use $L_{\mathbb{N}}$ to encode the number of hinges and the number of terms per hinge. We then transmit the correct assignment of terms T_j to parents in $\mathcal{S}$, and finally need $\log(|\mathcal{F}|)$ bits to identify the specific non-linear transformation that is used for each of the T_j terms in the hinge.

Encoding Parameters To encode the bias, as well as the set of parameters associated with function f_i in L_F, we use the proposal of Marx and Vreeken (2017) for encoding parameters up to a user specified precision p. We have

$$L_p(\theta) = |\theta| + \sum_{i=1}^{|\theta|} L_{\mathbb{N}}(s_i) + L_{\mathbb{N}}(\lceil \theta_i \cdot 10^{s_i} \rceil) \ ,$$

where s_i is the smallest integer such that $|\theta_i| \cdot 10^{s_i} \geq 10^p$. Simply put, $p = 2$ implies that we consider two digits of the parameter. We need one bit to store the sign, then we encode the shift s_i and the shifted parameter θ_i.

Encoding Residuals

Last, we need to encode the residual term, $L(\epsilon_i)$. Since we use regression functions, we aim to minimize variance of the residual—and hence should encode the residual ϵ as Gaussian distributed with zero-mean (Marx and Vreeken, 2017; Grünwald, 2007)

$$L(\epsilon) = \frac{n}{2} \left(\frac{1}{\ln 2} + \log 2\pi\hat{\sigma}^2 \right) \ ,$$

where we can compute the empirical variance $\hat{\sigma}^2$ from ϵ.

Combining the above, we now have a lossless MDL score for a DAG.

2.2.4 Consistency

Since MDL can only upper bound Kolmogorov complexity, but not compute it, it is not possible to directly derive strict guarantees from the AMC. We can, however, derive consistency results. We first show that our score allows for identifying the Markov equivalence class of the true DAG i.e. the partially

directed network for which each collider is correctly identified. Then, we show that under additional assumptions, we can orient the remaining edges correctly.

The main idea for the first part is to show that our score is consistent—simply put, *the likelihood term dominates in the limit.* For a score with such properties e.g. BIC (Haughton, 1988), Chickering (2002) showed that it is possible to identify the Markov equivalence class of the true DAG. To show that our score behaves in the same way, we need to make two additional assumptions for $n \to \infty$:

1. the number of hinges of $|H|$ is bounded by $\mathcal{O}(\log n)$, and
2. the precision of the parameters θ is constant w.r.t. to n and hence $L_p(\theta) \in \mathcal{O}(1)$.

Based on these assumptions, we can show that our score is consistent as it asymptotically behaves like BIC, meaning that the penalty term for the parameters only grows with $\mathcal{O}(\log n)$ complexity, while the likelihood term grows linearly with n and hence is the dominating term as $n \to \infty$.

Theorem 2.1 *Given a causal model as defined in Eq.* (2.2) *and corresponding data* $\boldsymbol{X}^n$ *drawn iid from joint distribution* P*. Under Assumptions (1) and (2),* $L(\boldsymbol{X}^n, M)$ *asymptotically behaves like BIC.*

With the above, we know that given sufficient data our score will identify the correct Markov equivalence class.

To infer the complete DAG, we need to be able to infer the direction for those edges that cannot be inferred using collider structures—i.e. single edges like $X - Y$. Closest to our approach is the work of Marx and Vreeken (2019) who showed that it is possible to distinguish between $X \to Y$ and $Y \to X$ using any L_0 regularized score—e.g. BIC, if we assume that the underlying causal function is near deterministic i.e. $Y := f(X) + \alpha N$, where f is a non-linear function and N is an unbiased, unit-variance noise regulated by a small constant $\alpha > 0$, and that $\alpha \to 0$. Since our score in the limit behaves like an L_0-based score (ref. Theorem 2.1), we can distinguish between Markov equivalent DAGs under this additional assumption. As an alternate to using the low-noise assumption, our guarantees would also hold if we assume that the function $f(X)$ is a non-invertible function, and therefore modeling the anti-causal relationship would incur a loss of information and result in a lower score than modeling the causal direction (Hoyer et al., 2009a). We use the low-noise assumption because it is more general in that it also covers the case of non-invertible functions.

Although our score is consistent and can be used to distinguish Markov equivalent DAGs, these guarantees only hold if we were to score all DAGs over $\boldsymbol{X}$. Since this is infeasible for large graphs, we propose a modified greedy DAG search algorithm to minimize $L(\boldsymbol{X}^n, M)$.

Algorithm 2.1: The Globe Algorithm

Data: Data $\boldsymbol{X}^n$ over $\boldsymbol{X}$
Result: Causal DAG G
1 $Q \leftarrow \text{EdgeScoring}(\boldsymbol{X}^n)$
2 $G \leftarrow \text{ForwardSearch}(Q, \boldsymbol{X}^n)$
3 $G \leftarrow \text{BackwardSearch}(G)$
4 *return* G

2.3 The Globe Algorithm

We now present Globe, a score-based method for discovering directed acyclic causal graphs from multivariate continuous valued data. Globe consists of three steps: edge scoring, forward and backward search, as shown in Algorithm 2.1. We subsequently describe these steps and provide pseudocode for them.

Edge Scoring To improve the forward search where we greedily add the edge that provides the highest gain, we first order all potential edges in a priority queue by their causal strength. We measure the causal strength of an edge, using the absolute gain in bits for orienting an edge in either direction in our model. Formally, let $e = (X_i, X_j)$ be an undirected edge between X_i and X_j, and further let $\vec{e}$ refer to the directed edge $X_i \to X_j$ and $\overleftarrow{e}$ the directed edge in the reverse direction. Now, let M be the current model. We write $M \oplus \overleftarrow{e}$ to refer to the model where we add edge $\overleftarrow{e}$, and $M \oplus \vec{e}$ for the model where we add $\vec{e}$. We define the gain in bits, δ, associated with edge $\overleftarrow{e}$ as

$$\delta(\ \overleftarrow{e}\ ) = \max\{0, L(\boldsymbol{X}^n, M) - L(\boldsymbol{X}^n, M \oplus \overleftarrow{e})\}$$

where $L(\boldsymbol{X}^n, M)$ is defined according to the causal model specified in the theory section, and define $\delta(\ \vec{e}\ )$ analogously. Based on $\delta(\ \overleftarrow{e}\ )$ and $\delta(\ \vec{e}\ )$, we define the directed gain $\Psi(\overleftarrow{e})$ for a given edge as

$$\Psi(\overleftarrow{e}) = \delta(\overleftarrow{e}) - \delta(\vec{e})\ ,$$

where $\Psi(\overleftarrow{e}) = -\Psi(\vec{e})$. The higher the value of $\Psi(\overleftarrow{e})$, the higher edge $\overleftarrow{e}$ is ranked. Intuitively, the larger the difference between the edge direction, the more certain we are that we inferred the correct direction. The algorithm for this step is straightforward and is shown in Alg. 2.2 — we pick each undirected edge e, calculate δ and Ψ (line 3) for $\overleftarrow{e}$ and $\vec{e}$, and add the edges to a priority queue (lines 4-5). We return this priority queue as the output of this step once

Algorithm 2.2: Edge Scoring

Data: **n** samples over $\boldsymbol{X}$
Result: priority queue of edges Q

```
1 Q ← ∅
2 foreach pair (u, v) ∈ X do
3     ψ ← δ(e_uv) − δ(e_vu)
4     Q ← Q ⊕ (e_uv, ψ)
5     Q ← Q ⊕ (e_vu, −ψ)
6 return Q
```

all edges have been ranked (line 6).

FORWARD SEARCH For forward search phase shown in Algorithm 2.3, we use the priority queue obtained from the edge ranking step to build the causal graph by iteratively adding the highest ranked edge (line 4,6). We reject edges that would introduce a cycle (line 5). After adding an edge $X_i \to X_j$ we need to update the score of all edges pointing towards X_j and re-rank them in the priority queue (lines 7-10). Due to the greedy nature of the algorithm, we may add edges in the wrong direction when we do not yet know all the parents of a node. Hence, after adding edge $X_i \to X_j$ to the current model—i.e. discovering a new parent for X_j—we check for all children of X_j, whether flipping the direction of the edge improves the overall score (lines 11-13). If so, we delete that edge $\vec{e}_{ji}$ from our model (line 15), re-calculate δ and Ψ for $\vec{e}_{ji}$ and $\overleftarrow{e}_{ij}$ (line 16), and push them again to the priority queue (line 18) (see Fig. 2.1 for an example scenario). We follow this up by again updating the score of all existing potential parents for the child node whose edge was removed (lines 19-22). The forward search stops when the priority queue is empty.

To avoid spurious edges, we check for significance of the gain. Let $k = \delta(\overleftarrow{e})$, based on the no-hypercompression inequality (Grünwald, 2007), the probability

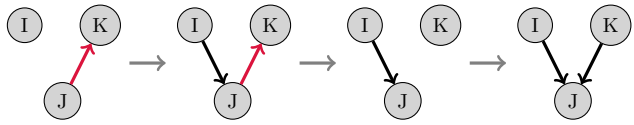


Figure 2.1: Edge reversal in the forward search: We start with the graph where we wrongly added edge $X_j \to X_k$, then we add the correct edge $X_i \to X_j$. Revisiting the children of X_j we see that flipping $X_j \to X_k$ improves our score and hence delete the edge. In the next step we add the correct edge.

Algorithm 2.3: Forward Search

Data: priority queue of edges Q, $\mathbf{n}$ samples over $\boldsymbol{X}$
Result: graph G

1 $E \leftarrow \emptyset$
2 $G \leftarrow (\boldsymbol{X}, E)$
3 **while** Q **not** *empty* **do**
4 $\quad e_{uv} \leftarrow$ *take top most entry from* Q
5 $\quad$ **if** $E \oplus e_{uv}$ *is not cyclic* **and** e_{uv} *is significant* **then**
6 $\quad\quad E \leftarrow E \oplus e_{uv}$
7 $\quad\quad$ **foreach** *incoming edge to* v, $e_{xv} \in Q$ **do**
8 $\quad\quad\quad \psi \leftarrow \delta(e_{xv}) - \delta(e_{vx})$
9 $\quad\quad\quad$ *update the value of* e_{xv} *in* Q *to* ψ
10 $\quad\quad\quad$ *update the value of* e_{vx} *in* Q *to* $-\psi$
11 $\quad\quad$ **foreach** *outgoing edge from* v, $e_{vy} \in E$ **do**
12 $\quad\quad\quad E' \leftarrow (E \ominus e_{vy}) \oplus e_{yv}$
13 $\quad\quad\quad G' \leftarrow (\boldsymbol{X}, E')$
14 $\quad\quad\quad$ **if** $Cost(G') < Cost(G)$ **then**
15 $\quad\quad\quad\quad E \leftarrow E \ominus e_{vy}$
16 $\quad\quad\quad\quad \psi \leftarrow \delta(e_{vy}) - \delta(e_{yv})$
17 $\quad\quad\quad\quad Q \leftarrow Q \oplus (e_{vy}, \psi)$
18 $\quad\quad\quad\quad$ *update the value of* e_{yv} *in* Q *to* $-\psi$
19 $\quad\quad\quad\quad$ **foreach** *incoming edge to* y, $e_{ky} \in Q$ **do**
20 $\quad\quad\quad\quad\quad \psi \leftarrow \delta(e_{ky}) - \delta(e_{yk})$
21 $\quad\quad\quad\quad\quad$ *update value of* e_{ky} *in* Q *to* ψ
22 $\quad\quad\quad\quad\quad$ *update value of* e_{yk} *in* Q *to* $-\psi$
23 *return* G

to gain k bits over the null model is smaller or equal to 2^{-k}. If for an edge the gain k is not significant—i.e. $2^{-k} > \alpha$, where α is a user defined significance threshold, we disregard the edge (line 5).

Backward Search To further refine the graph discovered in the forward search, we iteratively remove superfluous edges using the backward search procedure shown in Alg. 2.4. In particular, for each node X_j with $|\mathrm{pa}(X_j)| = k \geq 2$ we score all graphs for which we only use a subset of the parents of size $k - 1$

Algorithm 2.4: Backward Search

```
   Data: graph G
   Result: pruned graph G
1  foreach node v ∈ G do
2      while node updated and (|pa(v)| ≥ 2) do
3          (p, c) ← (pa(v) , Cost(v | pa(v) ))
4          foreach p′ ⊂ pa(v) s.t |p′| = |pa(x)| − 1 do
5              c′ ← Cost(v | p′)
6              if c′ < c then
7                  (p, c) ← (p′, c′)
8          pa(v) ← p
9  return G
```

(lines 4-5). If any of these graphs provides a gain in compression, we select the one that provides the largest gain and update the model accordingly (lines 6-8). We continue this process until we cannot find such a subset for any node and output the current graph as our predicted causal DAG (line 9).

2.3.1 Complexity Analysis

The edge ranking does one pass over the edges, it has a runtime of $\mathcal{O}(m^2)$. In the forward search, each edge can lead to at most $(m-1)$ ranking updates due to edge flips. Resulting in a total complexity in $\mathcal{O}(m^3)$. The backwards search has a loose upper bound of $\mathcal{O}(m^3)$, that results when the forward search returns a fully connected graph and we delete each of those edges in the backwards search. Hence, the overall complexity of Globe is in $\mathcal{O}(c(n) \cdot m^3 \cdot \log m)$, where $c(n)$ denotes the complexity of the regression approach used, and $\log m$ is the time complexity of updating the edge priority queue after each step. In practice, Globe is fast enough for networks as large as 500 nodes.

2.3.2 Instantiation

We instantiate Globe[1] using the open-source implementation in R of Multivariate Adaptive Regression Splines framework (Friedman, 1991). Since we

[1] Globe stems from discovering fully, rather than locally, oriented networks, as well as from it being based on Multivariate Adaptive Regression Splines (Mars), of which the public implementation is known as Earth.

could face issues like multi-collinearity (Farrar and Glauber, 1967) and unrealistic run times if we allow for arbitrary many interactions between parents, we restrict the maximum number of interaction terms to 2 for experiments.

2.4 Related Work

Causal discovery on observational data has drawn more attention in recent years (Bühlmann et al., 2014; Huang et al., 2018; Hu et al., 2018; Margaritis and Thrun, 2000) and remains an open problem. To give a succinct overview, we focus on the most related methods, ones that aim to recover a DAG or its Markov equivalence class from continuous valued data. We exclude methods that aim at weakening assumptions such as causal sufficiency or acyclicity (Spirtes et al., 2000b) as they do not learn a directed acyclic graph.

Most approaches can be classified as constraint based or score based. Both rely on the Markov and faithfulness conditions to recover Markov equivalence classes of the true DAG. Constraint based methods such as the PC and FCI algorithm (Spirtes et al., 2000b), their extensions (Colombo and Maathuis, 2014; Pearl et al., 1991) as well as the Grow-Shrink algorithm (Margaritis and Thrun, 2000) rely on conditional independence (CI) tests to first recover the undirected causal graph and then infer edge directions only up to the Markov equivalence class using additional edge orientation rules (Meek, 1995). The main bottleneck for those approaches is the CI test. The standard choice is the Gaussian CI test (Kalisch and Bühlmann, 2007). However, it cannot capture non-linear correlations. The current state-of-the-art uses kernel based tests such as HSIC (Gretton et al., 2005), which can capture non-linear dependencies.

Score based methods define a scoring function, $S(G, \boldsymbol{X}^n)$, that evaluates how well a causal DAG G fits the provided data $\boldsymbol{X}^n$. If the true causal graph G^* is a DAG, then given infinite data the highest scoring DAG is part of the equivalence class of G^* (Chickering, 2002). Score based approaches start with an empty graph and greedily traverse to the highest scoring Markov equivalence class that is reachable by adding, deleting or reversing an edge. Well-known algorithms in this category include the greedy equivalence search (Ges) (Chickering, 2002; Hauser and Bühlmann, 2012), its extensions (Ramsey et al., 2017), and the current state-of-the-art, generalized-Ges (GGes) (Huang et al., 2018) which uses kernel regression to capture complex dependencies.

In contrast, additive noise models (ANMs) aim to discover the fully directed graph (Hoyer et al., 2009b). The primary assumption is that the effect can be written as a function of the cause plus additive noise that is independent of the cause. Under this assumption, the function is only admissible in causal direction and not vice-versa (Hoyer et al., 2009b). Methods range from linear non-Gaussian (Lingam) (Shimizu et al., 2006), non-linear functions (Resit) (Peters et al., 2014) to mixtures of non-linear additive noise models (Hu et al., 2018).

The main caveat of ANMs is also the CI test. Fitting a non-linear function that maximizes the independence between the cause and noise is a slow process which restricts ANMs application to small networks (Hoyer et al., 2009b).

Most related to our work are methods based on regression error. Those methods have been shown to successfully decide between Markov equivalent DAGs under the assumption of having a non-linear function and low noise (Marx and Vreeken, 2017; Blöbaum et al., 2018b; Marx and Vreeken, 2019) or proven to correctly identify the causal ordering of all nodes (CAM) (Bühlmann et al., 2014). Directly comparing a causal ordering to a DAG is, however, not straightforward.

In this chapter, we combine the advantages of score based methods and methods based on regression error by discovering the fully oriented graph and allowing for complex non-linear dependencies, while being fast in practice.

2.5 EVALUATION

We evaluate GLOBE on both synthetic and real-world data with known ground truth. GLOBE is implemented in Python and both the source code, as well as the synthetic data are made available for reproducibility[2]. We compare GLOBE to the state-of-the-art from different classes of algorithms. We compare to RESIT (Peters et al., 2014) and LINGAM (Shimizu et al., 2006) as representative ANM-based methods, to GGES as the best score-based method (Huang et al., 2018), and to PC with the Hilbert Schmidt Independence Criteria, short PC_{HSIC} (Colombo and Maathuis, 2014; Gretton et al., 2005), as the state-of-the-art constraint-based method for causal discovery. Comparison with GES (Chickering, 2002; Ramsey et al., 2017) is ommitted since its performance was significantly worse than the other methods. We provide details on experimental setup as well as a case-study for our evaluations. GLOBE finished within ten minutes for each experimental instance except one pseudo-real-world dataset with 500 nodes, on which it took 3 days, whereas no other competitor was able to handle this data.

EVALUATION METRICS We evaluate the predicted and the ground truth graphs on the basis of their structural, as well as their *causal* similarity.

The Structural Hamming Distance (*SHD*) (Kalisch and Bühlmann, 2007), between two partially directed acyclic graphs (PDAGs) G and $\hat{G}$ is the the total number of edges where the two graphs differ. Denoting the edge adjacency matrix of G and $\hat{G}$ with X resp. $\hat{X}$ we have

[2] `https://eda.rg.cispa.io/globe`

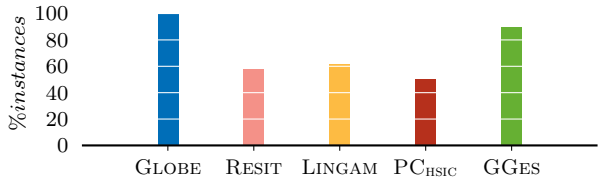


Figure 2.2: [Higher is better] Percentage of instances with no spurious edges reported for the independent data. GLOBE achieves a perfect score.

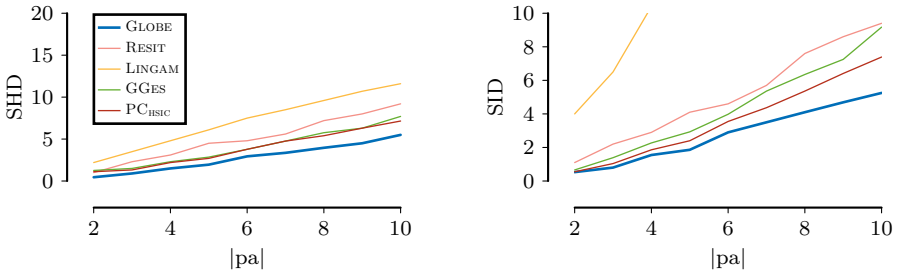


Figure 2.3: [Lower is better] *SHD* (left) and *SID* (right) for increasing number of parents.

$$SHD(G, \hat{G}) \coloneqq \sum_{1 \leq i < j \leq m} \mathbf{I}((X_{ij} \oplus \hat{X}_{ij}) \vee (X_{ji} \oplus \hat{X}_{ji})) \ ,$$

where $\oplus$ denotes an XOR operation and $\mathbf{I}(x)$ is 1 when the expression x is *true* and 0 otherwise.

However, *SHD* tells us nothing about the causal similarity between two graphs. Hence, we use the Structural Intervention Distance (*SID*) (Peters and Bühlmann, 2015) pre-metric. *SID* counts the pairs of nodes u and v such that the effect of intervention from u to v is falsely estimated by $\hat{G}$ with respect to G. In case a method outputs only the Markov equivalence class, *SID* is an interval, with smallest and largest scores indicating the best resp. worst scores for the DAGs in the given Markov equivalence class. For more details on *SID*, see Peters and Bühlmann (2015).

2.5.1 Synthetic Data

We start with a sanity check to ensure that GLOBE can reliably avoid false positives and build up to the case of varying sample sizes over a more complex network. We generated 100 instances each with 1 000 observations for the

n	GLOBE	RESIT	LINGAM	GGES	PC_{HSIC}
100	**0.28**	0.45	0.47	[0.18 , 0.48]	[0.28 , 0.54]
500	**0.26**	0.43	0.43	[0.17 , 0.48]	[0.21 , 0.55]
1000	**0.26**	0.42	0.42	[0.17 , 0.48]	[0.20 , 0.54]
1500	**0.27**	0.40	0.43	[0.17 , 0.48]	[0.19 , 0.53]
2000	**0.26**	0.40	0.40	[0.18 , 0.49]	[0.19 , 0.54]

Table 2.1: [Lower is Better] Averaged normalized *SID* for the methods. Interval for GGES and PC_{HSIC} indicates the best, resp. worst possible intervention distance for the DAGs in the discovered Markov equivalence class.

discussed structures, unless stated otherwise. We standardized the data to have zero mean and unit variance.

INDEPENDENT DATA As a sanity check, we test the methods on instances of a graph containing 10 independent nodes where the value of each node is sampled independently from a Gaussian distribution. We expect all the methods to report empty sets of edges for the instances in this experiment. GLOBE did not report a single spurious edge on *any* of the instances. On the other hand, LINGAM reported at least one spurious edge for 38%, RESIT for 42% and PC_{HSIC} and GGES for half resp. 10% of the instances.

EFFECT OF MULTIPLE PARENTS Next we test GLOBE on a simple case of a collider where we vary the number of parents from 2 up to 10. The collider node is calculated as a linear combination of non-linear parent functions given as

$$X_j = \sum_{X_i \in \text{pa}(X_j)} a_i \cdot (X_i + b_i)^{c_i} \quad . \tag{2.3}$$

Since it is possible to identify a collider structure using conditional independence tests, we expect GGES and PC_{HSIC} to discover a fully directed network. The results for both *SHD* and *SID* are shown in Figure 2.3. In case of *SID*, we compare favorably to both GGES and PC_{HSIC} by only reporting the *best possible* achievable score for their predicted graphs' Markov equivalence class. Even with this favorable comparison, GLOBE outperforms the competition.

DATA SAMPLED FROM A CAUSAL NETWORK Next, we show GLOBE's effectiveness in finding the causal relationships in a more general setting. Similar to Ghanbari et al. (2018), we consider multiple instances of a graph that contains all possible connections that could exist in a DAG. In this setting, each

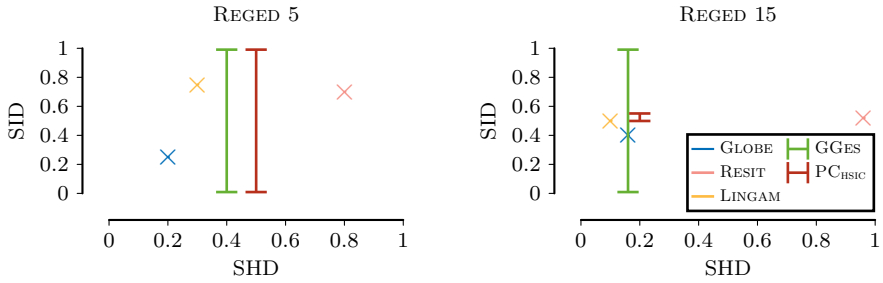


Figure 2.4: [Closer to Origin is Better] Comparison of Normalized *SHD* and Normalized *SID* for real world networks.

child node, X_j can alternatively be calculated using more complex multiplicative interactions between the parents given by

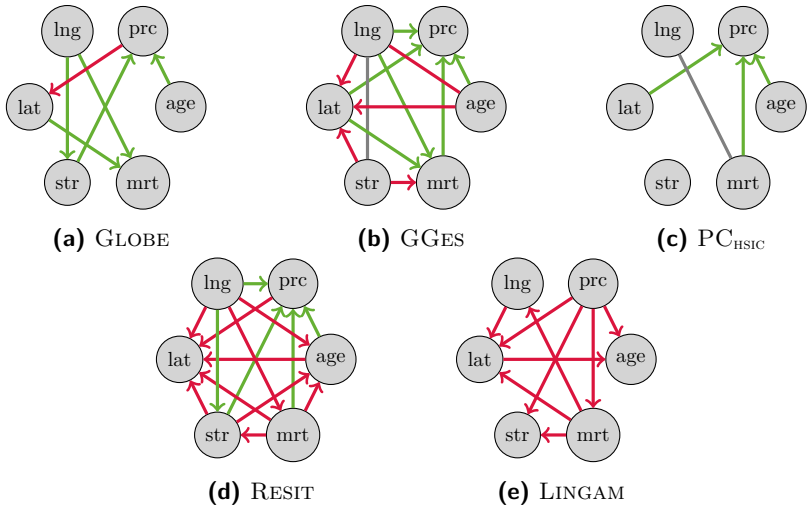
$$X_j = a_j \cdot \prod_{X_i \in \mathrm{pa}(X_j)} {X_i}^{c_i} + b_j \quad . \tag{2.4}$$

We generate data where we choose between Eq. (2.3) and (2.4) with probability 0.7 resp. 0.3 and report results over varying sample sizes. We report the values for *SID* in Table 2.1. Overall we see that Globe outperforms Resit and Lingam. The causal networks predicted by Globe have *SID* closer to the better end of the range of scores possible for $\mathrm{PC}_{\mathrm{HSIC}}$ and GGes. In terms of *SHD*, all the methods were found to be consistent over varying sample sizes, with Globe slightly outperforming the competition.

2.5.2 Real World Data

For real world data with known ground truth, we consider three distinct networks of sizes 5, 15 and 500 nodes from the reged dataset (Statnikov et al., 2015), each containing 1 000 rows. Looking at the results shown in Figure 2.4, we see that Globe is closest to the true causal network for both the 5 node (Reged 5) and the 15 node (Reged 15) network. For Reged 15, Globe reports a better *SID* than all the competitors. We see that for the Reged 15 network, GGes fails to orient most of the edges, which results in a graph where both extremes of the *SID* are possible.

For the 500 node network, Globe was the *only* algorithm to produce any kind of result in reasonable time (3 days), with a reported normalized *SID* and *SHD* of 0.1 resp. 0.01. While GGes failed to terminate within one month, all other methods could not process the data.



(a) GLOBE **(b)** GGES **(c)** PC_{HSIC}

(d) RESIT **(e)** LINGAM

Figure 2.5: Discovered DAGs on the real estate data set. Green edges are causal directions that agree with our domain knowledge, directed red edges are wrongly oriented causal dependency. Gray edges are associations that agree with domain knowledge but are left unoriented. Undirected red edges are associations that disagree with domain knowledge.

2.5.3 Case Study: Real Estate Data

To conclude our evaluation, we perform a case study on a real estate dataset (Yeh and Hsu, 2018; Dua and Graff, 2017) of market valuation of properties in the Sindian district, Taiwan. The data contains six continuous valued attributes: the age of the property (*age*), the distance to the nearest MRT station (*mrt*), the number of convenience stores reachable by foot from the house (*str*), the geographical coordinates (*lat*,*lng*) and the price of the property (*prc*). After additionally normalizing the data between zero and one, we run all the methods on this data and report the results in Figure 2.5. Overall, we see that the causal dependencies GLOBE discovers are in accordance with our domain knowledge: it finds that by changing the coordinates of the property (*lat*,*lng*), we can alter the distance to the train station (*mrt*) and that the latitude of the property determines the number of nearby stores. GLOBE also discovers three possible causal relations to the price of the property namely *age*, *str* and *lat*. However it wrongly orients the direction of price to latitude.

The other methods perform less well. We see that they either orient most edges against domain knowledge (RESIT, LINGAM), discover spurious causal relations (GGES), or report meaningful edges but only for a single variable (PC_{HSIC}).

2.6 Discussion

We considered discovering fully directed causal graphs from observational data. To tackle this problem, we built upon the algorithmic Markov condition and proposed a score based on MDL to approximate it from above. For non-linear mixture models with (low) additive noise assumption, our score allows for discovering the fully directed causal graph. To minimize the proposed score, we instantiate Globe, a greedy DAG search algorithm that iteratively builds a DAG to find a locally optimal solution while modeling functional dependencies using non-parametric regression functions. One elegance of Globe lies in its hyper-parameter-free nature — it is a straightforward end to end causal discovery algorithm. The user provides a dataset and gets a fully oriented causal network in return without the need to specify a hyperparameter a priori. From a user standpoint, the only thing that needs to be factored in is the significance threshold α that we use for the no-hypercompression inequality. The latter, nonetheless, only serves as a cut off point to discard insignificant edges, and does not result in drastic variance across different experiments over same dataset. This makes Globe a straightforward algorithm to use for causal discovery.

Throughout this chapter, we have build Globe on the assumption of causal sufficiency. While this assumption might holds for a number of controlled scenarios involving robotics and/or reinforcement learning, it is violated in various real-world settings as there might be unobserved confounders that explain two related variables within the data. Adding such a confounder induced edge to greedy DAG search can mislead the search procedure. Practically, we have tried to circumvent this by using relative gain $\psi(e)$ instead of absolute gain $\delta(e)$ for each edge e, in the hope that for such confounded variables neither edge direction would be visibly stronger and would therefore result in both these edges being ranked lower in importance. While we saw that this works well in practice, we do not have theoretical guarantees to back such an approach for causal discovery where causal sufficiency is violated. Working on extending Globe to entail such guarantees is one of the important lines of future work.

While we were able to scale Globe up to 500 variables using greedy DAG search, obtaining reasonable results during the process, there exist a number of alternate strategies that may work better. During experiments we encountered several individual cases where Globe got stuck in local optima due to a single wrong edge addition earlier in the search, which is a well known behavior for any greedy DAG search algorithm. For practical purposes we introduced the edge-flip phase to try and recover from such errors, to some degree of success. There exist, however, alternate strategies such as iterative sink/source selection (Peters et al., 2014) or permutation-based search (Squires et al., 2020). As opposed to edge-by-edge construction approach that we use, as future work, one could

investigate the conditions where sink/source selection or permutation-based approaches may be better suited than a greedy DAG search.

2.7 Conclusion

We developed an approach to discover fully directed causal graphs from observational data, using a score based on MDL to approximate Kolmogorov complexity. For non-linear mixture models with additive noise, this score identifies the Markov equivalence class and, with additional low variance assumption on the noise, finds the fully directed causal graph. We introduced Globe, a greedy DAG search algorithm that iteratively builds a locally optimal DAG by optimizing our proposed score. Through extensive experiments we showed Globe outperforms state-of-the-art methods, accurately orients edges with multiple parents, and efficiently infers networks up to 500 nodes.

Chapter 3

Information Theoretic Causal Discovery and Intervention Detection over Multiple Environments

In previous chapter, we introduced GLOBE to discover causal networks beyond Markov equivalence class. While GLOBE was a step forward, it worked with a somewhat restrictive assumption that we receive a *single*, *i.i.d* dataset as input. This is seldom the case in real-world scenarios where data may be collected in batches e.g. across different hospitals. In this chapter, we aim to relax this assumption. In particular, we consider the setting where we have multiple datasets generated by a shared underlying causal mechanism, but where each dataset is collected over a different environment. That is, each dataset obtains observations over the same set of variables, but with a different source distribution, or, may be generated through an intervention upon the underlying mechanism. Our goal is to jointly discover the overall causal network as well as the local interventions without knowing apriori which datasets are observational and which are interventional.

As a motivating example, suppose we are interested in learning the underlying causal process of some rare disease. A single hospital typically sees too few such patients as to collect sufficient data for drawing causal conclusions,

This chapter is based on Mian, Kamp, and Vreeken (2023b).

and hence we will have to consider data collected at multiple hospitals. It is at best cumbersome to centralize the data due to privacy regulations. Even if we could centrally collect the data, by their location, every hospital could have a different distribution of patients, and because of difference in machinery, etc., the parameters of the local data generating mechanisms will not all be exactly the same. If, for example, one hospital has a diagnostic device with an (undiscovered) internal anomaly, the data collected there will be from an intervention distribution, and pooling all data together in such cases can introduce bias in estimation (Lee and Tsui, 1982; Tillman, 2009).

While there exist approaches capable of discovering causal networks (Spirtes et al., 2000a; Chickering, 2002; Shimizu et al., 2006; Huang et al., 2018; Peters et al., 2014; Mian et al., 2021), they are designed to work only on a single dataset. Approaches that do take the multiple datasets into account work on strict assumptions such as having prior knowledge of intervention targets (Yang et al., 2018; Hauser and Bühlmann, 2012), can not match interventions on environments (Zhang et al., 2017) or impose strict assumptions on the underlying causal mechanisms that are unlikely to hold in practice (Shimizu, 2012; Ghassami et al., 2017).

To discover causal networks using data over multiple environments, we build our approach on the algorithmic model of causality. We use the postulate of Algorithmic Markov Condition (AMC) (Janzing and Schölkopf, 2010b) stating that the true causal factorization of the joint distribution has the lowest Kolmogorov complexity, which allows us to uniquely identify a fully directed overall causal networks and local interventions. As explained in Chapter. 2, Kolmogorov complexity is not computable itself, but can be instantiated in a statistically well-founded manner using MDL (Marx and Vreeken, 2021).

In this chapter, we define a theoretically sound MDL score for jointly discovering the causal model and local interventions, and provide a practical greedy-algorithm to optimize our proposed score. We explicitly do not assume any prior knowledge of which datasets are observational or interventional and neither assume anything about the functional form of causal relationships.

This chapter is organized as follows: we first introduce the essential notation in Section 3.1. Next we define the theory behind our approach as well provide consistency results in Section 3.2, and describe the Orion algorithm in Section 3.3. We discuss the existing approaches in Section 3.4 before providing empirical results and discussion in in Section 3.5 resp. Section 3.6.

3.1 Preliminaries

Setup and Notation For a set of random variables, $\boldsymbol{X} = \{X_1, \ldots, X_m\}$ with $X_i \in \mathbb{R}$, a *Structural Causal Model* (SCM) (Pearl, 2009) $\mathcal{S}$ models a joint distribution P over $\boldsymbol{X}$ corresponding to the observational distribution of the

system. For the scope of this work, we assume that we are given data $\boldsymbol{D} = \{D^1, \ldots, D^d\}$ from d environments, over $\boldsymbol{X}$, that share a common SCM. A causal DAG G over $\boldsymbol{X}$ is a graph in which the nodes represent random variables and edges identify the causal relationships as defined by $\mathcal{S}$. A directed edge between two variables $X_i \to X_j$ implies that X_i is a *direct cause* or *parent* of X_j, and X_j is a *child* of X_i. We denote the set of parents of X_j with pa_j and use $|\mathrm{pa}_j|$ to denote the size of the parent-set. Similarly we denote the set of children of X_j with ch_j and use $|\mathrm{ch}_j|$ to denote the size of the child-set. Given a sample $D \in \mathbb{R}^{m \times n}$ of size n from P, the goal of causal discovery is to identify the underlying causal *directed acyclic graph* (DAG) G entailed by $\mathcal{S}$ from this sample.

Similar to Chapter 2 we will work with the assumptions of 1) causal faithfulness, 2) the causal Markov condition (Spirtes et al., 2000a) and 3) causal sufficiency (Pearl, 2009), which makes it possible to discover causal networks from observational data up to the Markov equivalence class (Glymour et al., 2019). When we want to identify a fully oriented causal network we need additional assumptions (Peters et al., 2017), such as non-linear additive Gaussian noise models (Hoyer et al., 2009b) or the assumption of low-noise between causal pairs (Marx and Vreeken, 2019). We elaborate the latter in Sec. 3.2. Under these assumptions, fully directed causal networks can be identified and learned from observational data (Shimizu et al., 2006; Mian et al., 2021).

Intervention Detection An intervention set Υ over an SCM $\mathcal{S}$ defines any external perturbation that inhibits the influence of one or more parents of any $X_i \in \boldsymbol{X}$, resulting in a new joint distribution $\widetilde{P}$ over $\boldsymbol{X}$. If we were to know the true causal DAG G^* that models the observational distribution over $\boldsymbol{X}$ and have infinite samples from some new environment $\widetilde{D}$, it is straightforward to discover if $\widetilde{D}$ was generated from the original DAG G^* or an intervened DAG $\widetilde{G}$: First, we would discover $\widetilde{G}$ over $\widetilde{D}$. We can then simply consider the difference between the edge-sets $\mathcal{E}(G^*) - \mathcal{E}(\widetilde{G})$ to discover what are the intervened variables, if any.

In practice, neither do we have infinite data, nor do we know G^* in advance. Even if we could learn G^* from limited data D, we first need to ensure that there are no interventions present in D. This results in a cyclic dependency as learning what interventions are present in the data was our goal in the first place. The key question we hence need to answer is: How can we, given only limited data from multiple environments, *simultaneously* discover the true overall causal network, the local causal structures as well as the intervention targets within each environment? This we discuss next.

3.2 Causal Discovery from Data Drawn from Multiple Environments

In this section we build on the algorithmic Markov condition described in Section. 2.1 of the previous chapter, to identify the global resp. local causal models, as well as the intervention targets. Formally, our problem statement is:

Problem Statement 3.1 *Given samples* $\boldsymbol{D} = \{D^1, \ldots, D^d\}$ *over* d *environments that share a common SCM. Our goal is to (a) identify a single causal DAG* G^* *representing the true SCM; (b) identify which* $D^k \in \boldsymbol{D}$ *are interventional and which* $X_i \in D^k$ *are intervened upon; and (c) identify the local causal network,* G_k*, for each* D^k*.*

To address this, we first define our causal model, list down the assumptions necessary to prove identifiability and present a novel score. Then we show that the optimizer of this score identifies the true causal model and interventions in the limit.

3.2.1 Causal Model and Assumptions

We consider a setup where in each environment k, the value of each variable X_i is determined by a *non-linear* function f_i^k over its causal parents and additive independent Gaussian noise term with zero mean and unit variance N_i, regulated by a scaling factor α_i^k. For X_i in environment k we have

$$X_i := f_i^k(\mathrm{pa}_i) + \alpha_i^k \cdot N_i \ . \tag{3.1}$$

We assume that all N_i are jointly independent and that $N_i \perp\!\!\!\perp \mathrm{pa}_i$ for all $X_i \in D^k$. We assume that the number of parameters required to non-parametrically model f_i^k are upper-bounded by $O(\log n)$ (Mian et al., 2021).

Assumptions for Identifying Markov Equivalence Classes To discover causal networks up to Markov equivalence class we need to assume 1) the causal Markov condition, 2) the causal faithfulness (Spirtes et al., 2000a), and 3) causal sufficiency (Pearl, 2009). These assumptions allows us to guarantee identifiability up to the Markov equivalence class of DAGs, and not just partial ancestral graphs (PAGs) (Spirtes et al., 1999).

Assumptions for Identifying Fully Oriented Networks To ensure that we can orient edges between any pair of variables, and not just the edges coming into colliders, as is the case with the Markov equivalence class, we additionally need the low-noise assumption, meaning that the noise variance is sufficiently small for the causal *pairs* within a Markov equivalence class (Blöbaum

et al., 2018a) i.e. $\boldsymbol{\alpha} \to \mathbf{0}$, where $\boldsymbol{\alpha}$ is the vector consisting of scaling factors α_i^k for the bivariate causal edges and $\mathbf{0}$ is the null vector.[1] This, however, does not imply that the causal relationships are deterministic. For an extensive discussion on the low-noise assumption see Section 3 in Blöbaum et al. (2018a).

Assumptions for Identifying Interventions We assume that the true underlying causal network G that generates the data remains the same for all environments unless it is specifically changed by either (i) Hard-Interventions $\textsc{Hi}(X_j)$; or (ii) inhibiting Soft-Interventions $\textsc{Si}(X_j)$. A hard intervention on variable X_j eliminates the effect of pa_j on X_j, whereas a soft-intervention causes a *mechanism change* that sets the effect of a subset of pa_j to 0.

3.2.2 Encoding the Causal Model

To instantiate AMC (Eq. (2.1)) for our causal model (Eq. (3.1)) we need to define a lossless MDL score (Marx and Vreeken, 2021). The model class $\mathcal{M}$ that we consider for our proposed MDL score consists of all possible DAGs over $\boldsymbol{X}$, the set of local DAGs each environment, as well as the SCM that models f_i^k for all X_i in each $D^k \in \boldsymbol{D}$. The correct model $M \in \mathcal{M}$ is therefore one that minimizes $L(\boldsymbol{D}, M)$ such that

$$
\begin{aligned}
M^* &= \operatorname*{argmin}_{M \in \mathcal{M}} L(\boldsymbol{D}, M) \\
&= \operatorname*{argmin}_{M \in \mathcal{M}} \left(L(M) + \sum_{k=1}^{d} \sum_{i=1}^{m} L(X_i^k | pa_i^k, f_i^k) \right) \\
&= \operatorname*{argmin}_{M \in \mathcal{M}} \left(L(M) + \sum_{k=1}^{d} \sum_{i=1}^{m} L(\epsilon_i^k) \right)
\end{aligned}
$$

where pa_i^k are parents of variable X_i in dataset k according to the model M. We reformulate $L(X_i^k | pa_i^k, f_i^k)$ in the above equation by $L(\epsilon_i^k)$ to highlight that encoding each X_i once f_i^k and the parents are specified, comes down to storing

[1] Alternatively, we can make the assumption that these bivariate causal relationships are non-invertible. In this work, we make the low-noise assumption because it also covers the class of non-invertible causal relationships and is therefore a more general case of the two.

the residuals ϵ_i^k. We define the cost of the model as

$$L(M) = L_{str}(M) + \sum_{k=1}^{d} L_{mec}(M_k|M) \ ,$$

where L_{str} is the cost of storing the network structures and L_{mec} is the cost of storing the SCM once the structure is specified. Next, we describe what each of these costs are.

STRUCTURE The structure cost consists of the number of bits required to encode the global causal network as well as the interventions present in each environment. Formally we have

$$L_{str}(M) = L(G^*) + \sum_{k=1}^{d} L(G_k|G^*) \ ,$$

where we first encode the global causal network G^*, and for each G_k what are the interventions on G^*. Formally stated

$$L(G^*) = L_{\mathbb{N}}(d) + L_{\mathbb{N}}(m) + \sum_{i=1}^{m} L_{\mathbb{N}}(|pa_i|) + \log \binom{m}{|pa_i|} \ ,$$

where we first encode the number of environments, resp. variables, using $L_{\mathbb{N}}$, the optimal encoding for integers $z \geq 0$ (Rissanen, 1983). It is defined as $L_{\mathbb{N}}(z) = \log^* z + \log c_0$, where $\log^* z = \log z + \log\log z + \ldots$ and we consider only the positive terms, c_0 is a normalization constant to ensure the Krafft-inequality holds (Krafft, 1949). Then, for each of the m variables, we encode the number of parents $|pa_i|$ and identify pa_i from m using $\log \binom{m}{|pa_i|}$ bits.

Next we encode the local networks G^k once the interventions over G^* are provided, i.e. $L(G_k|G^*)$ is defined as

$$L(G_k|G^*) = \log(m) + \log \binom{m}{\tilde{m}^k} + \sum_{X_i \in \widetilde{X}^k} \log(|pa_i|) + \log \binom{|pa_i|}{|pa_i^k|}$$

For each local network, we encode the number, $\tilde{m}$ and identity $\widetilde{X}^k$ of intervened variables. Then, for each intervened variable, we identify the its active set of parents.

Combining the above, we have a lossless code for the causal structure.

MECHANISMS Next we define how to encode an SCM over M. Effectively we have to encode the function f_i^k for all X_i in each $D^k \in \boldsymbol{D}$. This is defined as

$$L_{mec}(M_k|M) = \sum_{i=1}^{m} L(f_i^k) \ .$$

Our causal model makes no assumption on the functional form of the causal relationship. We model each f_i^k *non-parametrically.* In particular we use multivariate regression splines (Friedman, 1991) of the form $X_i := \sum_{j=1}^{|H|} f_j(\mathcal{P}_j)$, where f_j is a hinge function applied to a subset $\mathcal{P}_j$ with size $|\mathcal{P}_j|$, of X_i's parents. Recall from Chapter 2 that a hinge function is of the form $f(\mathcal{P})j) = a \cdot \prod_{t=1}^{T} \max(0, g_t(\text{pa}_t) - b_t)$, where T denotes the number of multiplicative terms in the hinge, $\text{pa}_t \in \mathcal{P}$ is the parent associated with the t-th term, and g_t is a non-linear transformation from a finite function class $\mathcal{F}$ applied to pa_t. The cost to store the causal mechanism using multivariate regression splines can then be defined as

$$L(f) = L_{\mathbb{N}}(|H|) + \sum_{h_j \in H} \Big[L_{\mathbb{N}}(T_j) + \log \binom{|\mathcal{P}| + T_j - 1}{T_j} + T_j \log(|\mathcal{F}|) + L_p(\theta_j) \Big].$$

We use $L_{\mathbb{N}}$ to encode the number of hinges. Then for each hinge, we encode the number of terms per hinge, the correct assignment of terms T_j to parents in $\mathcal{P}$, the number of bits to identify non-linear transformations used for each term in the hinge, and parameters θ_j associated with th j-th term. We encode the parameters θ_j using $L_p(\theta_j)$ defined it Section 2.2. As the precision upto which each value is stored is fixed, computing L_p does not depend on sample size n.

RESIDUALS As a final step to obtaining a lossless score, we need to encode the noise that remains in the system once the specified model has captured the structure and generating mechanism of the data. Since we use regression functions, we aim to minimize the variance of the residual, and hence encode the residual ϵ as Gaussian distributed with zero-mean (Grünwald, 2007):

$$L(\epsilon_{i,k}) = \frac{n}{2} \left(\frac{1}{\ln 2} + \log 2\pi \hat{\sigma}_{i,k}^2 \right) \ ,$$

where we compute the empirical variance $\hat{\sigma}_{i,k}^2$ from the residual, ϵ_i^k.

Combining all of the above, we have a lossless MDL score by which we can instantiate the AMC. Next we establish theoretical guarantees entailed by the defined causal model and prove that the minimizer of $L(\boldsymbol{D}, M)$ identifies the

correct causal network and interventions in the limit.

3.2.3 Asymptotic Guarantees

We now provide formal guarantees and show that our proposed score in Section 3.2.2 is consistent when $n \to \infty$. We show that under the assumptions described in Section. 3.2.1, it identifies hard interventions as well as inhibiting soft-interventions.

We begin by showing that missing edges in local causal networks are the result of interventions.

Lemma 3.1 $\forall i, k \;\; HI(X_i^k) \iff pa_i^k = \emptyset \;,$ *and* $SI(X_i^k) \iff pa_i^k \subset pa_i$

To provide further identifiability results we first state the definition of a *conservative* set of interventions as stated by Hauser and Bühlmann (Hauser and Bühlmann, 2012).

Definition 3.2 ((Hauser and Bühlmann, 2012)) *A set of interventions* $\boldsymbol{\Upsilon}$ *is conservative, if* $\forall X_i \in \bigcup_{k=1}^{d} \Upsilon^k, \exists \Upsilon^k \in \boldsymbol{\Upsilon}$ *such that* $X_i \notin \Upsilon^k$.

Simply put, a set of interventions $\boldsymbol{\Upsilon}$ is conservative if for each variable X_i we can find at least one environment in which it is not intervened upon ($X_i \notin \Upsilon^k$). This implies that to reliably find the set of parents for X_i, we need to atleast observe it once without intervention.

Further, let G^* be the true global network and G^k be the network discovered for environment k.

Lemma 3.3 *If* $\boldsymbol{\Upsilon}$ *is conservative,* $\bigcup_{k=1}^{d} G_k = G^*$*, if* $\boldsymbol{\Upsilon}$ *is non-conservative,* $\bigcup_{k=1}^{d} G_k \subseteq G^*$.

Lemma 3.3 shows that under the conservative intervention assumption, we can discover the underlying global causal network and that under the violation of this assumption, the discovered network will be a subgraph of the true network.

Next, we provide our main result. We can show the following best resp. worst case result that we can guarantee for the causal model defined in Eq. (3.1). Moreover we can still identify the correct Markov equivalence class, even when the low noise assumption is violated.

Theorem 3.4 *Let* $\mathcal{Y}$ *be the set of all non-collider nodes. If* $\forall Y_i, k \; \alpha_i^k \to 0$*,* $L(\boldsymbol{D}, M)$ *will be the lowest for the true fully-oriented causal network.*

Theorem 3.5 $L(\boldsymbol{D}, M)$ *correctly identifies the collider structures in the underlying causal network.*

As a sketch of proof, note that to prove both Thm. 4 and 5, it suffices to show that $L(\boldsymbol{D}, M)$ is a valid L_0 regularized score (e.g. BIC). Note that showing $L(\boldsymbol{D}, M)$ is a valid L_0 regularized score suffices to prove Thm. 5 and the only additional step needed to prove Thm. 4 is the low-noise assumption as this lets us identify bivariate cases in the resulting Markov Equivalence class (Marx and Vreeken, 2019).

For an intuitive explanation of our main result, consider Thm. 3.5 first. Identifying collider structure means that our score identifies causal DAGs up to Markov equivalence class at the very least. This implies that any undirected edges that exist in the final network are between variables that are not colliders. For such case, our causal model simplifies to the pair-wise model of Marx and Vreeken (2019). They prove that under the low-noise assumption, orientation of such pair-wise edges is identifiable using an L_0 regularized score (e.g. BIC). Meaning, for the pair-wise model between variables X and Y, the BIC score for regressing Y onto X, resp. X onto Y, will be highest in the causal direction. Next, note that the BIC score is equal to the negative of the MDL criterion. Thus, if we were to score *all* Markov equivalent DAGs using an MDL based L_0 regularized score, the causal one will obtain the lowest score. Consequently, to prove Thm. 3.5, we reformulate $L(\boldsymbol{D}, M)$ to show that it is a valid L_0 regularized score. Using this score in conjunction with the low-noise assumption stated in Sec 3.2.2 lets us orient any remaining edges in the causal network, which proves Thm. 3.4.

It is worth noting that our proposed score identifies the fully oriented causal network, it neither requires using distribution-shifts nor introducing additional context variables to orient any remaining edges. These theoretical guarantees, however, only hold if we score all possible DAGs over the data. This quickly becomes infeasible for large graphs. Indeed, finding the exact Bayesian network is known to be NP-hard (Chickering et al., 2004). Hence, we propose a heuristic-based practical approach to minimizing $L(\boldsymbol{D}, M)$.

3.3 The Orion Algorithm

In this section we present a practical algorithm Orion for discovering causal DAGs from multivariate continuous valued data over multiple environments. Orion greedily adds and removes edges to the global resp. local causal networks such that it reduces $L(\boldsymbol{D}, M)$ most. Akin to Globe, it performs forward and backward search, repeated until convergence. We provide the algorithm outline in Alg. 3.1 and give detailed pseudocode. It learns a causal network by iteratively adding and removing edges to the global structure, and encoding interventions for the datasets that reject the globally introduced edges. As output, it returns the (intervened) local causal networks. We take union over these networks to reconstruct the predicted global causal network (Lem. 3.3)

Algorithm 3.1: The ORION Algorithm

Input: Datasets $\boldsymbol{D}$ over $\boldsymbol{X}$
Output: Array of causal networks $\boldsymbol{G}$

```
for k = 1 ... d do
    G_k ← ∅
G ← [G_1, ... G_d]
repeat
    G ← ForwardSearch(G, D)
    G ← BackwardSearch(G, D)
until convergence;
return G
```

Algorithm 3.2: Forward Search

Data: Environments $\boldsymbol{D}$ over $\boldsymbol{X}$, array of causal networks $\mathbf{G}$
Result: Array of updated networks $\mathbf{G}$

```
E*(G) ← all possible edges in G
E_cand ← E*(G) − E(G)
Q ← ScoreEdgeAddition(E_cand)
while Q not empty do
    e ← take top most entry from Q
    ch_e ← child variable for edge e
    if G ⊕ e not cyclic and e is significant then
        G ← G ⊕ e
        foreach edge e^k, connected to ch_e ∈ Q do
            update score of e^k ∈ Q to ψ(e^k)
return G
```

and take the difference between the edge-sets of global and local causal networks to determine the intervention targets (Lem. 3.1) As our score is lower-bounded at 0, and we only take steps that reduce our score, it is guaranteed to converge. Even though the guarantees of greedy DAG search are limited to causal trees, we show in Sec. 3.5 that ORION outperforms state-of-the-art search algorithms. Next, we describe the ranking mechanism and the search phases.

Algorithm 3.3: Score Edge Addition

Data: edgeset $\mathcal{E}$ over $\mathbf{G}$
Result: priority queue of edges Q

1 $Q \leftarrow \emptyset$
2 **foreach** *pair* $(u, v) \in \mathcal{E}$ **do**
3 $\quad \psi \leftarrow \delta^{\oplus}(e_{uv}) - \delta^{\oplus}(e_{vu})$
4 $\quad Q \leftarrow Q \oplus (e_{uv}, \psi)$
5 $\quad Q \leftarrow Q \oplus (e_{vu}, -\psi)$
6 *return* Q

EDGE GAIN To calculate the gain provided by each edge, we first measure the bits that we save by adding an edge in the current model. Formally, let $e_{ij} = X_i \rightarrow X_j$, and M be the current model. We write $M \oplus e_{ij}$ to denote the model with edge e_{ij} included. We define the absolute gain in bits δ associated with edge e_{ij} as

$$\delta(e_{ij}) = \max\{0, L(\boldsymbol{D}, M) - L(\boldsymbol{D}, M \oplus e_{ij})\} \ .$$

Next, we calculate the true gain for this edge by calculating the relative bits we gain over adding this edge in the opposite direction. Formally,

$$\psi(e_{ij}) = \delta(e_{ij}) - \delta(e_{ji}) \ .$$

Intuitively, the higher the value of $\psi(e_{ij})$, the more certain we are that we inferred the correct direction for this edge. This is motivated by the no-hypercompression inequality (Grünwald, 2007), which we use to test the significance of each edge. Let $s = \psi(e)$, the probability of gaining s bits over the null model is less than or equal to 2^{-s}. If we find that the gain for an edge is not significant— i.e. 2^{-s} is greater than the desired significance threshold— we do not add this edge.

FORWARD SEARCH In forward search, we maintain a priority queue containing the edges e_{ij} ordered by the gain in bits $\psi(e_{ij})$, when adding the edge to the model. This edge scoring is done by the SCOREEDGEADDITION function (line 3) shown in Alg. 3.3. We iteratively build the causal graph by adding the highest ranked edge from the priority queue to the global causal DAG (lines 5-6). We reject edges that introduce cycles in the network (line 7). Once an edge e_{ij} is added to the network, we re-rank all the candidate edges associated with variables X_j in the priority queue (lines 9-10). We repeat this until all the edges have been evaluated and no edge addition provides gain anymore.

Algorithm 3.4: Backward Search

Data: Environments $\boldsymbol{D}$ over $\boldsymbol{X}$, array of causal networks $\mathbf{G}$
Result: Array of updated networks $\mathbf{G}$

```
1 E_cand ← E(G)
2 Q ← ScoreEdgeRemoval(E_cand)
3 while Q not empty do
4     e ← take top most entry from Q
5     if e is significant then
6         G ← G ⊖ e
7         foreach edge e^k, connected to ch_e ∈ G do
8             update score of e^k ∈ Q to ψ(e^k)
9 return G
```

Algorithm 3.5: Score Edge Removal

Data: edgeset $\mathcal{E}$ over $\mathbf{G}$
Result: priority queue of edges Q

```
1 Q ← ∅
2 foreach pair (u, v) ∈ E do
3     ψ ← δ^⊖(e_uv)
4     Q ← Q ⊕ (e_uv, ψ)
5 return Q
```

We introduce each edge as part of the global network which means that the structure cost is shared across datasets. Each of the datasets, therefore, only need to pay a discounted cost of storing their causal mechanism in order to include this edge. If the discounted cost is not enough to register a gain, an intervention is encoded for this dataset.

Backward Search Since we greedily add edges during the forward search phase, some parents of variable X_j may become redundant as forward search progresses. This is because a subset of these parents may be able to explain X_j better. To remove these redundant parents, we need a backward search as shown in Alg. 3.4. We populate a priority queue containing the edges e_{ij} ordered by the gain in bits $\psi(e_{ij})$, when removing the edge from the model. This edge deletion scoring is done by the ScoreEdgeRemoval function (line

2) shown in Alg. 3.5. We iteratively remove that edge from the network which improves score the most (lines 4-6). After removing the edge, we update the costs associated with remaining parents in the priority queue (lines 7-8). We remove edges until no edge removal improves $L(\boldsymbol{D}, M)$ anymore.

Complexity Analysis We first make a pass over the entire edge-set for each environment to determine the initial edge gains. This requires $\mathcal{O}(cdm^2 \log m)$ steps where c denotes the complexity of the regression approach that is used. In forward search, each edge can lead to at most $m-1$ ranking updates, each of which require $\mathcal{O}(\log m)$ time when priority queue is implemented as a heap. Resulting in a complexity of $\mathcal{O}(cdm^3 \log m)$. The backwards search has a similar upper bound of $\mathcal{O}(cdm^3 \log m)$. Hence, the overall complexity is in $\mathcal{O}(cdm^3 \log m)$. Orion compares favorably to the worst-case complexities of Pc , $\mathcal{O}(2^m)$, Ges, $\mathcal{O}(2^m)$, CdNod, $\mathcal{O}(n^3)$. Orion is inherently parallelizable over both edges and environments, and we implement it as such. It is therefore quite fast in practice.

3.4 Related Work

There exist many proposals for discovering causal networks from a single (typically observational) i.i.d. dataset (Spirtes et al., 2000a; Chickering, 2002; Huang et al., 2018; Compton et al., 2021), which discover partially directed causal networks. While Globe (Mian et al., 2021) discussed in Chapter 2 discovers fully directed networks, it is restricted to a single dataset and can not handle interventions. Initial proposals that discover causal networks over multiple environments focused on single target variables (Peters et al., 2016; Yu et al., 2019a) and can not trivially be extended to discover causal networks. Many methods assume we either know the intervention targets (Hauser and Bühlmann, 2012; Triantafillou and Tsamardinos, 2015; Yang et al., 2018), or the environments that were intervened upon(Squires et al., 2020; Brouillard et al., 2020). Recently, Faria et al. (2022) proposed an approach to relax the assumption of known intervention environments.

Approaches that do not need prior knowledge of interventions substitute it with other restrictive assumptions such as assuming a single type of intervention (Cooper and Yoo, 1999; Kocaoglu et al., 2019) or fixing a functional form between cause and effect (Eaton and Murphy, 2007; Shimizu, 2012). In practice we often neither know which environments are interventional, nor do we know intervened variables, nor the causal functional forms.

The task of discovering causal networks over multiple environments without assuming any prior knowledge of interventions has been addressed by introducing an additional *context* variable that takes a fixed value within each environment (Zhang et al., 2017). While a single context variable allows to identify

intervention targets across different environments, one can not single out the environment where the intervention happens. Mooij et al. (2016) propose the unifying Joint Causal Inference (JCI) framework that can be implemented using any constraint-based causal discovery algorithm. JCI proposes to introduce one context variable per environment, thereby allowing localization of intervention targets within each context. JCI, however, outputs the overall global causal network and the intervention targets. It does not give us information about what are the local causal networks within environment, or what type of intervention has been performed. Finally, Jaber et al. (2020) recently provide a graphical characterization for testing whether two causal graphs with potentially different intervention targets belong to the same equivalence class. They, however, works under the assumption that the underlying structure stays the same for all the environments.

3.5 Evaluation

In this section we empirically evaluate ORION, we are mainly interested in answering the following three questions – (1) Does ORION accurately discover causal networks over data from multiple environments? (2) How well does ORION perform on real world networks where our assumptions may not hold? and (3) Does ORION reliably identify intervention targets? We first describe our experimental setup and then answer these questions in the subsequent set of experiments.

Setup We compare to state-of-the-art approaches from the classes of ANM, constraint, and score-based methods. As the representative ANM-based method, we compare to MultiGroup-LINGAM (Shimizu, 2012) which is an extension of the original LINGAM (Shimizu et al., 2006) to multiple datasets. For constraint-based methods, we compare to CDNOD (Zhang et al., 2017), and to the JCI framework of Mooij et al. (2016) using PC (Spirtes et al., 2000a) resp. FCI (Spirtes et al., 1999). For score-based approaches, we compare to the permutation-based greedy search approach, UT-IGSP (Squires et al., 2020), the GES algorithm (Chickering, 2002; Ramsey et al., 2017) using the two-layer approach proposed by Eaton and Murphy (2007), which we refer to as EGES. As baseline, we compute results over vanilla fast-GES (FGES) (Ramsey et al., 2017) by taking a union over locally discovered networks. Furthermore, we include our previously proposed approach GLOBE to these evaluations to see how its performance changes as we introduce new data that breaks the assumptions behind GLOBE. To learn network using latter, we stack all data together and give it as input to GLOBE.

We evaluate the quality of the discovered networks in terms of structural similarity using the Structural Hamming Distance (*SHD*) (Kalisch and Bühlmann, 2007) which measures the number of edges in which two networks

d	ORION	LINGAM	UT-IGSP	PC	CDNOD	GLOBE
3	**0.45**	0.58	0.58	[0.47, 0.67]	[0.48, 0.55]	0.37
5	**0.44**	0.55	0.58	[0.45, 0.67]	[0.44, 0.48]	0.30
7	**0.42**	0.53	0.57	[0.42, 0.65]	[0.56, 0.66]	0.27
9	**0.43**	0.52	0.57	[0.44, 0.63]	[0.60, 0.70]	0.22

Table 3.1: [Lower is Better] Averaged normalized *SID* for synthetic data with $m = 10$. Intervals indicates the best, resp. worst possible intervention distance for methods that output the Markov equivalence class of the causal network. GLOBE shows superior performance with increasing environments as it first stacks all data together. For this experiment, all samples follow the i.i.d. assumption, which works in favor of GLOBE.

differ. *SHD*, however, tells us nothing about the difference in networks' causal implications. To measure this causal similarity, we use the Structural Intervention Distance (*SID*) (Peters and Bühlmann, 2015). *SID* counts those pairs of variables X_i and X_j, such that the effect experienced by X_j due to an intervention on X_i differs between two networks. For comparability over different datasets, we normalize *SHD* and *SID* between 0 and 1. To avoid practical issues like var-sortability (Reisach et al., 2021), we standardize all data. We make our code and data available for research purposes[2].

Q1. Does Orion accurately discover causal networks over data from multiple environments? We start with a simple setting where we generate multiple datasets using the same underlying distribution. We simulate DAGs using the Erdős-Rényi model. We model effect as a function of its causes using polynomial functions in half of the cases. For other half we use randomly initialized 2-layer neural networks to model the mechanism. We average the resulting *SID* over 100 different runs and report the results in Table. 3.1. We omit JCI-FCI because it almost always returns empty networks, and FGES as it reports *SID* intervals too wide to convey meaningful information. We find that ORION reports the best *SID*, at least as good as the lowest score over the equivalence classes that PC resp. CDNOD report. For this case, we see that GLOBE actually has the best performance among *all* methods by a large margin, which only gets better with more environments being observed. This result is not suprising given that each dataset is i.i.d. and stacking them together does not introduce any biases. In fact, such stacking works in favor of GLOBE since it has larger number of samples to work with, which results in a clearly superior performance.

[2]https://eda.rg.cispa.io/orion

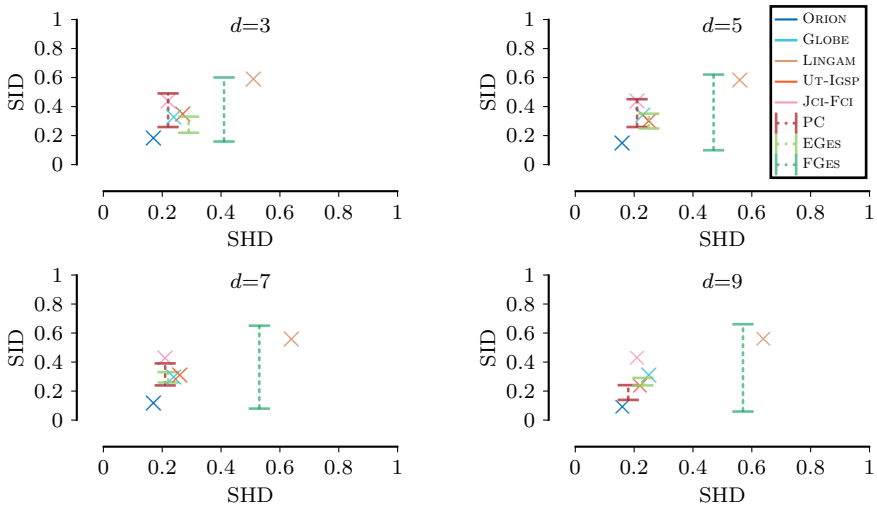


Figure 3.1: [Closer to origin is better] Comparison of normalized *SHD* and *SID* for environment sizes $d \in \{3, 5, 7, 9\}$ when all environments contain data from a different intervention distribution over the same causal network. Dotted lines indicate the uncertainty interval over *SID* for PC, EGES and FGES. GLOBE deteriorates considerably as stacking the data for this setting introduces biases in result because the i.i.d. assumption no longer holds.

Next, and more interestingly, we generate each environment using *different* intervention distributions from a fixed underlying causal network. This means that the data for each environment comes from a different (sub)network, about which we know neither the type nor the targets of intervention. We report the results in Fig. 3.1 where we see that ORION performs best whereas GLOBE this time deteriorates noticeably, which results in a performance that is worse than PC and GES. This corroborates the fact that stacking data together from different environments can indeed introduce bias during learning (Lee and Tsui, 1982; Tillman, 2009), thereby rendering the results from GLOBE unreliable for such scenario. We omit CDNOD as it is unable to handle the cases involving hard interventions.

Q2. How well does Orion perform when assumptions may not hold? To this end, we use the re-simulated Lung-cancer gene expression, REGED network (Statnikov et al., 2015). We extract two non-overlapping connected components of 5 resp. 15 variables, which we refer to as REGED 5 and REGED 15. For both networks, we randomly divide the data into 3 environments containing 250 samples each.

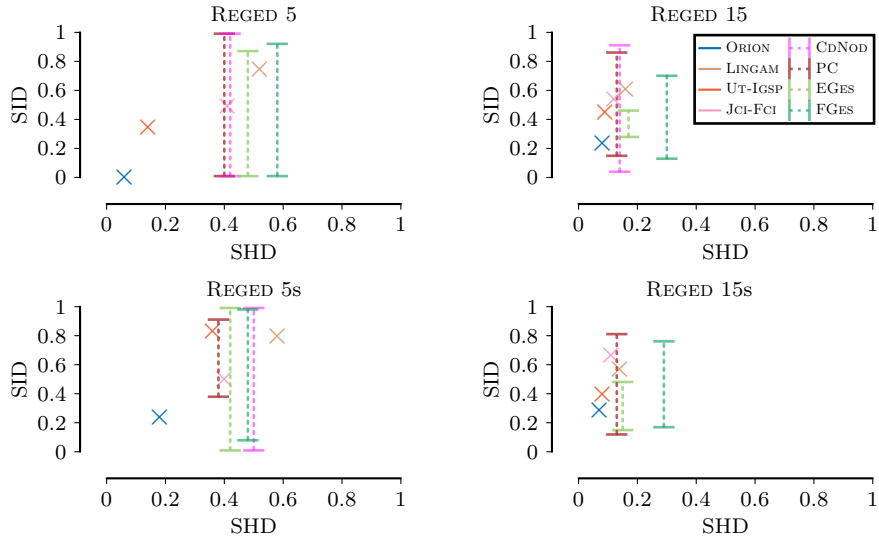


Figure 3.2: [Closer to origin is better] Comparison of normalized *SHD* and *SID* for the Reged networks without selection bias (Reged5, Reged15) and with selection bias (Reged5s, Reged15s). Dotted lines indicate the interval over *SID* for PC, EGes and CdNod.

Next, we introduce selection bias in the data by sorting on one of the variable and dividing the resulting dataset into three partially overlapping datasets of 200 samples each. We repeat this for each variable thereby giving us a total of 5 resp. 15 separate experiment instances for each network. We refer to these datasets as Reged5s resp. Reged15s.

We show the results for both aforementioned setups in Fig. 3.2 where we see that Orion performs the best overall. Moreover, we see that EGes, CdNod and PC have very wide *SID* intervals, which restricts us from drawing useful causal conclusions from the discovered networks.

Q3. Can Orion reliably identify intervention targets? We test how well Orion can identify both direct and indirect intervention targets over multiple environments. We use the same structure as used by Zhang et al. (2017) for their experiments and report the F1-scores for this experiment in Fig. 3.3. We see that Orion gets an F1-score average of 0.63, which is twice as good as Lingam and PC. Surprisingly, FGes, although only a baseline, performs better than both Lingam and PC.

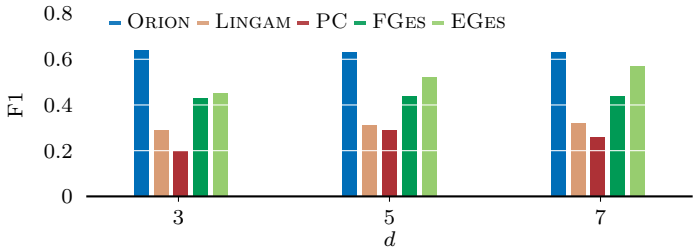


Figure 3.3: [Higher is better] F1 scores for ORION, LINGAM, PC, EGES and FGES for identifying intervention targets in synthetic data over different environment sizes, d. We omit CDNOD as it can not match intervention targets to environment.

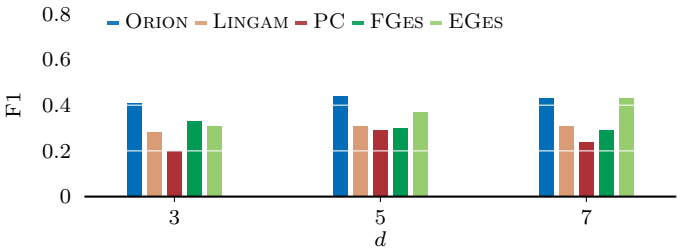


Figure 3.4: [Higher is better] F1 scores for identifying direct intervention targets in synthetic data over different environment sizes, d. We omit CDNOD as it does not contain a mechanism to identify intervention targets within each environment.

3.6 Discussion

We proposed a novel score for the discovery of causal networks over multiple environments based on the algorithmic Markov condition and its approximation via MDL. Our analysis proved that optimizing this score identifies the true DAG and all local interventions in the limit. This allows us to simultaneously discover the underlying causal mechanism and local interventions over multiple datasets. We proposed a practical algorithm ORION which, through extensive experiments, we showed that it outperforms the state of the art at discovering the true causal networks given multiple datasets, even when all the environments contain data generated from unknown intervention distributions over the same network, and reliably identifies intervention targets.

Although non-trivial, it is a promising direction to investigate implementing the GES (Chickering, 2002) procedure using ORION score as a line of future work. Such an implementation will extend theoretical guarantees entailed by our proposed score to also hold for the practical implementation. This ex-

tension while theoretically appealing, may not be straightforward as performing an efficient search using GES requires that our proposed score satisfies the score-equivalence property (Chickering, 2002), while we explicitly use non-score equivalence to quantify and exploit the asymmetry in explaining cause from effect. Using non-efficient GES can allow us to overcome this limitation, at the expense of the worst-case runtime becoming exponential in the number of variables.

Another worthwhile area of exploration would be to investigate evolving our proposed score to handle edge-introducing interventions alongside inhibiting interventions that we already consider. Maintaining identifiabilty guarantees while doing so is rewarding yet challenging line of future work as this would require redefining what we mean by "true" underlying causal network and reformulate identification guarantees with respect to our new definition.

3.7 Conclusion

In this chapter, we introduced novel scores for discovering causal networks across multiple environments grounded in algorithmic Markov condition and MDL approximation. Our analysis showed that optimizing our proposed score accurately identifies the true Causal DAG and local interventions in the limit, allowing for the discovery of underlying causal mechanisms across various datasets. As practical instantiation we developed the ORION algorithm, which extensive experiments demonstrate, outperforms current state-of-the-art methods in identifying true causal networks and reliably pinpointing intervention targets, even with data from unknown intervention distributions.

Chapter 4

Learning Causal Networks from Episodic Data

Until this point we have considered those settings where all data is fixed, be it in a single dataset or pre-specified static multiple datasets that never get updated. While the latter might still be applicable to certain domains where we have collected enough data across different environments, it may not be straightforward to update our knowledge systematically once new data becomes available. In contrast, a more realistic setting is one where we obtain observations in batches, or episodes, at different points in time, potentially forever. Not only does this mean that we need to learn and update our causal hypothesis over time, but each batch likely contains samples from a specific time period or sub-population, resulting in a biased distribution. Even the collective data distribution over all episodes is often not identically distributed since the causal interactions could differ across domains or change over time.

To motivate the episodic setting and illustrate its challenges, consider an example in environmental monitoring where we measure two markers X_1: *temperature* and X_2: *ozone concentration* over the course of a year. Suppose we obtain measurements at the end of each quarter, resulting in episodes $\{D^{(1)}, ...D^{(4)}\}$ at timepoints $\{t^{(1)}, ...t^{(4)}\}$. We show the data in Fig. 4.1, coloring samples by episode. In our example taken from the Tübingen cause-effect pairs (Mooij et al., 2014), X_1 is considered the cause of X_2 and the overall data suggest a roughly linear trend of the causal mechanism relating them. However, when we

This chapter is based on Mian and Mameche (2024) and Mian, Mameche, and Vreeken (2024).

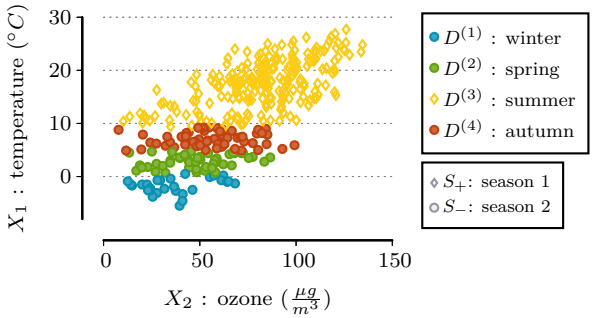


Figure 4.1: Cause X_1 and effect X_2 (Mooij et al., 2014) measured in episodes over time ($D^{(1)}$-$D^{(4)}$). Each episode comes from an unknown season (S_+, S_-), and unknown context (Switzerland in this case). We are interested in causal discovery with data arriving in episodes over time. For example, we could obtain episodes from different *seasons* where individual seasons (e.g. $D^{(1)}$) show a biased trend between two causally related variables (X_1, X_2). Episodes could also come from different *contexts* (a different location, e.g. Sahara Desert) where a different causal model applies.

consider the winter months $D^{(1)}$ (blue) on their own, it appears that both variables are uncorrelated. The same is the case for $D^{(2)}$ and $D^{(4)}$, and only once we include the summer months $D^{(3)}$ (yellow) do we obtain a complete picture. To show which part of the domain of X each episode covers, we can consider different subregions, for example, there is a high-temperature season S_+ where some samples are *observed* (diamond), others *missing* (circle), similarly a low-temperature season S_-. Episodes coming from such a specific subregion likely have a biased distribution.

While this simplistic example suggests that combining all episodes is a good practice to remove seasonal bias, this can lead to its own set of issues. Data could also arrive from a different geographical region or *context* where due to local measuring devices noise levels are different, or even the underlying causal relationship changes. For example, a phenomenon known as ozone suppression (Steiner et al.) leads to a different trend where ozone levels are no longer positively correlated with temperature. Ozone suppression only occurs above a temperature threshold, hence is not visible in the data obtained in Switzerland shown in Fig. 4.1 but likely the case if a future episode $D^{(5)}$ arrives from a region with exceptionally high temperatures. Overall, whereas episodes $D^{(1)} - D^{(4)}$ should be combined to remove seasonal bias, combining samples from different contexts $D^{(1)} - D^{(5)}$ obscures context-specific causal relationships (Zhang et al., 2017).

While recent work in causal discovery considers different contexts (Mooij et al., 2016; Zhang et al., 2017; Squires et al., 2020), it neither addresses episodes

nor allows for structural changes in the causal model across contexts. In contrast, we propose a causal modeling framework for episodes with selection bias where an unknown number of causal networks underlie the data-generating process. We show that in principle, we can use a consistent scoring criterion for causal discovery in this setting so long as we observe sufficiently many episodes.

From a practical perspective, existing algorithms for causal discovery (Pearl, 2009; Chickering, 2002; Mian et al., 2021) start from a single batch of data and hence would need to relearn the causal model whenever a new episode arrives. Not only is this computationally impractical, but a domain expert likely wants to gain preliminary insights into the causal relationships already based on data from the earlier episodes, and be ready to update these insights as new data becomes available.

To address these limitations, we develop the algorithm CONTINENT for discovering a faithful causal network, or multiple networks, over episodic data. Taking inspiration from continual learning, we hereby avoid fully re-learning the causal model upon the arrival of each episode but learn it in an online fashion. We propose a strategy to update the causal hypothesis as new episodes arrive, using distribution matching and an information-theoretic perspective of causality, and show that our updating strategy is consistent. We show in experiments that CONTINENT discovers causal networks reliably from data with episodic selection bias, under interventions, as well as structural changes in causal networks. Not only does it compare favorably to its competitors, but CONTINENT alone can learn the causal model adaptively over time, and can address a novel experimental setting where different causal networks underlie episodic data and we predict, for a new incoming episode, which causal network it is generated from. To summarize our main contributions, we

- introduce a causal modeling framework for episodic data,
- show under which conditions we can use a consistent information-theoretic scoring criterion to identify the underlying set of causal networks,
- develop the practical approach CONTINENT for efficient learning of the causal networks in a continual fashion,
- confirm in experiments that CONTINENT works in practice.

We structure our exposition according to the above, we first introduce notation and preliminaries in Section 4.2, then provide our causal model and practical algorithm in Section 4.3 resp. 4.4, before concluding with experimental evaluation and discussion in Sections 4.5 and 4.6.

4.1 Related Work

As we have seen from the previous chapters, causal discovery approaches typically fall into the categorizations of constraint-based methods, such as PC (Pearl, 2009), or score-based methods, such as GES (Chickering, 2002; Ramsey et al.,

2017). However, the examples given up to this point assume an i.i.d. data distribution where a single causal network can capture the causal interactions, and where neither selection bias nor contexts exist. We therefore, discuss the related work in those two aspects.

Selection Bias Missingness is a well-studied problem in statistical inference and in particular, many approaches exist for *correcting* for missingness and selection bias (Gretton et al., 2008; Sugiyama et al., 2007; Boeken et al., 2023); see Little and Rubin (2019) for an overview. Only very recent work studies assumptions for *identifying* whether selection bias holds in a given dataset (Kaltenpoth and Vreeken, 2023c). Our perspective is different as we are interested in *adapting* causal discovery to the presence of missingness. An important line of work studies *recoverability* (Bareinboim et al., 2014; Pearl, 2012) from selection bias in causal discovery, modeled through unobserved sink node S in the causal graph. We also adopt this model here using multiple missingness regions, and in addition consider the presence of multiple contexts in the form of varying causal mechanisms.

Different Contexts A wealth of recent literature studies causal discovery from different environments, experimental regimes, or contexts (Zhang et al., 2017; Squires et al., 2020; Jaber et al., 2020; Magliacane et al., 2018); prominent examples include the constraint-based JCI framework (Mooij et al., 2016), additive noise model based multi-group Lingam (Shimizu, 2012), and score-based approaches (Eaton and Murphy, 2007; Mian et al., 2023b) for discovering causal DAGs from multi-context data. However, existing work assumes that each context is an identically distributed (i.i.d.) sample with fixed causal model. We make this setting more general in that we obtain biased samples from each context, which need to be combined to result in i.i.d. data. To our knowledge, we are the first to allow a different causal model with episodic bias in different contexts, and also address the algorithmic challenges associated with discovering causal networks in an online fashion.

To demonstrate how classical and environment-based causal discovery approaches fare with episodic selection bias in practice, we next compare them against Continent.

4.2 Preliminaries

First, we outline our problem setting and review causal modeling techniques for independent and identically distributed (i.i.d.) data.

4.2.1 Notation and Problem Setting

Throughout this chapter, we consider a batch setting where we obtain observations as a sequence of perennially arriving datasets $\{D^{(1)}, D^{(2)}, \ldots\}$ at timepoints $\{t^{(1)}, t^{(2)}, \ldots\}$, and refer to dataset $D^{(i)}$ at time $t^{(i)}$ as an *episode*. We denote the dataset that combines all episodes up to time $t^{(d)}$ as $\mathbf{D}_d = \cup_{i=1}^{d} D^{(i)}$. In each episode, we observe a fixed set of continuous random variables $\boldsymbol{X} = \{X_1, \ldots, X_m\}$ with overall distribution $P(\boldsymbol{X})$.

Different Contexts Our main interest are the causal interactions between $\boldsymbol{X}$. As our motivating example suggests, these could change over time or across domains, hence unlike Globe or Orion, we do not assume a fixed causal model over $\boldsymbol{X}$. Rather, we consider different *contexts*, also known as regimes or environments, with a different causal model each. We denote the set of contexts as $\{C_0, \ldots, C_R\}$. Each episode $D^{(i)}$ is a member of a unique context, which we write as $C(D^{(i)})$ for short. We write X^r, P^r to refer to variables, resp. distributions, in the rth context. Novel to our work is that we neither know how many contexts R exist nor which context $C(D^{(i)})$ each episode comes from.

Biased Episodes In addition to coming from different contexts, i.i.d.-ness may not hold for our episodes. Each episode could for example preferentially include samples from a certain subpopulation. In other words, some samples from the population $P(X)$ are *missing* and others *observed* in a given episode. We can represent this using a binary variable S taking labels $S = 0$ for missing, $S = 1$ for observed samples. As an illustration, Fig. 4.1 shows the season S_+ with values $S_+ = \diamond$ for observed, $S_+ = \circ$ for missing samples, so that episode $D^{(3)}$ only includes the observed samples. We can also consider multiple such seasons, for example both $\{S_+, S_-\}$. In general, we consider a categorical variable S with values $\{s_1, \ldots, s_K\}$. In our example, colored samples from each season form a biased distribution, such as the summer season $P(X \mid S_+ = \diamond)$.

In this episodic setting, we want to discover how many and which causal models there are.

Problem Statement (informal). *Given datasets $\{D^{(1)}, \ldots, D^{(d)}\}$ where each episode $D^{(i)}$ is generated from the causal model in an unknown context C_r and by conditioning on an unknown value s_k of S, we want to discover the set of causal models over $\boldsymbol{X}$.*

Before we address this problem, we take a step back to recall how we did causal discovery in an i.i.d. setting and use that to introduce the concepts and assumptions that we build on.

4.2.2 Causal Discovery for I.i.d. Data

For now, consider the case of a single context without selection bias. We can specify a causal model over the variables X by a directed acyclic graph (DAG) $G = (X, E)$ with node set X and edges $(i, j) \in E$ whenever the variable X_i is a cause of X_j (Pearl, 2009). To denote the set of direct causes of X_j we write pa_j where we leave G implicit. Together with the network structure in G, we assume a structural causal model over the variables, where each effect is generated from its causes through a causal function or mechanism f_j,

$$X_j = f_j(pa_j, N_j)$$

where N_j is a noise variable implicit in G with $N_j \perp\!\!\!\perp X_j$.

To ensure identifiability we assume causal *sufficiency*, which states that no latent variable jointly causes any of the observed variables, as well as the *causal Markov* and *faithfulness* conditions, which together imply that edge separations in the graphical model G correspond to independence constraints in the observed distribution P. Under these assumptions, it is well known that identifiability holds up to the Markov Equivalence Class (MEC) of G (Hauser and Bühlmann, 2013).

As we also already show in case of Orion, identification of causal directions beyond the MEC is possible using additional information about how the system reacts to interventions (Hauser and Bühlmann, 2014; Zhang et al., 2017; Mameche et al., 2023). In the absence of such information, we need to make additional assumptions, such as restricting the functional dependencies f to nonlinear functions with additive noise (Bühlmann et al., 2014; Hoyer et al., 2009a; Marx and Vreeken, 2021). As an example of the latter, a family of methods build on the algorithmic framework of causation (Janzing and Schölkopf, 2010b) and derive consistent scoring criterion that can be used for causal discovery within a given class of functional models. This is the approach we will continue to follow in this chapter.

Continuing to build on our Information theoretic approach to causal discovery we assume, throughout this chapter, a given any MDL-based score L that decomposes as in Eq. (4.1) (as we already define in Chapter 2), and is consistent in the sense that it allows estimating a DAG $G \sim G^*$ that is Markov equivalent to G^* in the limit, $\lim_{n\to\infty} P(\hat{G} \sim G^*) = 1$ for i.i.d. data with sample size n.

$$
\begin{aligned}
M^* &= \operatorname*{argmin}_{M \in \mathcal{M}} L(\boldsymbol{X}^n, M) \\
&= \operatorname*{argmin}_{M \in \mathcal{M}} \left(L(M) + \sum_{i=1}^{m} L(X_i^n \mid \mathrm{pa}_i, M) \right) \\
&= \operatorname*{argmin}_{M \in \mathcal{M}} \left(L(M) + \sum_{i=1}^{m} L(\epsilon_i) \right)
\end{aligned} \tag{4.1}
$$

We refer to Chapter 2 for definitions of L in a multivariate setting and a consistent algorithm for discovering G in from an i.i.d. data distribution. As consistency results and practical algorithms have only been explored in the i.i.d. case (Mian et al., 2021) or interventional data (Mameche et al., 2023; Mian et al., 2023b), we turn to episodic data here.

4.3 Theory

Now that we have refreshed our knowledge of causal discovery on i.i.d. data, we introduce the concepts and assumptions for episodic data. We first define a causal model over episodic data and concretize the assumptions that we need and then provide theoretical guarantees to show that a consistent score can be used to discover causal networks in a setting with unknown contexts.

4.3.1 Causal Model

Unlike previous chapters where we only considered searching for a single causal graph, our causal model now comprises of a set of causal DAGs $\mathbf{G} = \{G_1, \ldots, G_R\}$ over a common set of variables $X \cup \{S\}$, where X are measured, continuous random variables of interest, and S is an unmeasured categorical variable with values $S = \{s_1, \ldots, s_K\}$. Each DAG G_r is a causal model over X^r, i.e., it describes the causal relationships in all episodes from a given context C_r. The additional variable S^r models that certain observations may be missing in each episode.

To do so, we extend upon a missingness framework commonly used to handle selection bias (Rubin, 1976; Bareinboim and Pearl, 2012). To explain, consider the nth observation, where we represent S using a one-hot encoding,

$$
\left(X_1^{(n)}, \ldots, X_m^{(n)}, s_1^{(n)}, \ldots, s_K^{(n)}\right)
$$

where we omit the dependency on the context to avoid clutter. Above, $X^{(n)}$ is associated to indicators s_k where $s_k = 1$ if $X^{(n)}$ is *observed*, else $s_k = 0$ if it is

missing in a distribution k. We obtain K biased distributions $P(X \mid S = s_k)$ where any number of samples of the support of X are missing.

Exactly which samples are observed could depend on X; in Fig. 4.1, for instance, $S_+ = \diamond$ holds for the temperature range $X_1 \geq 10$. In general, we assume that any unknown mechanism assigns S,

$$S = g(X, N_s), \qquad N_s \perp\!\!\!\perp S \ ,$$

where g maps each sample to an assignment of S using input X, which could be noisy, through N_s. We therefore include S in the causal model together with edges $X_j \to S$ for all X_j, and assume that S is a sink node. We do the above in any context, that is, include a sink node S^r in G_r. We assume causal sufficiency over $X^r \cup \{S^r\}$. To summarize, we work with the following causal model.

Assumption 4.1 (Causal model with contexts and selection) *Our causal model is given by a set of DAGs* $\mathbf{G} = \{G_1, \ldots, G_R\}$ *over* $\boldsymbol{X} \cup \{S\}$ *from a finite number of contexts* R *such that in context* C_r*, each observed variable* X_j *is generated as*

$$X_j^r = f_j^r(pa_j^r, N_j^r), \qquad N_j^r \perp\!\!\!\perp X_j^r \ , \tag{4.2}$$

where pa_j^r *denote the causal parents of* X_j^r *in* G_r *and* N_j^r *is an independent noise term. The latent variable* S *is generated as*

$$S^r = g^r(X_j^r, N_s^r), \qquad N_s^r \perp\!\!\!\perp S^r \ . \tag{4.3}$$

Equation (4.2) above describes an unbiased generating process where each variable X_j is a function of its causal parents pa_j and noise N_j. In addition, the mechanism g with noise N_s generates S as shown in Eq. (4.3). This generating process happens independently in each context. We further assume that episodes result from *conditioning* on a specific value of the unobserved selection variable.

Assumption 4.2 (Episodic data) *Under the causal model in Assumption 4.1, after generating an unbiased distribution* $P^r(X, S)$ *from the DAG* G_r *in each context* C_r*, all episodes* E *coming from context* $C(E) = C_r$ *have distribution* $P^r(X \mid S = s_k)$ *for some specific* $s_k \in \{s_1, \ldots, s_K\}$.

With no assumption on the selection mechanism g, number of contexts R, or number of selection regions K, our model can encompass general cases of episodic data. This invariably also makes it more challenging to discover the causal model from data. To do so, nevertheless, recall the algorithmic Markov condition first described in Chapter 2 and note that for our case we can rewrite it as follows.

Postulate 4.1 (Algorithmic Markov Condition) *Under Assumptions 4.1 and 4.2, a set of causal DAGs* $\mathbf{G} = \{G_1, \ldots, G_R\}$ *is only admissible as the causal hypothesis over* X *and* S *if*

$$K\big(P\big(X \cup \{S\}\big)\big) \stackrel{+}{=} \sum_{r=1}^{R} \sum_{j=1}^{m} K\big(P^r(X_j \mid pa_j)\big) + K(P^r(S \mid X)) \\ \stackrel{+}{=} K\big(P(X)\big) + K\big(P(S \mid X)\big)$$

where $\stackrel{+}{=}$ *holds up to an additive constant.*

As S is not included in any parent set, we can in principle consider the complexity of, and hence causal structure over, X *independently* of the complexity of S. This motivates the idea of using a consistent scoring criterion to find the causal structure over X in each context.

As a complication, we hereby need to discover the number of contexts. Suppose we obtained data $\mathbf{D}_n$ accumulated over n episodes. There could be any number R of different causal models, with $1 \leq R \leq n$. Thus, we need to consider any partition of our samples into R disjoint sets, which we write as $\Pi(\mathbf{D}_n) = \{X^1, \ldots X^R\}$. In each set, we propose discovering the causal DAG using the consistent score $L(X^r; G)$, and overall find the partition minimizing this score.

To summarize, our objective is as follows.

Problem Statement. *Given variables* X *and data* $\mathbf{D}_n$ *over* n *episodes, we aim to discover the partition* $\Pi(\mathbf{D}_n)$ *of* $\mathbf{D}_n$ *into contexts and the causal model* $\hat{G}r$ *in each context minimizing*

$$\min_{\Pi(\mathbf{D}_n)} \sum_{r=1}^{|\Pi(\mathbf{D}_n)|} \min_{G_r} L(X^r; G_r) \; . \tag{4.4}$$

where we write X^r *for the data in the* r*-th set of* $\Pi(\mathbf{D}_n)$*.*

This leaves us with two questions; first, ensuring that the above is a consistent way of identifying the causal model, and second, how to efficiently minimize it in practice.

4.3.2 Asymptotic Guarantees

We first want to establish conditions under which L can be used in a consistent way to discover the causal DAGs in all contexts.

This revolves around whether the *biased* distributions in each episode eventually allow us to estimate the relevant distributions in Postulate 4.1 in an *unbiased* way so that we can apply Eq. (4.4). That is, estimation of each causal

mechanism should not depend on the selection variable. We therefore need to make the following assumption.

Assumption 4.3 (Ignorability) *Under the causal model in Assumption 4.1 and given* $\mathbf{D}_d$ *over* d *episodes, in each context* C_r*, we assume the following* ignorability *of selection bias,*

$$X_j^r \perp\!\!\!\perp S^r \mid \mathbf{Z}^r$$

for each X_j^r *and conditioning set* $\mathbf{Z}^r \subseteq X^r \setminus \{X_j^r S^r\}$.

Examples of when the above holds are cases known as Missing At Random (MAR) or Missing Completely At Random (MCAR) (Rubin, 1976; Bareinboim and Pearl, 2012; Bareinboim et al., 2014), for example, when a biased $P(X \mid S = s_k)$ is a uniform sample from the population $P(X)$. A more realistic case is the one in Fig. 4.1 where the selection mechanism depends on temperature X_1. We can see that episodes from the cold season $P(X \mid S_- = \circ)$ indeed do not allow an unbiased view of the causal mechanism, however once we obtain enough episodes from both S_-, S_+ then ignorability holds. More generally, we ensure via Assumption 4.3 that we eventually obtain enough samples from the support of X.

With this, we can show that an MDL-based score L can be used for causal discovery with unknown contexts. For ease of exposition, we separate out the case of a single context with one causal model and show it first.

Lemma 4.1 (Consistency of L **for a single causal model)** *For the causal model in assumption 4.1 and assumption 4.2 with R=1 and data* $\mathbf{D}_n$ *over* n *episodes covering each value* s_k *of* S*, with a consistent scoring criterion* L *that decomposes as in Eq. 4.1 then* L *is consistent,*

$$\lim_{n\to\infty} P(\hat{G} \sim G^*) = 1 \,.$$

Using the result above, we move to our full causal model with multiple contexts.

Theorem 4.2 (Consistency of L **in the episodic setting)** *For the causal model in Assumption 4.1 and given data* $\mathbf{D}_n$ *over* n *episodes as in Assumption 4.2. Under Assumption 4.3, a consistent scoring criterion* L *that decomposes as in Eq. 4.1 remains consistent,*

$$\lim_{|D_n|\to\infty} P(\hat{G}_r \sim G_r^*) = 1 \qquad \textit{for all } r \in \{1, \ldots, R\} \,.$$

This, however, does not make it obvious how to apply L in practice. First, note that the result relies on enough episodes being observed so that selection

is ignorable, that is, we did not yet address how to deal with non-ignorable selection at each time point when we only observed a subset of episodes. Second, even when observing enough episodes, searching over the space of DAGs to minimize L as in Eq. 4.4 is prohibitive even for a single causal model due to the super-exponential search space over DAGs (Chickering et al., 2004). While there exist greedy algorithms to do so, such as the MDL-based GLOBE, applying such methods to any partition of the data with unknown number of contexts is not favorable as it could violate the i.i.d. assumption required for these methods. We address these issues in the following and subsequently propose an algorithm for causal discovery over episodic data.

4.4 The Continent Algorithm for Online Causal Discovery

Using the concepts defined in Section 4.3, we now introduce our algorithm to discover causal networks from episodic data. We show under which conditions our proposed algorithm, CONTINENT, entails correctness guarantees and can therefore find the correct causal network(s) for episodic data given enough evidence.

4.4.1 Overview

To motivate our algorithm setup, let us revisit our motivating example in Fig. 4.1 showing episodes obtained in winter $D^{(1)}$, spring $D^{(2)}$, summer $D^{(3)}$, and autumn $D^{(4)}$. We consider a fixed number of seasons, here S_+, S_-. All episodes $D^{(1)}$-$D^{(4)}$ shown come from a context C_1 but any number of future episodes could arrive from a different C_2.

Given a learner $\mathcal{A}$ for greedy DAG search with a consistent scoring criterion L, we aim to discover the underlying causal DAG G_1 over $D^{(1)}$-$D^{(4)}$, and possibly add a causal model G_2 if future episodes from a different C_2 arrive. Applying $\mathcal{A}$ to all episodes at each time point is not only impractical, but may also not be consistent given that selection bias is not ignorable until all episodes arrived. Instead, we propose an algorithm CONTINENT that maintains plausible causal models $\mathbf{G} = \{G_1, \ldots, G_R\}$ at each time t_i and uses a strategy for *updating* $\mathbf{G}$ when a new episode $D^{(i+1)}$ arrives.

MODEL UPDATING In our example, say that we obtained episodes $D^{(1)}$-$D^{(3)}$ and the current causal model is $\mathbf{G} = \{G_1\}$. As we already observed episodes from both seasons S_+ resp. S_- we likely already learned an unbiased model G_1. As the autumn episode $D^{(4)}$ arrives, we want to assign it to G_1 without re-learning the causal model from scratch. To this end, we propose using a two-sample testing procedure $\mathcal{T}$ for deciding whether a given episode matches

Algorithm 4.1: CONTINENT $(E, \mathcal{A}, \mathcal{T})$

input : episodes E arriving over time, residual test $\mathcal{T}$, causal discovery algorithm $\mathcal{A}$ with score L
output: causal model $\mathbf{G} = \{G_1, ...G_R\}$

$\mathbf{G} \leftarrow \{\}$
$\tau \leftarrow 0$
while a new episode $D^{(i)}$ arrives **do**
  $\mathbf{G} \leftarrow$ UPDATE $(\mathbf{G}, D^{(i)}, \mathcal{A}, \mathcal{T})$
  $\tau \leftarrow \tau + 1$
  if $\tau \geq \tau_{\max}$ **then**
    $\mathbf{G} \leftarrow$ MERGE $(\mathbf{G}, \mathcal{A}, \mathcal{T})$
    $\tau \leftarrow 0$
  end
end
$\mathbf{G} \leftarrow$ MERGE $(\mathbf{G}, \mathcal{A})$
return G

an existing causal model. Here, after checking with $\mathcal{T}$ that $D^{(1)}$-$D^{(4)}$ can be stacked we combine the data $D^{(1)}$-$D^{(4)}$ and keep the model G_1 as is.

On the other hand, say episode $D^{(5)}$ from a different context C_2 arrives[1] and $\mathcal{T}$ decides that it does not match any current causal model. Then we apply the learner $\mathcal{A}$ to learn a new model G_2 over $D^{(5)}$ and add it to our set of models, $\mathbf{G} = \{G_1, G_2\}$.

Note that the above assumes that we already learned an unbiased causal model over the available episodes. We also need to consider the case where a causal model is biased, such that we need to update it after *merging* data from multiple episodes.

MODEL MERGING Say that we observed episodes $D^{(1)}$-$D^{(2)}$ to learn a causal model G_0. From the winter seasons S_- alone, it appears that X_1, X_2 are uncorrelated, hence G_0 is biased. When $D^{(3)}$ from summer season S_+ arrives, we need to *merge* the data to the previous episodes and learn a new model G_1.

To do this, we attempt merging data over multiple episodes at regular time intervals. We again apply $\mathcal{T}$ to check whether a merge is possible, and if so, check whether merging any two causal models results in an improved model,

[1] This could be e.g. readings obtained from a different geographical region where causal mechanism between X_1 and X_2 is different/non-existent.

Algorithm 4.2: TestResidualEq $(G_r, D^{(i)}, D, \mathcal{T})$

input : causal model G, episode $D^{(i)}$, data D, residual test $\mathcal{T}$
output: test result
foreach X_j with parent set $\mathbf{Z}$ in G **do**
  $p_j \leftarrow \mathcal{T}$.Test$(H_0 : P^D(X_j \mid \mathbf{Z}) \equiv P^i(X_j \mid \mathbf{Z}); \alpha)$
end
$\mathbf{p} \leftarrow \mathcal{T}$.Correct$(\{p_1, ...p_m\})$
if $\mathcal{T}$.Significant$(\mathbf{p})$ **return** True **else return** False

judging by our score L. As stacking may be sufficient when we already gained sufficient evidence for a candidate model, in practice, we attempt merging at regular time intervals using pre-specified tolerance parameter τ_{max}.

Combining the model updating and model merging described above, we have our proposed approach, Continent. We show the pseudocode of Continent in Alg. 4.1. We maintain a set of models $\mathbf{G}$ throughout, where we associate each $G \in \mathbf{G}$ to a dataset D of episodes, initially empty (line 1). As new episodes arrive, we update $\mathbf{G}$ at each time step using the Update function (line 4). In short, it checks using hypothesis testing whether a new episode $D^{(i)}$ matches the data D under an existing model, in which case we stack the datasets $D^{(i)}$ and D; else we apply $\mathcal{A}$ to $D^{(i)}$ to discover a new model G_i which we add to $\mathbf{G}$. We show our hypothesis test in Alg. 4.2, and Update in Alg. 4.3.

After a pre-specified number of episodes, we attempt merging existing models (line 7), with a tolerance parameter τ keeping track of the time since a merge last happened (line 8). In essence, Merge performs pairwise comparison of models G, G'. If appropriate, it learns a new model $G_\cup$ after pooling the resp. datasets D, D' of the pair. During the algorithm, we only allow such a merge if $\mathcal{T}$ marks the residual distributions of D, D' as compa‘tible, for which we again apply our hypothesis test in Alg. 4.2. We provide the pseudocode for the Merge in Alg. 4.4.

Our alternation of updating and merging continues as long as new episodes arrive. We conclude with a final merge (line 11). Compared to merge steps throughout our algorithm which we protect by $\mathcal{T}$, we consider all remaining possible merges of model pairs G, G' in this step given that no more episodes arrive (line 11).

4.4.2 Consistency

Naturally, we want to make sure that our adaptive strategy is consistent. At any time point $t^{(i)}$, however, we only have access to a subset of the episodes so

Algorithm 4.3: UPDATE $(\mathbf{G}, E, \mathcal{A}, \mathcal{T})$

input : episode E,
causal model $\mathbf{G}$,
causal discovery algorithm $\mathcal{A}$ with score L,
residual test $\mathcal{T}$
output: updated causal model $\mathbf{G}$

```
accepted ← FALSE
foreach G_r over data D in G do
    if TESTRESIDUALEQ (G_r, E, D, T) then
        accepted ← TRUE
        D ← D.STACKDATA (E)
    end
end
if not accepted then
    G ← A.LEARN(E)
    G = G ∪ {G}
end
return G
```

that ignorability in Assumption 4.3 unlikely holds, and hence any causal model inferred using $\mathcal{A}$ may be incorrect. Nevertheless, we need to avoid merging episodes with different underlying models. We now show that we can do so without knowing the true models. To do so, we assume a hypothesis test $\mathcal{T}$ testing

$$H_0 : P^1(X_j \mid \mathbf{Z}) \equiv P^2(X_j \mid \mathbf{Z})$$

for a given variable X_j, conditioning set $\mathbf{Z}$ and two datasets P^1, P^2. Given any causal DAG, we test H_0 for each variable given its estimated parent set and include a multiple testing correction, as shown in Alg. 4.2. We can show that our updating strategy protected by this test is consistent under the following condition.

Assumption 4.4 (Detectable selection) *We assume that selection* detectable *for a variable X_j and pair of contexts C_r, C_r' meaning*

$$P^r(X_j \mid pa_j) \neq P^{r'}(X_j \mid pa_j)$$
$$\Rightarrow P^r(X_j \mid pa_j, S = s_k) \neq P^{r'}(X_j \mid pa_j, S = s_k)$$

holds for each value s_k of S.

Algorithm 4.4: MERGE $(\mathbf{G}, \mathcal{A}, \mathcal{T})$

input : causal model $\mathbf{G}$,
causal discovery algorithm $\mathcal{A}$ with score L,
residual test $\mathcal{T}$
output: updated causal model $\mathbf{G}$

repeat
  foreach G over data D in $\mathbf{G}$ **do**
    $D^\star \leftarrow D$
    $G^\star \leftarrow G$
    $L^\star \leftarrow G.\text{SCORE}(D)$
    foreach G' over data D' in $\mathbf{G}$ not seen yet **do**
      if not TESTRESIDUALEQ $(G', D, D', \mathcal{T})$ **continue**;
      $D^\cup = D \cup D'$
      $G^\cup \leftarrow \mathcal{A}.\text{LEARN}(D^\cup)$
      $L^\cup \leftarrow G^\cup.\text{SCORE}(D^\cup)$
      if TESTSCOREDIFF $(L^\cup, L^\star)$ **then**
        $D^\star \leftarrow D^\cup$
        $L^\star \leftarrow L^\cup$
        $G^\star \leftarrow G^\cup$
      end
    end
    if $G^\star$ is not G **then**
      replace corresponding G, G' with $G^\cup$ in $\mathbf{G}$
    end
  end
until convergence;
return G

Unlike ignorability in Assumption 4.3 which requires full independence of the causal mechanism and selection mechanism, i.e. ensures that we can estimate the causal mechanism for each variable in a fully unbiased way, Assumption 4.4 only requires that distribution differences of $P(X)$ hold also in the biased distribution $P(X \mid S = s_k)$. Given that the latter are subsamples of the overall distribution, this is reasonable in practice. With this, we can show that our updating strategy is consistent.

Theorem 4.3 (Consistency of updating using $\mathcal{T}$) *With discrepancy test*

$\mathcal{T}$ we will never merge a new episode $D^{(i+1)}$ with a set $\hat{X}^r$ from an incorrect context where $C(D^{(i+1)}) \neq C(E)$ for some $E \in \hat{X}^r$.

This shows that our updating step is safe in the sense that we always discover subsets of the correct contexts. When we observed all episodes, we can also recover the exact sets of contexts if ignorability holds, based on Thm. 4.2.

Corollary 4.4 (Consistency of Continent) *Given a consistent DAG search algorithm $\mathcal{A}$ and score L, under assumption 4.3 our algorithm is consistent, so that*

$$\lim_{|D_n| \to \infty} P(\hat{G}_r \sim G_{r*}) = 1 \qquad \text{for all } r \in \{1 \dots, R\}$$

holds after we obtain n episodes D_n and perform the merge step.

As the final step in this section, we address practical considerations around our algorithm.

4.4.3 Instantiation

We conclude this section by giving details on the components of Continent .

Causal Discovery Algorithm $\mathcal{A}$ We assume a score-based causal discovery algorithm $\mathcal{A}$ that allows discovering a causal DAG G from an i.i.d. dataset D. While in principle, this could be any score-based method with a consistent scoring criterion L decomposing according to Eq. (4.1), we use an MDL-based approach in our practical instantiation as it allows for a principled way for model comparison. We instantiate $\mathcal{A}$ with Globe (Mian et al., 2021) which is an efficient algorithm for discovering causal networks. It models causal functions through non-parametric multivariate regression with additive noise.

Residual Test $\mathcal{T}$ Our method can also work together with any hypothesis test $\mathcal{T}$ for differences in conditional distributions under a causal model. As Globe models causal functions through non-parametric spline regression, a natural choice is testing residual distributions under a given model for equality. As we apply a test per each variable, we perform Bonferroni correction to obtain a p-value from the test results $\{p_1, \dots, p_m\}$. Unless otherwise stated, we apply the non-parametric Kolmogorov-Smirnov (AN, 1933; Smirnov, 1948) test in our evaluations.

4.5 Evaluation

Since to the best of our knowledge, there is no specific algorithm designed for causal discovery from continually arriving episodic data, we look at the nearest possible modifications of existing algorithms for comparison. As baseline we compare to Globe (Mian et al., 2021), Resit (Peters et al., 2014) and Ges (Chickering, 2002; Ramsey et al., 2017). We modify these algorithms as follows — we first learn a causal network over each individual incoming episode of data, and then take a union over the edges. This is correct, under the assumption underlying each of these approaches, that each episode comes from the *same* causal network (Mian et al., 2023b). We also compare to multi-environment causal discovery approaches such as the JCI-framework (Mooij et al., 2016) using the Pc algorithm (Spirtes et al., 2000a), the Orion algorithm Mian et al. (2023b), as well as Multi-Group Lingam (Lingam) (Shimizu, 2012). The latter three approaches, however, require that all episodes are available to learn a causal network. Hence, we provide all episodes in one go to these approaches. This constitutes an advantage as they can learn from complete data from the very start.

To measure the quality of the predicted causal structures we use the Structural Hamming Distance (Shd) (Kalisch and Bühlmann, 2007) , the Structural Intervention Distance (Sid) (Peters and Bühlmann, 2015), as well as Orientation-F1 score over learned networks. Shd counts the number of edges where the predicted causal network differs from the true causal network, Sid counts pairs of variables for which intervention estimation differs across predicted resp. true causal network and F1 score allows us to see how accurately are the edges oriented in the learned network. Next, we discuss results over both synthetic and real-world data.

4.5.1 Synthetic Data

For each of the proposed experiment setups, we generate random graphs using Erdős-Rényi model for network sizes $d = \{5, 10, 15\}$, and generate data for effects using functions of the following form, $X_i = \sum_{x \in pa_i} f(x) + \mathcal{N}_i$, where $f(x)$ is either a polynomial function or a combination of sine and cosine functions defined over each parent $x \in pa_i$ of X_i, and $\mathcal{N}_i$ is either Gaussian or Uniform. For each graph/function combination, we generate a total of $10,000$ samples and then split them into 10 episodes of size 1000 each. We *transmit* these episodes to each algorithm one at a time. After each episode, we note the updated causal network for each of the methods. As PC and Lingam are provided all episodes together, we only measure performance over the final network. Primarily, we investigate over the following questions:

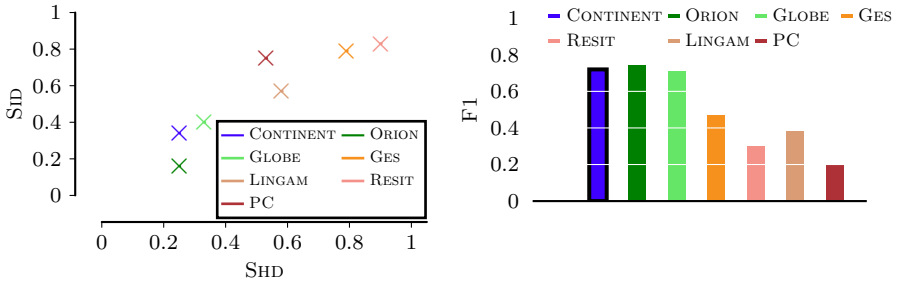


Figure 4.2: Normalized Shd and Sid [Left, Closer to origin is better] and Orientation F1 [Right, Higher is better] for networks learned over episodic data with selection bias.

Q1 Can Continent reliably discover causal networks when the incoming episodes come from the same underlying causal network?

Q2 How well does Continent perform when episodes contain unknown interventions?

Q3 Can Continent identify causal networks from episodic data containing *different* causal mechanisms?

Q4 How does Continent 's performance change over time as episodes arrive?

Continent is designed without the assumption that each data comes from the same underlying causal network, and therefore maintains a list of candidate networks for groups of episodes. For comparability to other approaches, we *force* Continent to predict a single causal network for cases $Q1$ and $Q2$ by taking a union over the edges in candidate models as this should result in the correct causal network in the limit (Mian et al., 2023b). We further provide an analysis of the individually learned causal networks for evaluation in $Q3$. We release all our code and data for research purposes[2]. Next, we show results for each of the four questions.

Q1. Identical Networks We first test all methods on the cases where each incoming episode comes from the same underlying causal network, both for i.i.d. as well as selection-bias. Interesting for us is the latter where episodes can contain selection bias. We generate this case by choosing a variable at random from our dataset and sorting the entire data over that variable before splitting the data into episodes and transmitting it. We show the results for this in Fig 4.2 where we see that Continent shows superior performance to the competition. It is second in terms of Sid only to Orion, which can be attributed to the latter having the full picture from the beginning. Nonetheless,

[2]https://eda.rg.cispa.io/prj/continent/

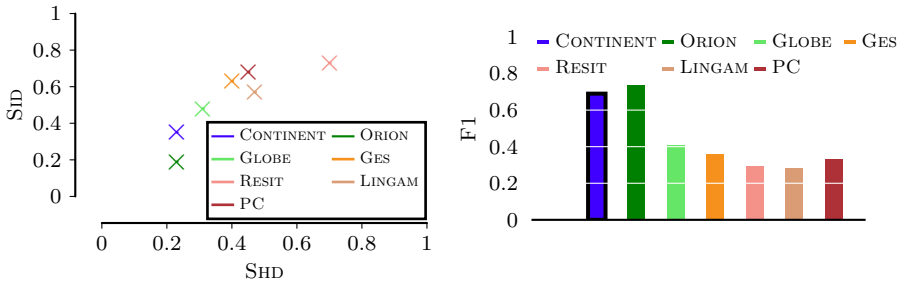


Figure 4.3: Normalized SHD and SID [Left, Closer to origin is better] and Orientation F1 [Right, Higher is better] for networks learned over episodic data with unknown interventions. CONTINENT gives almost on-par performance to ORION, which gets all data in one go.

we see that both ORION and CONTINENT have the same SHD, indicating that they likely find the same underlying causal skeleton and the difference in SID is therefore due to CONTINENT finding it harder to orient edges in some of the cases. Moreover, among the methods that do not get access to full data from the start, CONTINENT not only discovers causal network structurally closer to the ground truth, but also clearly performs well when orienting the edges as can be seen by the F1 score in Fig. 4.2.

Q2. INTERVENTIONS After our sanity check using i.i.d. data and dominant performance over data with selection bias, we level up the difficulty by introducing episodes that contain interventions. To do so, we generate 3 datasets. The first dataset is observational, whereas for the other two, we select a subset of at most $\log_2(d)$ variables and perform a *do*-intervention (Pearl, 2009) on that subset, before generating the data. This gives us data sampled from three different distributions. We further split each of these datasets into episodes before transmitting them. We never provide information about these interventions to any of the methods beforehand.

We show the results of this experiment in Fig. 4.3, where we see that GLOBE already degrades significantly as can be seen by the drop in F1 score, CONTINENT's performance does not degrade compared to the setup in $Q1$. CONTINENT, in fact, continues to clearly outperform the baselines for episodic discovery and is almost on par with ORION when it comes to methods that get access to complete data.

Experiment	Nodes	Shd	Sid	F1	Shd	Sid	F1
		CONTINENT			ORION		
Interventions	5	**0.23**	0.15	0.68	0.24	**0.14**	**0.70**
	10	0.25	0.34	0.54	**0.24**	**0.30**	**0.62**
	15	0.29	0.50	0.43	**0.26**	**0.45**	**0.49**
Mechanism Changes	5	**0.21**	**0.15**	**0.74**	0.54	0.23	0.46
	10	**0.26**	**0.38**	**0.64**	0.56	0.52	0.41
	15	**0.36**	**0.65**	**0.41**	0.6	0.7	0.4

Table 4.1: Normalized SHD [Lower is better], normalized SID [Lower is better] and Orientation F1 [Higher is better] for networks predicted by CONTINENT and ORION for held-out episodes for interventional data as well as mechanism changes. CONTINENT consistently performs well across both settings. ORION works well in case of interventional data where there is a single underlying causal network, but fails when incoming episodes come from networks with different causal mechanisms.

Q3. CHANGING MECHANISMS. As the next challenging step, we introduce episodes containing different causal networks/mechanisms over the same variables. To evaluate CONTINENT in this setting, we additionally generate a hold-out set of episodes that we do not learn over. Once CONTINENT has learned over the training episodes, we try to *predict* the causal network for hold-out episodes, without learning it explicitly, using the existing learned models. We do so by simply taking the model that compresses this hold-out episode best (ref. Eq. (4.1)) and compare the predicted network to the ground truth. Note that this rules out using any of our competitors except ORION as they do not maintain a list of plausible causal networks. For ORION we simply check at the end, which of the learned interventional sub-networks compress data best and use that as the predicted causal network.

We show the results in Table. 4.1 where we observe that CONTINENT shows competitive performance for the case where each episode comes from same underlying causal network but different interventional sub-networks, and is infact superior to ORION when incoming episodes come from causal structures that may change across episodes. In contrast ORION, being biased towards finding a single global causal network, can not handle changing mechanisms. Furthermore, we see that for the more challenging setting with changing mechanisms, CONTINENT can find a reasonable skeleton (lower SHD) but conflicting mechanisms may cause it to get edge directions wrong more often (higher SID). Nevertheless, we see that CONTINENT's performance does not degrade, even in this challenging case.

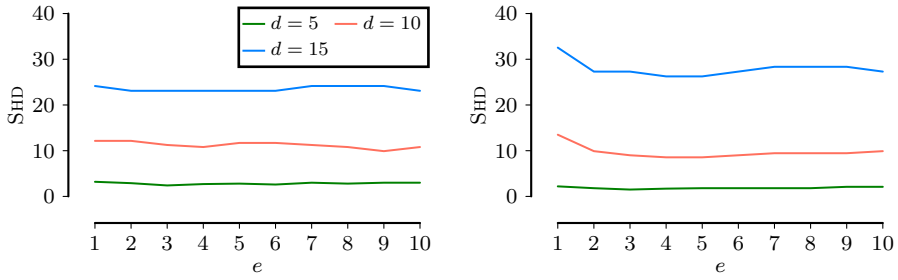


Figure 4.4: [Lower is better] Change in SHD over increasing number of episodes e for data with selection bias (left) and unknown interventions (right) for graph sizes $d = \{5, 10, 15\}$.

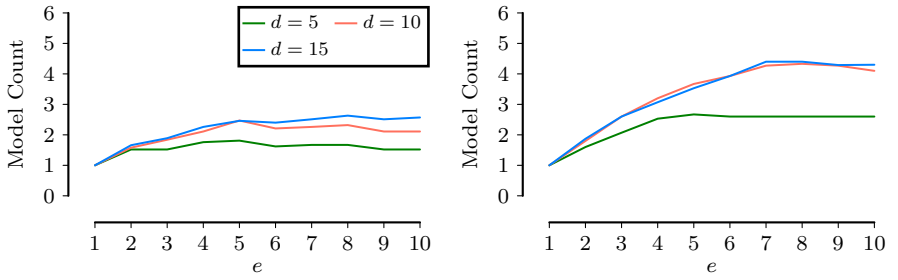


Figure 4.5: Model Count over increasing episodes e for data with selection bias (left) and data with (unknown) interventions (right) for graphs of size $d = \{5, 10, 15\}$. There are 1 resp. 3 true underlying models for bias resp. intervention cases.

Q4. PERFORMANCE OVER TIME. We measure how the individual models present inside CONTINENT evolve over time. To that end, we show how the SHD (Fig. 4.4) as well as the model count (Fig. 4.5) progresses as we receive new episodes. For the case of SHD, we find that CONTINENT always ends up with a lower SHD at the final episode, than the one it starts with, this effect is more profound for networks of size $d = 15$ than $d = 5$ as it might be harder to identify the correct network over a larger number of variables in the beginning. We see that CONTINENT is able to improve as the number of episodes increase. For data with selection bias, we see that CONTINENT keeps on average 2 models throughout the learning as shown in Fig. 4.5. More interestingly CONTINENT ends up converging to almost 4 models for interventional data as shown in Fig. 4.5, which is very close to the the actual number of different networks (3) present across episodes.

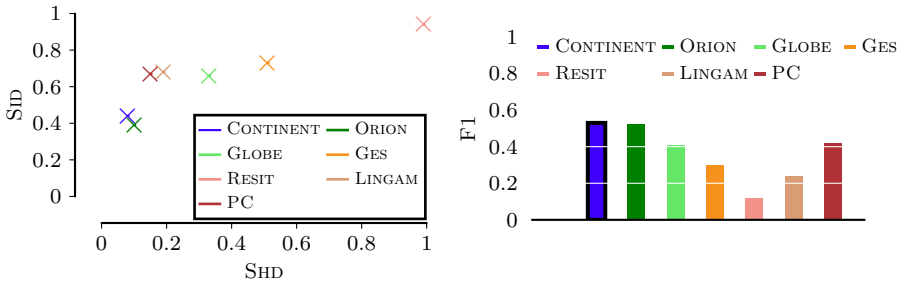


Figure 4.6: Normalized Shd and Sid [Left, Closer to origin is better] and Orientation F1 [Right, Higher is better] for networks learned Reged Lung cancer gene expression dataset.

4.5.2 Lung cancer gene expression data

After measuring the efficacy of our approach using synthetic data, we turn to (pseudo) real-world Reged dataset (Statnikov et al., 2015) containing $20,000$ samples over 500 variables for lung cancer gene-expressions. We split the samples into ten non-overlapping episodes and consider two non-overlapping networks of sizes $d = 5, 15$ within the ground truth network and run a total of 10 experiments as follows. First, we randomly choose a subset of 5 episodes, merge them and introduce selection bias over stacked data akin to $Q2$, before splitting it back. We show the results for Reged dataset in Fig. 3.2 where we see once again Continent comes out on top of the baselines.

4.6 Discussion

Our interest in this work is determining causality when data arrives progressively over time in multiple episodes. Each representing sub samples of the population or subregions of the data that need to be pooled *together* to avoid bias. At the same time, we address that the causal relationships may not be stationary over time, and treat episodes from different contexts under a *seperate* causal model. To address this setting, we propose a causal model over a set of latent contexts leading to a set of different causal networks, as well as model episodic bias through a hidden selection variable. Unlike Globe and Orion, Continent, relaxes the causal sufficiency assumption slightly in that it does not require selection variables to be explicitely observed, and can still discover the correct causal networks for the observed variables provided that the hidden variables are all sink nodes. This, however, does not address the case when the hidden selection variable is in fact a source node for any two

observed variables. For such unobserved confounding, theoretical guarantees do not directly hold and it remains an involving line of future work to address the more general case of (hidden) selection variables.

While we have shown that information-theoretic scoring criteria remain consistent for our defined model in the limit if we obtain sufficiently many episodes so that selection bias becomes ignorable, practical issues might still exist. This includes merging of two models from different contexts due to finite-sample behavior of the residual test being used. Since CONTINENT does not employ a SPLIT step, a wrong merge can mislead the search process. While not trivial, including steps to identify and rectify incorrect MERGE operations is one of the lines of future work.

CONTINENT is capable of learning causal networks in a somewhat realistic setting where episodes arrive one by one over time with non-ignorable selection. It does this by maintaining a set of causal networks over all episodes and incorporates new episodes into the model, using a residual testing strategy to avoid combining episodes from different contexts. We can, however, address non-ignorability further, by using correction or extrapolation techniques, which is an ongoing continuation of our research on this topic.

4.7 CONCLUSION

In this chapter we considered the case of causal discovery over datasets that perennially arrive over time. We showed why the conventional approached in causal discovery that expect a single, observational dataset can not be applied to such scenarios. Namely, not only because each episode may be a *biased* sample of the population but also because multiple episodes may *differ* in the causal interactions underlying the observed variables. We addressed these issues using notions of episodic selection bias and context switches and showed under which conditions we can apply information-theoretic scoring criteria for causal discovery in episodic settings while preserving consistency. To discover the causal model progressively over time in practice, we proposed the CONTINENT algorithm which, taking inspiration from continual learning, discovers the causal model in an online fashion without having to re-learn the model upon arrival of each episode.

Through extensive experimental evaluation, we showed that our method performs reliably in the presence of selection bias, under unknown interventions, and even when different causal models underlie the data generating process, which to our knowledge no existing methods can address.

Chapter 5

Privacy-preserving Federated Causal Discovery

In privacy sensitive applications such as healthcare, we cannot pool data. Causal discovery in such critical applications, therefore, comes with its own set of additional challenges and constraints. In these cases we usually have multiple sites each with their own private data. Learning causal networks over such data presents a challenging setting where we don't just want to discover the underlying causal network in a federated manner, we can also not compromise on privacy. We therefore consider the problem of discovering a global causal network over distributed datasets with a fixed set of variables — in a privacy preserving manner.

While a plethora of approaches for discovering causal networks are designed for single datasets (Spirtes et al., 2000a; Chickering, 2002; Shimizu et al., 2006; Peters et al., 2014; Huang et al., 2018), most state-of-the-art causal discovery approaches that can work with multiple datasets, including ORION proposed in Chapter 3 and CONTINENT proposed in Chapter 4, require that we have access to the data (Mooij et al., 2016; Zhang et al., 2017; Mian et al., 2023b). This requirement makes them inapplicable to our new privacy-critical setting. Multi-dataset approaches that do not require data to be pooled, work only for a single target variable (Peters et al., 2016) at a time — rendering them inapplicable for overall structure discovery, or place strict assumptions on the causal mechanisms that are unlikely to hold in practice (Shimizu, 2012; Ghassami

This chapter is based on Mian, Kaltenpoth, and Kamp (2022) and Mian, Kaltenpoth, Kamp, and Vreeken (2023a).

et al., 2017).

On the other hand, state of the art federated learning approaches allow to train models in a distributed manner without sharing any data, but their application to causal discovery is not straightforward. A naive approach is discovering individual causal models for each local dataset, pooling those models and computing the likely global causal model governing the process that generated all local datasets. Sharing models, however, is not guaranteed to be privacy-preserving, since one can make inferences about local datasets from model parameters (Geiping et al., 2020; Lyu and Chen, 2021; Singhal et al., 2021). Another naive approach is to discover *local* causal networks for each dataset and compute their union. This has two major issues: (i) For finite dataset sizes, locally discovered causal models can vary substantially from the true network and may contain spurious edges, leading to a bad performance, and (ii) this still requires us to explicitly communicate the local causal networks for pooling, which may compromise privacy guarantees (Geiping et al., 2020; Wang et al., 2020).

In this chapter, we propose to discover the global causal network without sharing any data, model parameters, or even local causal networks— using regrets. Intuitively, the regret measures how much worse a given causal network is, compared to the best causal network for a given dataset. We first propose a simple algorithm that can be used to find the underlying causal structure for distributed, private datasets by minimizing over worst-case regret. We further show that minimizing the worst-case regret over these distributed datasets allows us to define a scoring criterion that, under mild assumptions, is guaranteed to be consistent. This implies that we can now employ worst-case regret as a score within Greedy Equivalence Search (GES) (Chickering, 2002) to discover the global causal network with correctness and privacy guarantees, by only using regrets obtained from local datasets. We obtain this as follows: we first let each site discover the best network for its dataset using GES with any consistent scoring criterion (e.g., MDL or BIC), and then optimize the worst-case regret, once again using GES, with respect to the locally discovered causal networks. Throughout the entire learning process, the optimizing algorithm neither sees the data, nor knows the local model parameters. To ensure privacy of local data, we show that using the Laplace mechanism on the shared regrets guarantees ϵ-differential privacy.

To perform federated causal discovery, we instantiate our proposed approach, which we call PERI[1], using three well known consistent scoring criteria. Through extensive experiments we show that PERI discovers causal networks of

[1]In astronomy, Peri is the point at which an orbiting object is closest to the center of mass of the body it is orbiting (such as a planet). In our approach, we aim to discover that network which is collectively closest to the local networks of all environments.

higher quality than the state of the art on both synthetic and real-world data, scales upto 100 distributed environments while requiring orders of magnitude less communication.

We organize this chapter as follows. We start with a review of existing literature in Sec. 5.1, and preliminaries in Sec 5.2. Next we explain the concept of regret in Sec. 5.3 and propose a naive algorithm to perform regret-based federated causal discovery using beam-search in Sec. 5.4. We then show in Sec. 5.5 that using regret as a score within well known causal discovery algorithm preserves consistency guarantees which lets us build a theoretically sound, scalable federated causal discovery algorithm, Peri, in Sec. 5.6. For this proposed algorithm Peri, we describe how to additionaly provide privacy guarantees in Sec. 5.7. Finally, we provide experimental evaluation in Sec. 5.8 before providing concluding discussion in Sec. 5.9.

5.1 Related Work

Many methods have been proposed to discover causal networks given a single dataset (Spirtes et al., 2000a; Chickering, 2002; Shimizu et al., 2006; Peters et al., 2014; Blöbaum et al., 2018a; Huang et al., 2018; Zheng et al., 2018a; Mian et al., 2021), much fewer for doing so given data collected from multiple environments (Zhang et al., 2017; Mooij et al., 2016), and only a small handful for doing so when the data cannot be gathered centrally (Ng and Zhang, 2022).

Methods that can consider only a single dataset are not applicable in our setting; even if we ignore all privacy aspects and were to centrally collect and pool all data, it is well known that naively pooling the data can introduce unwanted bias in estimation (Tillman, 2009). Methods that can consider multiple datasets, such as when data has been collected from different environments (Yang et al., 2018; Squires et al., 2020), come one step closer to the scenario we consider in this paper. The most prominent approaches still combine all data, adding one or more context variables to distinguish the rows of the combined datasets, and then perform causal discovery on the augmented data (Zhang et al., 2017; Magliacane et al., 2018). A very general such approach is the Joint Causal Inference (Jci) framework proposed by Mooij et al. (2016), which permits any constraint-based causal discovery algorithm to work with data from multiple environments. Each of these approaches require that all data is available at one site, which is prohibitive in our setting.

Federated learning allows for learning without the need for centralized data. Rather than sharing data with other nodes, the key idea in federated learning is that we share (partial) local results. The topic of federated causal discovery is relatively young. Proposals for federated causal inference (Xiong et al., 2021) and federated causal discovery (Shimizu, 2012) require strong parametric assumptions. Recent approaches avoid these, either by sacrificing convergence

guarantees (Gao et al., 2021) or by sharing additional learning parameters (Ye et al., 2022; Ng and Zhang, 2022). Although these methods do not directly share data, by sharing completely specified local causal models they can provide attackers sufficient information to reconstruct local data (Geiping et al., 2020; Singhal et al., 2021).

In this paper we propose a framework for federated causal discovery that, rather than parameters, only shares regret values. We build upon the idea of regret-based learning to propose a theoretically sound score that comes with strong privacy guarantees and achieves lower communication costs while scaling up to 100 environments.

5.2 Preliminaries

5.2.1 Notation and Assumptions

We consider data consisting of m variables $\boldsymbol{X} = \{X_1, \ldots, X_m\}$ with $X_i \in \mathbb{R}$, split into d different environements $\boldsymbol{D} = \{D^1, \ldots, D^d\}$ of sizes $n^{(1)}, \ldots, n^{(d)}$. We assume that each D^i is drawn i.i.d. from a distribution $P_i(\mathbf{X})$, which are all are entailed by the same true causal network G^* but where the parameters associated with G^* may be different between D^i. Our goal is to solve the following problem.

Problem Statement 5.1 (Informal) *Given data $\boldsymbol{X}$, discover the true causal network G^* in a federated (without pooling data) and privacy-preserving (without sharing any models fit over individual datasets) manner.*

Akin to previous chapters we need to assume 1) the causal Markov condition (Spirtes et al., 2000a), 2) causal faithfulness, and 3) causal sufficiency (Pearl, 2009), which makes it possible to discover causal networks from observational data up to the Markov equivalence class (MEC). When all of the above assumptions hold, algorithms such as Greedy Equivalence Search (Chickering, 2002) can discover causal networks, for a single dataset, up to Markov equivalence (Glymour et al., 2019) i.e. partially oriented causal networks where all collider structures are correctly identified. Unlike Chapters 2 and 3, however, causal sufficiency and faithfulness assumptions may not always be necessary and we provide a discussion in Sec 5.9 on how they can be avoided.

5.2.2 Greedy Equivalence Search

Greedy Equivalence Search (Ges) (Chickering, 2002) is a score-based causal discovery approach that learns a causal network $\hat{G}$ from observational dataset $\boldsymbol{X}$. To do so it uses a scoring criterion L to measure how well a network

G describes $\boldsymbol{X}$. Starting from an empty network, GES iteratively builds a causal network through repeated forward respectively backward-search. In each step of the forward search, GES chooses a single edge addition to the current best network such that the edge improves score the most and uses the new network as the best network for the next step. Similarly, in each step of the backward search, single edge deletions that improve score the most are chosen. Each phase ends when no modifications of the current network improve score anymore. GES is guaranteed to return the correct Markov equivalence class as $n \to \infty$ if the following two conditions hold:

1. L is decomposable.
2. L satisfies the (global) consistency property.

DECOMPOSABLE SCORE A given score L is decomposable means that L can be expressed as

$$L(\boldsymbol{X}; G) = \sum_{j=1}^{m} l_j(X_j; pa_j^G)\,,$$

where pa_j^G are the parents of variable X_j in G and l_j is only a function of X_j and its parents.

SCORE CONSISTENCY Chickering (2002) defines consistency property of a given score L as follows.

Definition 5.1 (Chickering (2002), Consistent Scoring Criterion) *Let G, H be any pair of DAGs, $\boldsymbol{X}$ be a set of data consisting of n records that are i.i.d. samples from some distribution $P(\cdot)$. A (minimizing) scoring criterion L is consistent if in the limit $n \to \infty$, the following two properties hold:*

1. *If H contains P and G does not contain P, then $L(\boldsymbol{X}; H) < L(\boldsymbol{X}; G)$*
2. *If H and G both contain P, and G contains fewer parameters than H, then $L(\boldsymbol{X}; G) < L(\boldsymbol{X}; H)$,*

where *contains* means that G has the exact independence constraints implied by P.

Despite its greedy nature, if L is consistent, GES is guaranteed to find a graph in the MEC of the true G in the large sample limit, although (in the worst-case) this discovery could require runtime super-exponential in the number of variables. Examples of decomposable consistent scores include the Akaike's Information Criterion (AIC) (Akaike, 1974), Bayesian Information Criterion (BIC) (Schwarz, 1978) and scores defined using Minimum Description Length (MDL) (Grünwald, 2007; Mian et al., 2021).

GES, however, is limited to finding causal networks over a single dataset and can therefore only be used to learn individual networks G_i for each dataset

D^i. To extend GES to a federated setting, we require that we can measure the score of a global network G relative to a locally learned G_i *without* knowing what the local networks are. To do so, we introduce the concept of regret.

5.2.3 Regret

Given data D and some model M, from a model class $\mathcal{M}$ that explains the data, let $L(D; M)$ be a score function that is minimized when M is the true model for X. Regret $R(M)$ for a given model M with respect to data D is defined as the difference in scores when evaluating D using M instead of the best model M^* for D. Formally stated

$$R(M) \coloneqq L(D; M) - \min_{M^* \in \mathcal{M}} L(D; M^*)\,, \tag{5.1}$$

where we drop the dependence on the data D and write $R(M)$ instead of $R_{_D}(M)$ to simplify notation. Simply put, regret measures how much worse the proposed model M is compared to the *best* model for the data. If both M and M^* are present in $\mathcal{M}$, $R(M)$ is lower bounded by 0, which is achieved when $M \equiv M^*$.

5.3 Learning from Regrets

In this section we show that we can use regret defined in Eq. (5.1) to build a score for federated causal structure discovery. Using such a regret-based score, we can propose a straightforward algorithm to search for global causal network over multiple private datasets *without* ever looking at the data or any of the locally learned information. For our model class $\mathcal{M}$ defined in Eq (5.1), we consider the space of all Directed Acyclic Graphs (DAGs), $\mathcal{G}$. Hence for our proposed setup we can write Eq. (5.1) as

$$R_i(G) \coloneqq L(D^i; G) - \min_{G_i \in \mathcal{G}} L(D^i; G_i)\,,$$

where $R_i(G)$ is the regret associated with dataset D^i when using network G.

Now it becomes easy to see the merit of using regret from a federated learning perspective: Given a server S that aims to learn a global causal network using d different sites, each with their own private datasets $D^1, \ldots, D^d$, S can send a network G and a scoring criterion L to each site and optimize over regrets that it receives back. To do so, S needs to consolidate these regret values received back from each site into a meaningful score. We propose this to be

the worst-case regret calculated over *all* environments,

$$\begin{aligned} L_F(G) &\coloneqq \max_i R_i(G) \\ &= \max_i \big(L(D^i; G) - L(D^i; G_i)\big) \end{aligned} \tag{5.2}$$

where G_i is the minimizer for $L(D^i; \cdot)$.

Using the aforementioned formulation, the goal of the server is to find that network G that minimizes the worst-case regret among all the networks. Formally stated

Problem Statement 5.2 *Given samples $\mathcal{X} = \{\boldsymbol{X}^1, \dots, \boldsymbol{X}^d\}$ corresponding to d environments $\boldsymbol{D} = \{D^1, \dots, D^d\}$ that share a common underlying causal DAG, find $\widehat{G}$ such that*

$$\widehat{G} = \operatorname*{argmin}_{G \in \mathcal{G}} \max_i R_i(G) \; . \tag{5.3}$$

This obtained network $\widehat{G}$ is the one which trades off errors relative to one local network G_i to another local network G_j and tries to jointly minimize them. Such a $\widehat{G}$ is the least bad network relative to any of the local networks. To simplify notation, we use $\boldsymbol{X}^i$ and D^i interchangeably from next section onward.

5.4 Optimizing over Worst-case regret

Using the concept of regrets in Eq (5.1) and a consistent scoring criteria, we can already propose a straightforward algorithm for federated causal structure learning. We refer to this basic ideas as Regret-based Federated Causal Discovery (RFCD). To this end, let $\mathcal{A}$ be any score-based structure discovery algorithm, e.g. GES (Chickering, 2002) or GSP (Solus et al., 2017) and let L be any consistent score used within $\mathcal{A}$ such as the BIC-score (Schwarz, 1978) or MDL-based score (Mian et al., 2021). Then we can replace the terms $\min_{G_i} L(X_i; G_i)$ in the regret term as follows

$$\begin{aligned} \widehat{G} &= \operatorname*{argmin}_{G} \max_i \left(\widehat{R_i}(G)\right), \text{ where} \\ \widehat{R_i}(G) &\coloneqq L(D^i; G) - L(D^i; \widehat{G_i}) \, , \end{aligned}$$

where $\widehat{G_i}$ is the graph learned by $\mathcal{A}$ on D^i, and $L(D^i; G)$ is the score that evaluates how well does G fit the data D^i, The idea is that when $\mathcal{A}$ is a consistent algorithm with respect to L then as $n^{(i)} \to \infty$ we find that $\widehat{G_i} \to G^*$,

and subsequently $\text{argmin}_{G_i} L(D^i; G_i) \rightarrow G^*$. This means that for sufficiently large datasets the above replacement is harmless.

Next, given the goal in Eq. (5.3), how do we estimate $\widehat{G}$? We can, in theory, find the true causal network by exhaustively searching over the space of DAGs and taking the one minimizes Eq. 5.4. Such an approach, however quickly becomes infeasible as the space of DAGs grows super-exponentially in the number of nodes and the loss landscape associated with L does not exhibit any structural regularities, which makes optimal Bayesian structure discovery NP-Hard (Chickering et al., 2004).

Nevertheless, in practice, we can implement the above approach as a beam search which is guaranteed to find the optimum, given large enough beam size, and also allows for reducing beam size to trade optimality for runtime. That is, starting from the empty network G_0 we evaluate at every step every one-edge extension G of the b best networks $G_{t,1}, \ldots, G_{t,b}$ from the previous step and keep the b best networks from the current step $G_{t+1,1}, \ldots, G_{t+1,b}$. We repeat this until no further extensions of any of the networks $G_{t,j}$ improve upon the best network already found. Then we set

$$\widehat{G}_{\mathcal{B}} = \underset{G_{t,j}}{\text{argmin}} \max_i \left(L(D^i; G_{t,j}) - L(D^i; \widehat{G_i}) \right) ,$$

to be the best performing network discovered so far.

We can use the above formulation to perform federated causal discovery as shown in Algorithm 5.1. Given a server S and d different sites $D^1, \ldots, D^d$, the server communicates the algorithm $\mathcal{A}$ and the scoring metric L to each of the clients. Each client then runs $\mathcal{A}$ on its own data to learn the local $\widehat{G_i}$ (line 4). The server then sends an empty network G_0 to each client and receives the regrets $r_0^{(i)}$ w.r.t locally learned networks back from each $C^{(i)}$. Next, the server calculates the worst-case regret r_0 (lines 5-7) and initializes a beam B of size b with the state (G_0, r_0) (line 8). Then the search process begins. At each search-step, all possible single edge extensions of each DAG in the beam are enumerated and their worst-case regret calculated via communication between the server and clients (lines 11-14). The top b extensions with lowest worst-case regret are then retained as the new beam (line 15). An immediate advantage of our search procedure is that it is guaranteed to converge. This is because regret defined using a consistent score L used within $\mathcal{A}$ can never go below 0, and we only take steps that reduce regret. Hence we continue the search until convergence. During the entire learning process, the server neither sees the data, nor the locally learned causal networks for each dataset. The only communication that takes place between server and client is the regret value r_G for a query DAG G.

Setting beam size to $\binom{m^2}{(m^2-m)/2}$ is equivalent to an exhaustive search where

Algorithm 5.1: Causal Discovery using RFCD

Input: Algorithm $\mathcal{A}$, Consistent scoring criteria L, beam size b
Output: Causal network $\mathbf{G}$

$B \leftarrow \emptyset$
$r_0 = 0$, $G_0 \leftarrow \emptyset$
for $i = 1 \dots l$ **do**
  $c[i]$.LEARN$(\mathcal{A}, L)$
  $r_0^i \leftarrow c[i]$.REGRET(G_0, L)
  if $r_0^i > r_0$ **then**
    $r_0 \leftarrow r_0^i$
$B \leftarrow B \oplus (G_0, r_0)$
repeat
  $Q \leftarrow B$.COPY()
  $\mathbb{G} \leftarrow$ all admissible single edge extensions of DAGs in B
  foreach $G \in \mathbb{G}$ **do**
    $r_G \leftarrow$ MAX$(c[i]$.REGRET$(G, L))$ *for* $i = 1 \dots l$
    $Q \leftarrow Q \oplus (G, r_G)$
  $B \leftarrow$ first b entries in Q
until convergence;
$\mathbf{G}^* \leftarrow$ first entry in B
return $\mathbf{G}^*$

we are guaranteed to find the global optimum. This, however, is only suitable for networks with small number of variables. Alternatively, setting $b = 1$ results in a greedy DAG search algorithm which is only guaranteed to discover correct causal network if the underlying structure is a tree. In practice, we find that setting beam-sizes as small as 10 already performs well even though our search is only guaranteed to find local optima in those cases.

Minimizing worst-case regret using beam search using RFCD as shown in Alg. 5.1 is a simple and straightforward idea. This, however, neither entails theoretical guarantees for fixed beam size nor does it scale. To build an algorithm that is scalable and entails correctness guarantees require us to prove consitency property for our proposed score. This we do next.

5.5 Consistency Guarantees for Worst-case Regret

An obvious consideration to solve both above mentioned limitations is to use worst-case regret as a score *within* well established causal structural learning algorithms such as Ges (Chickering, 2002; Ramsey et al., 2017). We can not, however, directly plug in our proposed score into Ges as the latter requires the score to fulfill certain properties. Therefore, we must first prove that worst-case regret is both a decomposable, as well as consistent scoring criterion thereby implying that the minimizer for Eq. (5.3) is the true causal network. This we can prove for a class of regularization-based scores that we describe next.

To prove that the minimizer for Eq. (5.3) is the true causal network, we consider scores, $L(D^i; G)$, of the form

$$L(D^i; G) = L(G) + L(D^i|G) \ ,$$

where $L(G)$ is a function penalizing the complexity of the network G and the parameters associated with the class of generating functions e.g. linear or spline relationships between each variable and its parents, and $L(D^i|G)$ is the log-likelihood of the data given the G.

We can now show that in the limit, when every site uses the same consistent score L and obtains arbitrarily much data then our method is guaranteed to find the correct MEC.

Theorem 5.2 *Let G^* be the true underlying causal network for all $P(D^i)$ and let $n^{(1)} \ldots, n^{(d)} \to \infty$. Further let L be a consistent and decomposable score. Then*

$$\lim_{n^{(1)},\ldots,n^{(d)}\to\infty} P\left(\widehat{G} \sim G^*\right) = 1 \, .$$

That is, $\max_i R_i(G)$ is consistent when all $n^{(i)} \to \infty$.

We can further relax Thm. 5.2 to not require that every site's amount of data grows over time. In fact, as long as even one of the datasets grows, we nevertheless find all edges.

Theorem 5.3 *Let G^* be the true causal network for all $P(D^i)$ and let $N \coloneqq \max_i n^{(i)} \to \infty$. Further let L be a consistent and decomposable score. Then*

$$\lim_{N\to\infty} P\left(\widehat{G} \sqsupseteq G^*\right) = 1 \, .$$

For scores L, like AIC, the correct MEC is generally impossible to recover precisely because the penalty for additional edges does not scale with the number of data points. In contrast, for the BIC score this is not an issue.

Algorithm 5.2: PERI for federated causal discovery

Input: Scoring criterion L
Output: Causal network G

```
for i = 1 ... d do
    site[i].GES(L)
G* ← ∅
Define L_F(G) := max_i [L(X_i, G) − L(X_i, G_i)]
repeat
    G* ← server.FORWARDEQVSEARCH(G*, L_F)
    G* ← server.BACKWARDEQVSEARCH(G*, L_F)
until convergence;
return G*
```

Corollary 5.4 *Let the assumptions of Thm. 5.3 hold and let L be the BIC score. Then*

$$\lim_{N\to\infty} P\left(\widehat{G} \sim G^*\right) = 1\,.$$

That is, the score $\max_i R_i(G)$ *is consistent when L incorporates a BIC-penalty for parameters and $N \to \infty$.*

The proof of Cor. 5.4 applies equally to any other consistent criterion where the parameter-penalty grows strictly with sample size, e.g., MDL-based scores such as those used in GLOBE or ORION. In Sec. 5.9 we discuss how to extend our work to other types of scores. These results imply that $R_i(G)$ remains both a decomposable *and* a consistent scoring criterion as long as L used within $R_i(G)$ is consistent. We can hence, as explained in Sec. 5.4, define $R_i(G)$ using any consistent L and perform a search for the underlying causal network G by exhaustively evaluating all possible causal networks and choosing one that minimizes Eq (5.3). Moreover using our derived consistency guarantees, we can now instantiate our search more efficiently making it scale with increasing number of variables, using any of the well known causal discovery algorithms.

Using these results we show in the next section how we can instantiate an efficient regret-based causal learning framework, while maintaining correctness guarantees.

5.6 The Peri Framework

Using consistency results derived in 5.5 we now describe PERI, a score-based federated causal discovery approach for distributed environments. For this

explanation we consider the well-known GES algorithm. PERI, however, can be applied to any score-based algorithm with likelihood-based scoring criteria.

Let L be any consistent score used within GES, such as BIC, and let L_F be the composition that calculates the worst-case regret using L as defined in Eq.(5.2). Then L_F can be used as a consistent score within GES (Thm. 5.3, Cor. 5.4) to discover causal networks in a federated fashion. As a result, we can perform federated causal discovery as shown in Algorithm 5.2. Given a server S and d different sites, each with their own private datasets $D^1, \ldots, D^d$, the server communicates L to each of the sites. Each site then learns a local network G_i using GES (lines 1-2). The server then instantiates L_F as defined in Eq. (5.2) (line. 4) and runs its own GES using L_F. In the forward pass (line. 6), the server communicates the best discovered network G_t, at iteration t, to all sites. Each site converts G_t to the MEC $\mathcal{E}_t$ and calculates regret over all possible single edge extensions of $\mathcal{E}_t$. The list of these scores is communicated back to the server. Next, the server chooses the network G_{t+1} with the lowest worst-case regret among all these extensions and sets this network as the best network for the next iteration. The forward search ends when no extensions of G_t improve the score anymore. The backward search (line. 7) is analogous to the forward search except that the regret scores are calculated over single edge deletions of $\mathcal{E}_t$ at each iteration. We repeat the search process until convergence (line. 8). During the learning process, the server neither sees the data, nor knows the local models for any site. The only communication that takes place is the list of regret values for networks in the MEC for the query DAG G_t.

This proposed approach has several advantages: First, the regret for a query network can be calculated locally at each site and returned back to the server, requiring no communication of model parameters — the job of the server is to choose the worst-case regret for a given network G. Second, PERI is *guaranteed* to converge. This is because regret is lower-bounded by 0, and we only take steps that reduce regret. Third, we do not need any additional assumptions except the ones required for L — to be used within GES we require L to be decomposable and consistent. PERI, in fact, can be viewed as a generalization of GES to multiple datasets as PERI for a single site simplifies to GES.

5.7 Privacy Guarantees

With the framework explained, we now describe how we can go an extra step and guarantee differential privacy using PERI. Intuitively, sharing only regrets reveals less about local data than sharing model parameters and causal networks: Attackers can infer membership in local datasets from model parameters (Shokri et al., 2017; Ma et al., 2020) and even reconstruct local datasets from model updates (Zhu and Han, 2020). Moreover, model parameters allow an attacker to craft poisoning and backdoor attacks (Sun et al., 2019). Shar-

ing only causal graphs still does not fully protect local data, since "a causal graph can leak information about participants in the dataset" (Wang et al., 2020). PERI shares only regret values, but local causal networks can be reconstructed by optimizing Eq. (5.3) with respect to the target site, which in principle remains NP-hard (Chickering et al., 2004).

By applying the Laplace mechanism (Dwork et al., 2006), i.e., adding appropriate noise to the regret values, we can guarantee that sensitive local data is protected in terms of ϵ-differential privacy. To prove this guarantee holds, it suffices to show that all regrets R_i have bounded sensitivity. For that, we assume that G corresponds at each site i to a parameter vector $\theta^{(i)}$ such that X_j is modeled via $X_j^{(i)} = f(\mathrm{pa}_j, \epsilon_j; \theta^{(i)})$ with independent noise ϵ_j. We assume that our score L is well-behaved in the following sense: when X_i is of size n and $X'^{(i)}$ differs in one element from X_j then the corresponding optimizers for L differ by $\left\|\theta^{(i)} - \theta'^{(i)}\right\|_1 \propto 1/n$. This assumption holds for many learning algorithms, e.g. convex empirical risk minimization with finite VC-dimension or Rademacher complexity (Von Luxburg and Schölkopf, 2011).

Lemma 5.5 *Assume that $P_i(x;\theta)$ is uniformly lower-bounded bounded by r, i.e., $\forall x \in \mathcal{X}\, \forall \theta \in \Theta$: $P_i(x;\theta) \geq r$, that $\|\theta\| \leq M$ for all local model parameters $\theta \in \Theta$, and that the score L is partially differentiable with respect to θ. Let $X^{(i)}$ and $X'^{(i)}$ be datasets that differ in a single element, i.e. $X^{(i)} \setminus X'^{(i)} = x_k$, θ and θ' the respective local parameters, and $\widehat{R_i}(G)$ and $\widehat{R_i'}(G)$ the respective regrets. Assume that $\|\theta - \theta'\|_1 \leq 2M/n$. Then the sensitivity $\Delta\widehat{R_i}$ of the regret is bounded by*

$$\max \left| \widehat{R_i}(G) - \widehat{R_i'}(G) \right| \leq (4M+1) \log r + \mathcal{O}\left(\frac{\log n}{n}\right) \quad .$$

With this, it follows from the Laplace mechanism (Dwork et al., 2006) that adding Laplacian noise to regrets before sending them to the server guarantees ϵ-differential privacy.

Proposition 5.6 *Assume that each local regret $\widehat{R}_i$ has sensitivity $\leq Q$. Then PERI with i.i.d. Laplace noise with scale $\lambda = Q/\epsilon$ added to each $\widehat{R}_i$ is ϵ-differentially private.*

In practice, adding noise can deteriorate the training process, but we show in Sec. 5.8 that the practical performance of PERI is robust against noise added to local regret values and that it performs well under privacy requirements.

5.8 Evaluation

5.8.1 Setup

We instantiate Peri using three consistent scoring criteria, which are: the AIC (Sakamoto et al., 1986), BIC (Schwarz, 1978) and spline-based MDL score (Mian et al., 2021). We refer to these instantiations as Peri-Aic, Peri-Bic and Peri-Mdl respectively. Since Ges could get stuck in local-optima when discovering local causal networks with limited sample sizes (Lu et al., 2021), for practical reasons we run Peri in two rounds to prevent it from being misled due to incorrectly discovered local networks: first we use Peri to learn $\widetilde{G}$ using the local G_i for each environment. Next, we learn the actual G^* using Peri by enforcing $\widetilde{G}$ as the local model for all environments.

We compare to RFcd (Mian et al., 2022) as representative score-based approach. As representative ANM based method we compare to Direct-Lingam (Shimizu, 2012), which is a modified version of the original Lingam (Shimizu et al., 2006) for causal discovery over multiple groups. We compare to the nonlinear version of NOTEARS-ADMM (NT-Admm) (Ng and Zhang, 2022) as continuous optimization based federated causal discovery approach. Both of the above approaches require that the model parameters be communicated between server and sites. As baseline, we use Ges (Chickering, 2002) to locally discover causal networks within each environment and take a union over the discovered networks to predict the global causal network. While no parameter exchange takes place, the local causal networks are still shared with the server. For this particular work we cannot compare to approaches like Orion (Mian et al., 2023b), Continent (Mian et al., 2024), CdNod (Zhang et al., 2017) or Jci (Mooij et al., 2016) as these methods require that we first pool all data and are therefore not applicable to our setting.

We evaluate the predicted networks in terms of structural similarity using the Structural Hamming Distance (*SHD*) (Tsamardinos et al., 2006) — which counts the number of edges where two networks differ. For comparability across multiple experiments, we normalize *SHD* to be in the range $[0, 1]$. To measure correctness of edge orientations in the predicted networks, we use the $F1$ score. For synthetic data, we terminate all experiments that do not finish within 24 hours. We standardize all data to have zero mean and unit variance to avoid practical issues like var-sortability (Reisach et al., 2021) and make all code and data available for research purposes.[2]

[2] https://eda.rg.cispa.io/prj/peri/

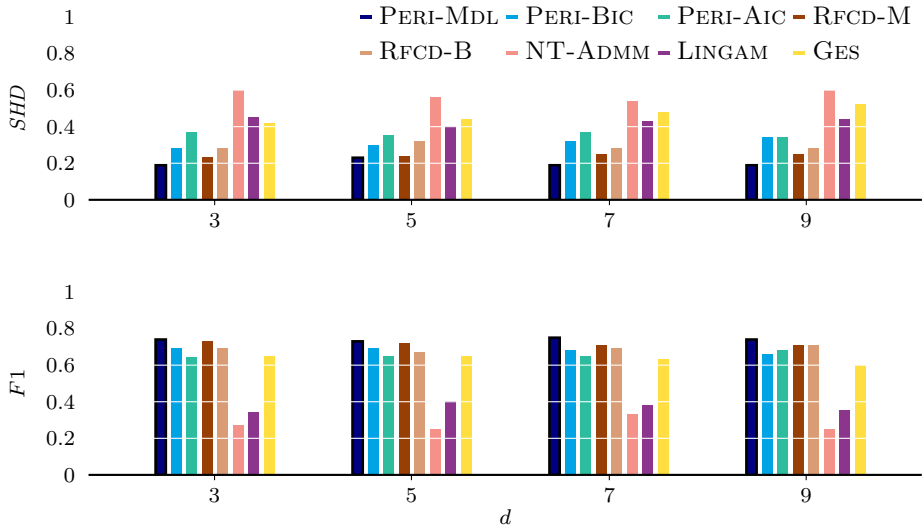


Figure 5.1: [Top, Lower is better] *SHD* and [Bottom, Higher is better] $F1$ over environment size $d = \{3, 5, 7, 9\}$. Peri-Mdl performs the best overall.

5.8.2 Results

Next, we provide empirical results of our work on Peri. We extensively test Peri using both synthetic and real-world data and evaluate Peri's performance on five distinct aspects: 1) causal discovery in our intended setting 2) causal discovery when only a subset of environments are available at each learning iteration, 3) performance under privacy considerations, 4) communication efficiency, and 5) causal discovery on real-world data.

Causal discovery in our intended setting We start with the simplest setting where we generate multiple datasets using the same underlying distribution. We have number of environments $d \in \{3, 5, 7, 9\}$, number of variables $m \in \{5, 10, 15\}$, and samples per environment $n = 5000$ as our experimental setting. We perform a total of 52 experiments for each m. We simulate DAGs using the Erdős-Rényi model and generate each effect, X_i from its parents pa_i using functions of the form $X_i = f(\text{pa}_i) + \epsilon_i$, where f is a non-linear function defined over pa_i, and ϵ_i is independent additive noise Gaussian noise with zero mean. We generate complex causal relationships by defining f to be a randomly initialized 2-layer neural network, using the causal discovery toolbox (Kalainathan and Goudet, 2019).

We report the results across varying number of environments in Fig. 5.1

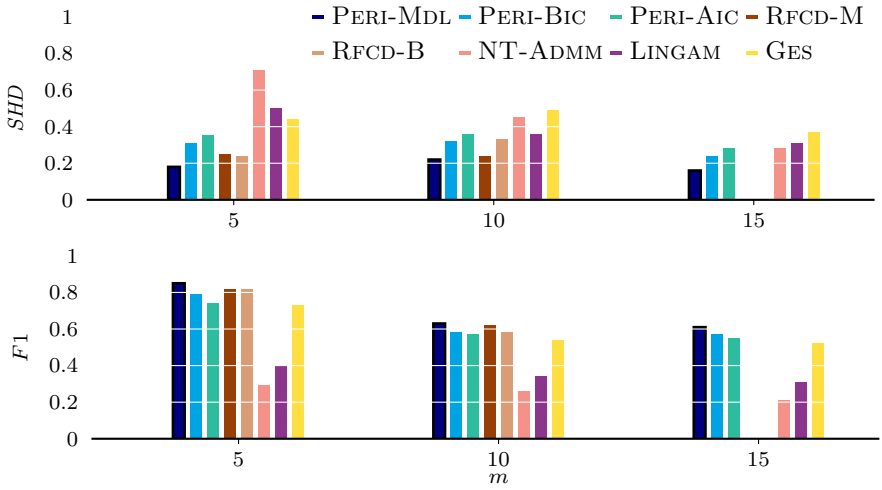


Figure 5.2: [Top, Lower is better] *SHD* and [Bottom, Higher is better] $F1$ over networks with variable count $m = \{5, 10, 15\}$. PERI-MDL consistently performs the best overall. RFCD does not terminate within 24 hours for any 15 variable networks.

and for different sized networks in Fig. 5.2. We see that overall PERI-MDL outperforms all other approaches in terms of both *SHD* as well as orientation-$F1$. One reason for this is that spline-based MDL score uses non-parmetric regression to model causal relationships and is therefore able to identify the causal parents with higher accuracy. This is in contrast to PERI-BIC and RFCD-B, both of which use the BIC score with a lenient parameter penalty which could support inclusion of spurious edges. We see in Fig. 5.2 that both RFCD variants, despite their competitive performance, fail to scale to networks with $m = 15$. Moreover we find that baseline GES has better $F1$-scores than LINGAM and NT-ADMM..

DISCOVERING NETWORKS WHEN ONLY A SUBSET OF ENVIRONMENTS ARE AVAILABLE As our next experiment, we generate data using two well known causal structures, namely the ASIA (Lauritzen and Spiegelhalter, 1988) and WASTE (Lauritzen, 1992) networks. We generate a total of 10 experiments, each containing 100 unique environments. At each round of update, we allow the methods to only query a fraction $s \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$ of randomly chosen environments. We average the results over 30 iterations of each experiment for PERI-MDL, PERI-BIC and PERI-AIC whereas for NT-ADMM,RFCD-M and RFCD-B we average over 10 iterations due to longer run times. We omit LINGAM as it does not contain a mechanism to query a subset of environments.

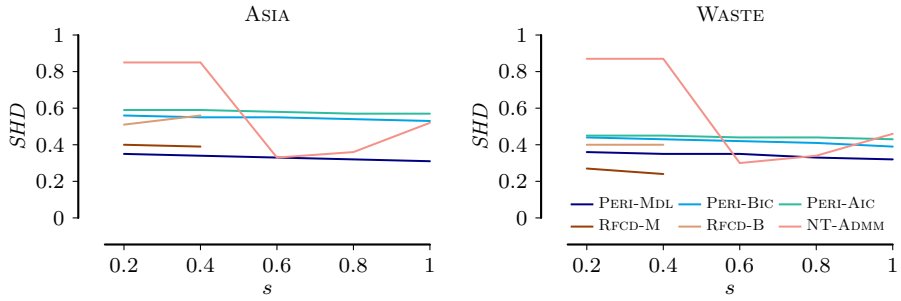


Figure 5.3: [Lower is better] Averaged *SHD* over Asia (Left) and Waste (Right) networks when querying only a subset s of environments. Peri-Mdl performs best. Results for Peri progressively improve as more environments are allowed to be queried. Rfcd-B and Rfcd-M do not finish within 24 hours for any experiments with $s > 0.4$.

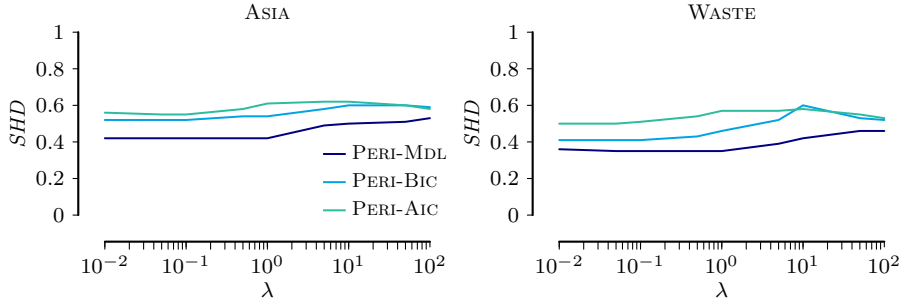


Figure 5.4: [Lower is better] Averaged *SHD* over Asia (Left) and Waste (Right) networks with $d = 100$ and Laplace noise on regret values with scale parameter $\lambda \in [0.0, 0.01, 0.05, 0.1, 0.5, 1, 5, 10, 100]$. Peri-Mdl deteriorates the slowest. Rfcd is omitted as it does not produce any output after 24 hours for this experiment.

We show the results in Fig 5.3 where we see that Peri-Mdl performs the best overall. All of the Peri approaches show improvement in results as the available number of environments increase. Surprisingly, NT-Admm shows inconsistent performance which initially improves with increasing environment, but subsequently worsens even when all of the sites are available.

Performance under privacy considerations We test the effect of adding Laplacian noise with 0 mean and increasing scale λ over the range $[0.01 - 100]$ to the values of regret before communicating the regret values to the server.

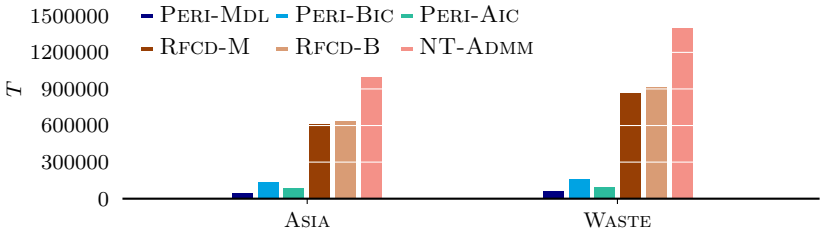


Figure 5.5: [Lower is better] Average number of parameter values, T, communicated to infer causal structures over Asia and Waste networks at $d = 100$. For baseline Ges, the number of parameters is always 6400 and 8100 for Asia resp. Waste networks. Neither Rfcd-M nor Rfcd-B produce any results within 24 hours for $d = 100$. We therefore report their results for $d = 40$.

The results in Fig. 5.4 indicate that Peri is robust to Laplacian noise. Indeed, the performance of Peri does not change significantly with λ up to 1; and not even with $\lambda = 10$ when we use MDL. Since the larger noise corresponds to stronger privacy guarantees, this implies that Peri performs well under privacy requirements. We find that neither Rfcd-M nor Rfcd-B produced any output after 24 hours for any of the settings in this experiment.

Communication efficiency To measure communication efficiency between server and sites, we investigate the total rounds of communications required by each approach to infer a causal network. Overall Peri-Mdl is able to discover the causal network on average 15 rounds of communications for the Asia network, with Peri-Bic and Peri-Aic following closely with 21 resp. 23 rounds. This is much less than NT-Admm which always terminates after the max iteration cap of 176 rounds set by Ng and Zhang (2022). This means that the number of parameters that Peri exchanges during the course of learning for both Asia and Waste networks are significantly less than NT-Admm and Rfcd as we show in Fig. 5.5.

Real world Data To see how well Peri performs on real-world data, we consider three distinct real-world networks. We consider two non-overlapping networks of sizes $\{5, 15\}$ from the Lung cancer gene-expression dataset (Reged) (Statnikov et al., 2015). For each of the Reged networks we generate 9 distinct environments without any sample overlap, each with 2000 samples per environment. Third, we consider the Sachs protein signaling network (Sachs et al., 2005) consisting of 11 variables, already measured over 9 distinct environments. The Sachs dataset provides a challenging setting since each environment has

Table 5.1: [Lower is better] *SHD* for multiple real-world networks. PERI-MDL discovers the exact ground truth for both REGED 5 and REGED 15.

	REGED 5	REGED 15	SACHS
PERI-MDL	**0**	**0**	18
PERI-BIC	1	25	18
RFCD-M	**0**	5	**17**
RFCD-B	1	37	18
LINGAM	4	26	**17**
NT-ADMM	6	23	23
GES	2	55	25

its data generated from a *different* intervened-upon causal network. This violates our assumption of a common, shared ground truth network.

We see from the results in Table 5.1 that PERI-MDL discovers the exact ground truth for both REGED 5 and REGED 15 networks and is marginally outperformed by LINGAM on the assumption-breaking case of SACHS dataset. We find that RFCD-M, which also uses a spline-based MDL score, recovers the correct causal network for REGED 5 but fails to do the same for REGED 15.

5.9 DISCUSSION

We considered the problem of discovering causal networks in a federated setup. We have proposed a new method PERI that allows us to discover causal networks in a privacy-preserving manner. Extensive experiments on diverse settings show that PERI outperforms the state of the art in federated causal discovery, both in quality of the discovered causal networks and communication efficiency while providing privacy guarantees on top of it.

We considered three different scores to instantiate PERI: AIC (Akaike, 1974), BIC (Schwarz, 1978), and spline-based MDL score (Mian et al., 2021). We found that while all three work well, the MDL score overall works best in practice. One of the reasons for the superiority is the ability of the proposed MDL score to model causal relationships non-parametrically in combination with an adaptive penalty for the parameters, rendering the method robust even when large noise values are added to regret.

We discover the global DAG by sharing only regrets, but we do not obtain the global models for each causal relationship; methods that share local model parameters do obtain them, at the cost of privacy and communication. We could additionally measure the regret with respect to global parameters θ. In such a scenario, the server proposes both G and θ to each site, instead of

sending G alone. The conditions under which the parameter space Θ can be efficiently searched remains an open question.

We have used orientation-$F1$ to measure the correctness of edge orientation. Alternatively, one could consider the use of Structural Intervention Distance *SID* (Peters and Bühlmann, 2015), which measures the number of intervention distributions where two networks differ. It is, however, not straightforward to interpret *SID* between two Markov equivalence classes. This is because, as opposed to *SHD*, the *SID* of the ground-truth Markov equivalence class with itself is almost always non-zero. This makes *SID* dependent on the underlying Markov equivalence class and incomparable across experiments.

While in this work we instantiate Peri using Ges, regret-based federated causal discovery framework is agnostic of the underlying causal discovery algorithm: For any score-based causal discovery algorithm $\mathcal{A}$ and a consistent score $\widetilde{L}$ with respect to $\mathcal{A}$, if L_F defined in Eq. (5.3) can be proven to be consistent for $\widetilde{L}$, we can simply replace Ges in Algorithm. 5.2 with $\mathcal{A}$ and perform federated causal learning using $\widetilde{L}$ as the score. This implies that, unlike Ges, if $\mathcal{A}$ does not require the faithfulness assumption as in the case of Gsp (Solus et al., 2017), we can perform causal discovery without the latter. How to preserve guarantees for such score-based approaches, as well as for the ones that consider a mixture of observational and interventional data (Yang et al., 2018; Squires et al., 2020; Brouillard et al., 2020) is an engaging line of future work.

5.10 Conclusion

We proposed to perform privacy-preserving federated causal discovery by distributed min-max regret optimization. To do so we proposed, Peri, an approach that can be instantiated using Ges in a score-agnostic fashion. We have designed Peri such that clients only need to communicate local regret values, instead of model parameters, to the server, through which we ensured the privacy of sensitive local data. We instantiated our proposed framework using AIC, BIC and MDL scores. Through extensive experiments, we showed that Peri beats the state of the art by a clear margin and reliably discovers causal networks without ever looking at local data or local causal structures and requires orders of magnitude lower communication cost.

Chapter 6

Conclusion

In this thesis, we focused on developing sound causal discovery algorithms that can be applied to practical scenarios because of their mild assumptions. We specified common scenarios that may occur in practice, and subsequently proposed causal discovery methods applicable to them. Specifically, these included discovering fully oriented causal networks, discovering causal networks under unknown interventions, online causal discovery, as well as privacy preserving causal discovery. Each of these practical aspects constituted a sub-quest of our overarching goal. In the following sections, we summarize our contributions with respect to each of the above-mentioned practical scenarios as well as provide an outlook of limitations that still exist and where the road leads from here.

6.1 Summary of Contributions

The first practical aspect that we addressed was discovering fully oriented causal networks i.e. going beyond Markov equivalence classes. Existing methods, to this end, needed limiting assumptions that hampered their performance in practice. We defined our causal model to consist of non-linear functions with additive Gaussian noise, defined a lossless MDL encoding to compress data under this model, and proved that this score identifies the correct fully oriented causal network in the limit. We then tackled this problem using the algorithmic model of causality using the algorithmic Markov condition (AMC) as introduced by Janzing and Schölkopf (2010a). As AMC depends on measuring Kolmogorov complexity, it is not directly computable. We can, however, approximate it from above using the Minimum Description Length (MDL)

principle (Grünwald, 2007). Keeping true to our goal of practically applicable causal discovery, we proposed a practical GLOBE algorithm that greedily searches for fully oriented causal networks and runs in time polynomial to the number of nodes. To avoid assumptions on the functional form, we modeled the causal relationships using non-parametric regression splines. We showed through extensive experiments that our proposed method works well in practice and beats the state-of-the-art in discovering fully oriented networks, even though the theoretical guarantees entailed by our proposed score were limited to causal trees in case of greedy search.

While it worked well in practice, it is not a silver bullet and we identify room for further advancements from here on. One such avenue is to improve the search strategy: our experiments indicated that the nature of GLOBE can cause it to get stuck in local optima. While exhaustive search is infeasible due to NP-hardness of the problem, continuous-optimization-based approaches like NOTEARS (Zheng et al., 2018b), or iterative sink/source selection approaches (Peters et al., 2014) offer promises as alternate search approaches. GLOBE assumes causal sufficiency, i.e., there are no unobserved confounders and, whenever this assumption is broken, could lead to incorrect results. Recent work from Kaltenpoth and Vreeken (2019, 2023a,b) establishes important identifiability results towards causal discovery with hidden confounding. This opens up interesting avenues of research for practical causal discovery where we can relax the sufficiency assumption for certain scenarios. We could for example do this by using a different score, whose correctness does not depend on assuming causal sufficiency, within GLOBE.

Next, we considered the case where potentially non i.i.d. data over the same set ofvariables may come from multiple, different sources. As each of these datasets may have non-identical distributions and could contain (unknown) interventions, it is not possible to stack such data together without violating the i.i.d. assumption required by existing approaches. To this end, we considered a setting where we learn both the underlying causal network resp. interventions from such data. We turned to the algorithmic model of causation for the second time, and built on GLOBE to develop a theoretically sound MDL score for jointly discovering the causal model and local interventions. Moreover, we provided a practical, highly parallelizable algorithm, ORION, to optimize this score. Unlike existing work, we explicitly avoided assuming prior knowledge of which datasets were observational or interventional and made no assumptions about the functional form of causal relationships. Through extensive evaluation we showed that ORION predicts both structurally and causally better networks than the state-of-the-art in multi environment setting.

Again, there is still room for further development. In addition to exploring different search strategies and handling sufficiency violations like with GLOBE, we can work on enhancing our proposed score to incorporate other types of in-

terventions such as stochastic or edge-introducing interventions. There exists work that establishes theoretical results on when such interventions are identifiable (Correa and Bareinboim, 2020a,b) and a few recent methods that can identify mechanism change interventions (Jaber et al., 2020; Mameche et al., 2022, 2023). Causal discovery for edge introducing interventions, however, is still an open line of work. As a starting point, we could consider defining an encoding scheme that assumes that interventions occur with low probability and thus treat only the most "frequent" edges as part of the true causal graph. Verifying whether such a score retains soundness guarantees is an interesting line of future work. On one hand, developing approaches for edge introducing interventions while maintaining identifiability guarantees is challenging in that it necessitates redefining our definition of a "true" underlying causal network. On the other hand, it can move us one step forward in modeling real world causal relationships better.

The above problem setups required that we had static, fully-specified data sets, all generated by the same underlying causal network. In practice data often arrives in batches over time. Not only does this mean that we need to learn and update our causal hypothesis over time, but each episode likely contains samples from a specific time period, or worse, from a different causal network. This became our next quest, and to achieve this, we proposed an approach that could avoid learning the causal model from scratch upon the arrival of each episode, and could instead learn it in an online fashion. We proposed a consistent strategy to continually update the causal hypothesis, using distribution matching and an information-theoretic perspective of causality. To the best of our knowledge, our method CONTINENT is the first causal discovery approach that can learn causal networks in an online fashion. Using CONTINENT, we could address a novel experimental setting where different causal networks underlie episodic data and we predicted, for a new incoming episode, which causal network it is generated from without explicitly having to learn a network over the incoming episode.

While online learning of causal networks is a step forward, there remain some practical limitations: while in theory we would never "merge" two datasets with different underlying causal networks, in practice this happens quite often. These wrong merges, consequently, mislead the search process. While not trivial, including steps to identify and rectify incorrect episode merges is an interesting line of future work. Moreover, we saw that we can recast our search problem as learning from (pre-specified) missing data, where our guarantees only hold for missing completely at random (MCAR) or missing at random (MAR) types of missingness in data. Going forward we could investigate a more general setup where data is missing not at random (MNAR) and therefore, episodic selection bias may exist even in the limit. Causal inference (Mohan et al., 2013) and causal discovery (Tu et al., 2019; Ma and Zhang, 2021;

Gao et al., 2022; Kitson et al., 2023) from missing data has recently been an active field of research and these results point us in an interesting direction on how to build causal discovery methods for online learning by extending such approaches.

Orthogonal to the setting of online learning was the privacy consideration that lies at the heart of many real world applications such as healthcare. In such settings we could neither pool data, nor expect it to arrive over time. This introduced its own set of additional challenges and constraints. To develop a causal discovery algorithm for privacy sensitive scenarios, we focused on how we can discover the global causal network without ever sharing any data, model parameters, or even local causal networks— using regrets. We developed a general framework that can be instantiated using any score-based causal discovery approach, and a consistent score therein, all while optimizing over worst-case regret. Crucially, we showed that using the Laplace mechanism on the shared regrets guarantees ϵ-differential privacy. To keep true to our goal of practical causal discovery, we showed that our method discovers causal networks of higher quality than the baseline on both synthetic and real-world data, even in stochastic communication setting, for as many as 100 distributed environments — while requiring orders of magnitude less communication.

Regret-based causal discovery in itself promises to be a powerful notion and opens up avenues for us to harness it further. Going forward we could investigate how we can extend this to learning not just the causal structures but also the model parameters such that we can also learn the underlying SCM across different domains. Naturally, this will have its own set of additional challenges such as assuming that each site has the same network resp. SCM, as well as the consideration on how to maintain privacy guarantees once model parameter communication starts to happen. Furthermore, we could investigate how to extend our approach to discovering causal networks in a setting where unknown interventions exist within each private site.

6.2 Future Research Directions

In this work, we made efforts to develop practically applicable causal discovery approaches. While we have taken steps forwards, a number of questions and practical issues remain wide open. In the following, we highlight some of these existing questions and discuss directions we could take.

An important, looming practical assumption that remains to be addressed is that of causal sufficiency. Neither does this assumption usually hold in practice nor is it trivial to verify. Yet, we need it so that the edges in a learned DAG can be interpreted to carry a causal meaning. In practice, this assumption can be very tricky to handle. One of our proposed approaches, Continent, relaxes this assumption slightly in that it does not require selection variables to

be explicitly observed provided that they are all sink nodes. The latter, however, is only a specific case. There exist approaches to directly check whether selection bias may be present in data (Kaltenpoth and Vreeken, 2023c), as well as to identify if a set of given variables cause a target variable, or if all the variables are jointly confounded (Kaltenpoth and Vreeken, 2019). Recent work from Kaltenpoth and Vreeken (2023a,b) extends this idea further to establish results on assumptions under which we can simultaneously recover confounders and learn a causal network from given observational data. This line of work is a useful launchpad when it comes to developing practically applicable approaches that relax sufficiency assumption. A regret-based algorithm, for example, where we develop a "confounder-aware" score and optimize over worst-case regret is an interesting future direction. Another aspect that can be investigated is using this confounder knowledge to learn causal networks under partial observability, i.e. where different datasets have a partial variable overlap. Then, knowledge of confounders can be consolidated with existing variables to draw inference over the causal structure governing the full set of variables. Investigating whether we can do the same for multiple datasets with unknown interventions, could prove to be a more challenging yet intriguing line of future work.

The methods that we developed in this work are designed to work with continuous-valued data. Extending these to discrete data is straightforward (just use an equivalent MDL score for discrete type data) and may even allow for stronger theoretical guarantees (Budhathoki and Vreeken, 2017). Doing the same for mixed type data, however, is far from simple. In theory, one could develop a sound MDL score to encode variables based on their type. In practice, this might not work An evidence for this was also shown by Marx and Vreeken (2018) where the scale of MDL score could be disparate between continuous and discrete type values. This was due to encoding of "noise" requiring elevated number of bits for continuous valued variables as compared to their discrete counterparts. To the best of our knowledge, this remains an open research question and it is possible that the solution to such a problem lies in investigating existing methods inspired from kernalized conditional independence tests (Fukumizu et al., 2007; Zhang et al., 2014). One such score-based method using kernel regression has been proposed by Huang et al. (2018). While their method can still only give us partially oriented causal networks, the authors note that under their kernelized regression framework the true causal network within the underlying Markov equivalence class frequently ranks higher in terms of "likelihood-score" than its counter-parts. This hints that we may be able define MDL scores based on kernel regression, for mixed type data, to circumvent score disparity across different data types, while still being able to learn a fully oriented causal DAGs.

Among other practical problems that discrete optimization discovery algorithms like Globe and Ges suffer from, is the super exponential growth of

search space with increasing number of variables. It is well known that exhaustive search is NP-hard (Chickering et al., 2004). To have practical, scalable implementations, existing methods employ either edge-greedy (Chickering, 2002; Bühlmann et al., 2014; Mian et al., 2021), or node-greedy (Peters et al., 2014; Squires et al., 2020) search. This comes at the cost of undermined theoretical guarantees, as greedy algorithms are not guaranteed to find the global optimum of an objective unless we place convexity assumptions on the objective plane. Recent work by Zheng et al. (2018b) gives us an alternate way to solve this partly by reformulating the structure learning problem as a purely continuous optimization problem over real matrices, completely avoiding the combinatorial constraint. While this helps with scalability, work needs to be done to prove causal consistency of methods using such approaches. Even though a number of methods exploit this result (Yu et al., 2019b; Kyono et al., 2021), they lack correctness guarantees, as they wrongly assume the acyclicity constraint alone to imply causality. This gives us an engaging research area to explore, where we could investigate developing methods based on continuous optimization *while maintaining consistency guarantees*. An ambitious (maybe even wishful) but exciting direction of work would be to marry approaches that investigate behavior of loss surfaces with approaches for continuous optimization structure learning to see if we can find conditions under which we can find that causal structure which achieves global optimum, using continuous optimization.

In addition to the directions proposed above, several other aspects have a potential for investigation. Causal discovery for time series, for example, poses unique challenges and needs specialized algorithms. Unlike i.i.d. data, temporal data involve a time aspect, making traditional i.i.d. causality definitions philosophically incompatible with those for time series. Causal discovery approaches for time series (Chu et al., 2008; Papana et al., 2016; Nauta et al., 2019; Runge, 2020; Assaad et al., 2022) have been an active area of research lately and could give us challenging new practical problems to investigate. We could say the same can for causal discovery from missing data (Tu et al., 2019; Ma and Zhang, 2021; Gao et al., 2022; Kitson et al., 2023). For missing data, in particular, finding causally consistent data imputations can be an interesting practical application and is an intriguing line of future work.

To conclude, in this dissertation we made a humble attempt to allow the power of causal reasoning to become more applicable to more real world applications. We did so by pointing out practical limitations of existing methods and proposing novel approaches that could circumvent these limitations. Despite our efforts, we still have work to do. We hope to further explore the future work avenues discussed in this chapter in our attempts to continue making causal discovery practically applicable for an even wider range of applications.

Appendix A

Proofs

A.1 Discovering Fully Oriented Causal Networks

Theorem 2.1 *Given a causal model as defined in Eq.* (2.2) *and corresponding data* $\boldsymbol{X}^n$ *drawn iid from joint distribution* P*. Under Assumptions (1) and (2),* $L(\boldsymbol{X}^n, M)$ *asymptotically behaves like BIC.*

Proof: *[Score Consistency]First note that we can rewrite the encoding of the residuals* $L(\epsilon)$ *as*

$$c_1 n \log \hat{\sigma}^2 + \mathcal{O}(1),$$

where the additive constant is independent of the model.

Next, we upper bound $L(M)$*. From Assumption (1) we get that* $|H| \in \mathcal{O}(\log n)$*. Per hinge we need to encode the number of multiplicative terms* $L_{\mathbb{N}}(T_j)$*, the function type per term* $T_j \log |\mathcal{F}|$*, the number of possible assignments from terms to parents* $\log \binom{|\mathcal{S}|+T_j-1}{T_j}$ *and the parameter vector per hinge* $L_p(\theta(h_j))$*. Each parameter vector is constant, by Assumption (2). Since the number of parents are independent of* n *as they are fixed for a certain network, the number of possible interacting terms* T_j *is also constant w.r.t.* n*, which means that for large* n $L_{\mathbb{N}}(T_j)$*,* $T_j \log |\mathcal{F}|$ *(for a finite function class) and* $\log \binom{|\mathcal{S}|+T_j-1}{T_j}$ *are also constants. Since we encode for each non-source node a function where we need to encode each hinge, we get an asymptotic complexity of*

$$c_2 \log n + \mathcal{O}(1).$$

In addition, we need to encode the parents and number of hinges for each node, which adds to the constant term. Combining the above statements, we arrive

at

$$c_1 n \log \hat{\sigma}^2 + c_2 \log n + \mathcal{O}(1) \ .$$

If we set $c_1 = 1$ *and* $c_2 = \frac{d}{2}$*, where* d *is the number of degrees of freedom of the model, we arrive at the BIC score.* □

A.2 Causal Discovery over Multiple Environments

Lemma 3.1 $\forall i, k \ \ HI(X_i^k) \iff pa_i^k = \emptyset$, *and* $SI(X_i^k) \iff pa_i^k \subset pa_i$

Proof: *[Identifiability of Interventions]Assume that we are given the true causal network* G^* *for an SCM as well as the dataset* D^k *over the same SCM for which* $SI(X_i)$ *holds.*

First, we prove the direction $pa_i^k \subset pa_i^* \longrightarrow SI(X_i^k)$*. Assume that* $pa_i^k \subset pa_i^*$ *holds but* $SI(X_i)$ *does not, then* X_i *in* D^k *is calculated as*

$$X_i^k := \sum_{j=1}^{p} f_j^k(\mathcal{S}_j^k) \ , \tag{A.1}$$

with $p = 2^{|pa_i^k|}$ *and* h *and* $\mathcal{S}$ *defined according to our causal model in Section 3.2 of the main text, whereas* X_i *in* D^* *is calculated as*

$$X_i^* := \sum_{j=1}^{q} f_j^*(\mathcal{S}_j^*) \ , \tag{A.2}$$

with $q = 2^{|pa_i^*|}$*. Under our assumption that the causal model does not change unless an intervention is performed, equations* (A.1) *and* (A.2) *should be equal and we can therefore write.*

$$\sum_{j=1}^{p} f_j^k(\mathcal{S}_j^k) = \sum_{j=1}^{q} f_j^*(\mathcal{S}_j^*) \ , \tag{A.3}$$

Without loss of generality, we can re-write r.h.s of the equation. (A.3) *as two summations as follows,*

$$\sum_{j=1}^{p} f_j^k(\mathcal{S}_j^k) = \sum_{j=1}^{p} f_j^*(\mathcal{S}_j^*) + \sum_{r=p+1}^{q} f_r^*(\mathcal{S}_r^*) \ , \tag{A.4}$$

where the summation $\sum_{j=1}^{p}$ *on both sides of the equation, corresponds to the same indices of the generating functions as well as the same corresponding*

subset of parents. The summation over r on the r.h.s of eq. (A.4) *contains all the remaining subsets over the power set of* pa_i^*. *Note that the set of non-linear functions h, over all possible combinations of parents in the power set* $\mathcal{P}(pa_i)$ *of* X_i*'s parents form a basis and therefore are linearly independent, this implies that the first summation term on the r.h.s is equal to the summation on the l.h.s which in turn implies*

$$\sum_{r=p+1}^{q} f_r^*(\mathcal{S}_r^*) = 0 \ .$$

This is possible in one of the two cases: (1) if the basis functions are a linear combination of each other or (2) if the coefficients associated with each of the basis functions is 0. *The former we have already ruled out, whereas the latter implies that the coefficients of all the basis* $f_r^*(\mathcal{S}_r^*)$ *are zero, which implies that there is no edge incoming to* X_i *in* G^* *for this set of parents, which is a contradiction.*

Next we prove the direction $\textsc{Si}(X_i^k) \longrightarrow pa_i^k \subset pa_i^*$ *for Lemma 2. Assume that* $\textsc{Si}(X_i^k)$ *holds,* pa_i^k *are the actual set of* X_i*'s parents in* D^k *after* $\textsc{Si}(X_i^k)$ *but we instead find* pa_i' *such that* $pa_i' = pa_i^*$.

Recall that since we are using regression, our aim for $X_i \in D^k$ *is to minimize*

$$\mathbb{E}\left[\left(X_i - \sum_{j=1}^{q} f_j(\mathcal{S}_j)\right)^2\right] \ .$$

Without loss of generality, we can divide the summation term in two parts, the first part consists of the basis containing only pa_i^k *and the second part consists of the remaining set of basis.*

$$\mathbb{E}\left[\left(X_i - \sum_{j=1}^{p} f_j(\mathcal{S}_j) - \sum_{r=p+1}^{q} f_r(\mathcal{S}_r)\right)^2\right] \ . \tag{A.5}$$

Since the true generating mechanism for X_i *only comprises of basis in the first summation term, we are only left with the noise term* ϵ_i *associated with* X_i. *Hence can further simplify eq.* (A.5) *to*

$$\mathbb{E}\left[\left(\epsilon_i - \sum_{r=p+1}^{q} f_r(\mathcal{S}_r)\right)^2\right] \ . \tag{A.6}$$

The minimum for eq. (A.6) *is achieved when* $\sum_{r=p+1}^{q} f_r(\mathcal{S}_r) = \mathbb{E}(\epsilon_i)$. *By our modelling assumptions, we know that* $\mathbb{E}(\epsilon_i) = 0$. *Therefore, by the same rea-*

soning used to prove reverse direction, we can conclude that the coefficient associated with each of the basis functions in $\sum_{r=p+1}^{q} f_r(\mathcal{S}_r)$ *is zero. This implies that* $pa_i' \subset pa_i^*$*, which is a contradiction.* □

Lemma 3.3 *If* $\boldsymbol{\Upsilon}$ *is conservative,* $\bigcup_{k=1}^{d} G_k = G^*$*, if* $\boldsymbol{\Upsilon}$ *is non-conservative,* $\bigcup_{k=1}^{d} G_k \subseteq G^*$*.*

PROOF: *[Consistency of Orion under Conservative Interventions]If* $\boldsymbol{\Upsilon}$ *is conservative,* $\forall X_i \in \boldsymbol{X}\ \exists D^k \in \boldsymbol{D}$ *such that* $pa_i^k = pa_i^*$*. We get* $\forall X_i\ \bigcup_{k=1}^{d} pa_i^k = pa_i^*$*, which implies that* $\bigcup_{k=1}^{d} \mathcal{E}(G^k) = \mathcal{E}(G^*)$*.*

If $\boldsymbol{\Upsilon}$ *is non-conservative,* $\exists X_i \in \boldsymbol{X}$ *such that* $\forall D^k \in \boldsymbol{D}\ pa_i^k \subset pa_i^*$*. This implies that* $\exists X_i$ *s.t.* $\bigcup_{k=1}^{d} pa_i^k \subseteq pa_i^*$*, which implies that* $\bigcup_{k=1}^{d} \mathcal{E}(G^k) \subseteq \mathcal{E}(G^*)$*.* □

Theorem 3.4 *Let* $\mathcal{Y}$ *be the set of all non-collider nodes. If* $\forall Y_i, k\ \alpha_i^k \to 0$*,* $L(\boldsymbol{D}, M)$ *will be the lowest for the true fully-oriented causal network.*

Theorem 3.5 $L(\boldsymbol{D}, M)$ *correctly identifies the collider structures in the underlying causal network.*

PROOF: *[Score Consistency]We can write* $L(\boldsymbol{D}, M)$ *as*

$$\begin{aligned} L(\boldsymbol{D}, M) &= L_{str}(M) + \sum_{k=1}^{d} L_{mec}(M^k|M) + \sum_{i=1}^{m} L(\epsilon_i^k) \\ &= L_{str}(M) + \sum_{k=1}^{d} \sum_{i=1}^{m} L(f_i^k) + L(\epsilon_i^k) \end{aligned}$$

Since $L_{str}(M)$ *only stores the structure of the global network, which is independent of the number of samples n, therefore it is constant w.r.t n. Hence we get*

$$L(\boldsymbol{D}, M) = \mathcal{O}(1) + \sum_{k=1}^{d} \sum_{i=1}^{m} L_F(f_i^k) + L(\epsilon_i^k)\ .$$

Next, let us look at the cost of encoding a specific environment, k which is given as

$$\sum_{i=1}^{m} L(f_i^k) + L(\epsilon_i^k)\ .$$

Encoding Residuals *Note that we can rewrite the encoding of the residuals* $L(\epsilon)$ *as*

$$b_k n_k \log \hat{\sigma_k}^2 + \mathcal{O}(1) \ ,$$

where the additive constant is independent of the model.

Encoding Functions *Next, we upper bound* $L(f)$*. We get that* $|H| \in \mathcal{O}(\log n)$ *from our assumptions. Per hinge we need to encode the number of multiplicative terms* $L_{\mathbb{N}}(T_j)$*, the function type per term* $T_j \log |\mathcal{F}|$*, the number of possible assignments from terms to parents* $\log \binom{|\mathcal{S}|+T_j-1}{T_j}$ *and the parameter vector per hinge* $L_p(\theta_j)$*. Each parameter vector is constant. Since the number of parents are independent of* n_k *as they are fixed for a certain network, the number of possible interacting terms* T_j *is also constant w.r.t.* n_k*, which means that for large* n_k $L_{\mathbb{N}}(T_j)$*,* $T_j \log |\mathcal{F}|$ *(for a finite function class) and* $\log \binom{|\mathcal{S}|+T_j-1}{T_j}$ *are also constants. In addition, we need to encode the number of hinges for each node, which adds to the constant term. Hence, we can rewrite* $L_F(h)$ *as*

$$c_k \log n_k + \mathcal{O}(1) \ .$$

Combining the residual and function cost for a specific environment, we arrive at

$$b_k n_k \log \hat{\sigma_k}^2 + c_k \log n_k + \mathcal{O}(1) \ .$$

If we set $b_k = 1$ *and* $c_k = \frac{d_k}{2}$*, where* d_k *is the number of degrees of freedom of the model, we arrive at the BIC score.*

Since we compute the same score individually for each environment we can compute the sum over these scores and arrive at

$$L(\boldsymbol{D}, M) = \sum_{k=1}^{d} b_k \log n_k + c_k * n_k \log (\sigma_k^2) + \mathcal{O}(1) \ .$$

□

A.3 Episodic Causal Discovery

Lemma 4.1 (Consistency of L for a single causal model) *For the causal model in assumption 4.1 and assumption 4.2 with R=1 and data* $\mathbf{D}_n$ *over* n *episodes covering each value* s_k *of* S*, with a consistent scoring criterion* L *that decomposes as in Eq. 4.1 then* L *is consistent,*

$$\lim_{n \to \infty} P(\hat{G} \sim G^*) = 1 \ .$$

PROOF: *[Consistency of L for a single causal model end]In the underlying causal model in assumption 4.1 with $R = 1$, we denote the true DAG over $X \cup \{S\}$ as G^*. By consistency of L, we know that when S is observed,*

$$\lim_{n\to\infty} P(G^* \sim \arg\min_G L(X \cup \{S\}; G)) = 1 \ .$$

Using that L is decomposable as in Eq. 4.1, we can write

$$\begin{aligned}
\min_G L(X \cup \{S\}; G) &= \min_{G(X,S)} \Big(L(G(X,S)) + \sum_{j=1}^{m} L(X_j \mid pa_j(G)) + L(S \mid X) \Big) \\
&= \min_{G(X)} \Big(L(G(X)) + \sum_{j=1}^{m} L(X_j \mid pa_j(G)) \Big) + \\
&\quad \min_{G(S|X)} \Big(L(G(S \mid X)) + L(S \mid X) \Big) \\
&= \min_{G(X)} \ L(X; G(X)) + \\
&\quad \min_{G(S|X)} L(S; G(S \mid X))) \ .
\end{aligned}$$

Above, we separated the graph structure G into two subgraphs: $G(X)$ over X, and $G(S \mid X)$ which includes the remaining edges towards S. We can do so as S is a sink node and L is decomposable. Hence, when S is observed, the subgraph $G(X)$ can be identified with our objective. While this holds by construction of our causal where we include S as a sink node, when S is unobserved, we only access a biased $\tilde{X}$ in D_n.

In that case, assume we obtain $\tilde{G} = \min_{G(\tilde{X})} L(\tilde{X}; G)$ with $\tilde{G} \nsim G^*$ and $L(\tilde{X}; \tilde{G}) < L(X; G^*)$. Then for at least one X_j, $pa_j(\tilde{G}) \neq pa_j(G^*)$. We know however that $L(\tilde{X_j} \mid \tilde{pa}_j(G^*)) = L(X_j \mid pa_j(G^*)) < L(X_j \mid pa_j(\tilde{G})) = L(X_j \mid pa_j(\tilde{G}))$ under ignorabiliy in Assumption 4.3, which contradicts that $\tilde{G}$ is a minimizer. Therefore, we also have

$$\lim_{n\to\infty} P(G^* \sim \arg\min_G L(X; G)) = 1 \ .$$

□

Theorem 4.2 (Consistency of L in the episodic setting) *For the causal model in Assumption 4.1 and given data $\mathbf{D}_n$ over n episodes as in Assumption 4.2. Under Assumption 4.3, a consistent scoring criterion L that decom-*

poses as in Eq. 4.1 remains consistent,

$$\lim_{|D_n|\to\infty} P(\hat{G}_r \sim G^*_r) = 1 \qquad \textit{for all } r \in \{1,\dots,R\}\ .$$

PROOF: *[Consistency of L in the episodic setting end]First, in case the context is known, we can apply Lemma 4.1 in each context. That is, given R as well as* $\Pi(D) = \{X^1,\dots,X^R\}$ *into disjoint, non-empty sets* $X^r \subseteq D$ *such that* $\cup_r X^r = D$*, then*

$$\lim_{n\to\infty} P(G^* \sim \sum_{r=1}^{R} \min_{G_r(X)} L(X^r, S^r; G_r(X))) = 1\ .$$

Left to show is the case where R and $\Pi(D)$ *are unknown. We compare*

- *the true model* $\mathbf{G}^* = \{G^*_1,\dots,G^*_{R^*}\}$ *and subsets* $\Pi^*(D) = \{X^{*1},\dots,X^{*R^*}\}$*, and*
- *the estimated model* $\hat{\mathbf{G}} = \{\hat{G}_1,\dots,\hat{G}_{\hat{R}}\}$ *and subsets* $\hat{\Pi}(D) = \{\hat{X}^1,\dots,\hat{X}^{\hat{R}}\}$ *minimizing Eq. 4.4 with score* $L(\hat{\mathbf{G}})$*.*

For contradiction, assume that there is no exact correspondence between the true and estimated models, more precisely, that for at least one context r with true model X^{*r} *and* G^*_r *there is no other* r' *so that* $\hat{X}^{r'} = X^{*r}$ *and* $\hat{G}_{r'} \sim G^*_r$*. We can distinguish the following cases,*

1. *Case* $X^{*r} = \hat{X}^{r'}$ *for some* $r' \neq r$*: then also* $\hat{G}_{r'} \sim G^*_r$ *by Lemma 4.1 as* X^{*r} *is a dataset from a single context r, which however contradicts the above assumption.*
2. *Case* $X^{*r} \subset \hat{X}^{r'}$ *for some* $r' \neq r$*: Then the set* X^{*r} *is wrongly included under the incorrect model* $\hat{G}_{r'}$*. Then the decomposition of Eq. 4.4 will contain a suboptimal likelihood term*

$$L(X^{*r} \mid \hat{G}_{r'}) = \sum_{j=1}^{m} L(X^{*r}_j \mid pa^r_j(\hat{G}_{r'}))\ .$$

Using that L is decomposable, we can replace the above term in the decomposition of L as follows (keeping all other terms the same),

   (a) if $G^*_r \in \hat{\mathbf{G}}$*, we can replace* $L(X^{*r} \mid \hat{G}_{r'})$ *with* $L(X^{*r} \mid G^*_r)$*.*
   (b) if the $G^*_r \notin \hat{\mathbf{G}}$*, we can replace* $L(X^{*r} \mid \hat{G}_{r'})$ *with the full cost* $L(X^{*r}; G^*_r)$ *as the likelihood component dominates over* $L(G^*_r)$ *in the limit Mian et al. (2021).*

In both cases, we can replace $\hat{\mathbf{G}}$ *by* $\hat{\mathbf{G}} \cup \{G^*_r\}$ *and* $\hat{\Pi}(D)$ *by* $\{\hat{X}^1,..,X^{*r},\hat{X}^{r'},...,\hat{X}^{\hat{R}}\}$ *where we separate* X^{*r} *and* $\hat{X}^{r'}$ *and keep all other parts the same, resulting in a favorable model, contradicting that it is the minimizer of Eq. 4.4.*

3. *Case $X \subset \hat{X}^{r'}$ for some $r' \neq r$ and for a set $X \subset X^{*r}, X \neq \emptyset$: This means that a non-empty subset of X^{*r} is included under the incorrect DAG, in which case we can apply the same argument as in case (2).*

We can disregard the case $X^{*r} \cap \hat{X}^{r'} = \emptyset$ for all r' as then X^{*r} is not covered by the partition.

Thus, $\hat{R} = R^*$ and each $\hat{X}^r = X^{*r}$ and $G_r \sim \hat{G}_r$ (up to permuting the indices). □

Theorem 4.3 (Consistency of updating using $\mathcal{T}$) *With discrepancy test $\mathcal{T}$ we will never merge a new episode $D^{(i+1)}$ with a set $\hat{X}^r$ from an incorrect context where $C(D^{(i+1)}) \neq C(E)$ for some $E \in \hat{X}^r$.*

PROOF: *[Consistency of updating using T]We need to show that with a merge protected by $\mathcal{T}$, a merge of $D^{(i+1)}$ with any set $\hat{X}^r$ can only occur if $C(D^{(i')}) = C(D^{(i+1)})$ for all $i' \leq i$. For induction on the time step i, consider the following cases,*

1. *For the base case is $i = 2$, assume $C(D^{(1)}) \neq C(D^{(2)})$. We need to show that $\mathcal{T}$ never merges $D^{(1)}, D^{(2)}$ from C_1, C_2 From our causal model, we know there is at least one variable in G_1^*, G_2^* s.t.*

$$P(X_j^1 \mid pa_j^1) \neq P(X_j^2 \mid pa_j^2)$$

From Cor. 4.5 in Perry et al. (2022), this implies that also for any conditioning set $\mathbf{Z}$,

$$P(X_j^1 \mid \mathbf{Z}^1) \neq P(X_j^2 \mid \mathbf{Z}^2)$$

that is, we have a distribution shift even when $\mathcal{A}$ discovers an incorrect DAG $\hat{G}_1$. Left to show is that it holds also for the biased distributions

$$P(X_j^1 \mid \mathbf{Z}^1, S = s_k) \neq P(X_j^2 \mid \mathbf{Z}^2, S = s_{k'})$$

which holds under detectable selection. Hence, our test T will detect the difference for X_j given enough data from $D^{(1)}, D^{(2)}$ and reject merging.

2. *For the induction step, we can assume that $C(D^{(i')}) = C(D^{(i'')})$ for all i', i'', and apply the above pairwise argument to $D^{(i+1)}$ and each $D^{(i')}$.*

□

Corollary 4.4 (Consistency of Continent) *Given a consistent DAG search algorithm $\mathcal{A}$ and score L, under assumption 4.3 our algorithm is consistent, so that*

$$\lim_{|D_n|\to\infty} P(\hat{G}_r \sim G_{r*}) = 1 \qquad \text{for all } r \in \{1\dots,R\}$$

holds after we obtain n episodes D_n and perform the merge step.

PROOF: *[Consistency of Continent]Consider the estimated model $\hat{\mathbf{G}} = \{\hat{G}_1, \dots, \hat{G}_{\hat{R}}\}$ and subsets $\hat{\Pi}(D) = \{\hat{X}^1, \dots, \hat{X}^{\hat{R}}\}$ that we obtain with PERI at time step n. By the previous theorem, we know that episodes from different contexts were not merged incorrectly, $\hat{X}^r \subseteq X^{*r'}$ for some r' for each r where $\hat{R} \leq R$, which we write shorthand as $\hat{\Pi}(D) \subseteq \Pi^*(D)$. In case $\hat{R} < R$, we need to consider any remaining merges among sets in $\hat{X}^r$. If the assumptions of Thm. 4.2 hold, then we can use*

$$\min_{\Pi(D),\hat{\Pi}(D)\subseteq\Pi(D)} \sum_{r=1}^{|\Pi(D)|} \min_{G_r} L(X^r; G_r)\ .$$

The above will be minimized for $\Pi^*(D)$ and $\hat{G}_r \sim G_{r*}$ for each r as it considers a subset of the partitions that Thm. 4.2 considers. Hence minimizing L is a consistent way to discover the remaining merges. □

A.4 PRIVACY PRESERVING FEDERATED CAUSAL DISCOVERY

Theorem 5.2 *Let G^* be the true underlying causal network for all $P(D^i)$ and let $n^{(1)} \dots, n^{(d)} \to \infty$. Further let L be a consistent and decomposable score. Then*

$$\lim_{n^{(1)},\dots,n^{(d)}\to\infty} P\left(\widehat{G} \sim G^*\right) = 1\,.$$

That is, $\max_i R_i(G)$ is consistent when all $n^{(i)} \to \infty$.

PROOF: *[Peri Consistency]Since L is a consistent score, we know that $\lim_{n^{(i)}\to\infty} P(G_i = G^*) = 1$ for all i. Thus $P\left(\widehat{G} = \operatorname{argmin}_G \max_i \left(L(D^i; G) - L(D^i; G^*)\right)\right) = 1$, which is clearly minimized when $\widehat{G} \sim G^*$.* □

Theorem 5.3 *Let G^* be the true causal network for all $P(D^i)$ and let $N \coloneqq \max_i n^{(i)} \to \infty$. Further let L be a consistent and decomposable score. Then*

$$\lim_{N\to\infty} P\left(\widehat{G} \sqsupseteq G^*\right) = 1\,.$$

PROOF: *[Peri Weak Consistency]When* all $n^{(i)} \to \infty$, *Thm. 5.2 applies. We therefore consider the case where some $n^{(i)}$ remain bounded. Let $I = \left\{i : n^{(i)} < \infty\right\}$ and $M = \max\left\{\limsup n^{(i)} : i \in I\right\}$. Then we have $\max_G \max_{i\in I} R_i(G) \leq cM < \infty$ for some $c > 0$. Meanwhile for all i with $n^{(i)} \to \infty$ we have for all $G \subsetneq G^*$ that*

$$aL(D^i;G) - L(D^i;G_i) \approx L(D^i;G) - L(D^i;G^*) \propto n^{(i)} \to \infty\,.$$

Hence any smaller $G \sqsubset G^*$ achieves strictly worse minmax regret than any $G \sqsupseteq G^*$ as $N \to \infty$. □

Corollary 5.4 *Let the assumptions of Thm. 5.3 hold and let L be the BIC score. Then*

$$\lim_{N\to\infty} P\left(\widehat{G} \sim G^*\right) = 1\,.$$

That is, the score $\max_i R_i(G)$ is consistent when L incorporates a BIC-penalty for parameters and $N \to \infty$.

PROOF: *[Peri BIC Consistency]When L is the BIC score then for any dataset i such that $n^{(i)} \to \infty$ we have $R_i(G) \propto \log(n^{(i)}) \to \infty$ when $G \sqsupset G^*$ is too large. This grows larger than any finite penalty incurred from any of the datasets j with $n^{(j)} \leq M$ bounded, so that picking $\widehat{G} \sim G^*$ will be the best choice as $N \to \infty$.* □

Lemma 5.5 *Assume that $P_i(x;\theta)$ is uniformly lower-bounded bounded by r, i.e., $\forall x \in \mathcal{X}\, \forall \theta \in \Theta\ : P_i(x;\theta) \geq r$, that $\|\theta\| \leq M$ for all local model parameters $\theta \in \Theta$, and that the score L is partially differentiable with respect to θ. Let $X^{(i)}$ and $X'^{(i)}$ be datasets that differ in a single element, i.e. $X^{(i)} \setminus X'^{(i)} = x_k$, θ and θ' the respective local parameters, and $\widehat{R_i}(G)$ and $\widehat{R'_i}(G)$ the respective regrets. Assume that $\|\theta - \theta'\|_1 \leq 2M/n$. Then the sensitivity $\Delta\widehat{R_i}$ of the regret is bounded by*

$$\max\left|\widehat{R_i}(G) - \widehat{R'_i}(G)\right| \leq (4M+1)\log r + \mathcal{O}\left(\frac{\log n}{n}\right)\ .$$

PROOF: *[Bounds on Regret]Removing a single element from a local dataset $X^{(i)}$ changes also the local causal model, both in terms of the DAG $G^{(i)}$ and the local model parameters $\theta^{(i)}$. Therefore, the local score changes for two reasons: (i) the dataset the score is computed on changes, and (ii) the local causal model changes. That is, the sensitivity is*

$$\begin{aligned}\max \left|\widehat{R_i}(G) - \widehat{R_i'}(G)\right| = &\Big|L(X^{(i)}, G) - L(X^{(i)}, G^{(i)}) \\ &-L(X'^{(i)}, G) + L(X'^{(i)}, G'^{(i)})\Big| \\ = &\left|L(X'^{(i)}, G'^{(i)}) - L(X^{(i)}, G^{(i)})\right| \quad .\end{aligned}$$

Thus, it suffices to bound $|L(X', G') - L(X, G)|$ for datasets X and X' that only differ in a single element and corresponding different DAGs G, G' and local model parameters θ, θ'. This difference encompasses both the difference in DAGs and local model parameters. Since the difference in DAGs is determined by the difference of θ and θ', we for convenience write $L(X, G) = L(X, \theta)$ and show that the difference $|L(X, \theta) - L(X', \theta')|$ is bounded. Since

$$\begin{aligned}|L(X, \theta) - L(X', \theta')| \leq &|L(X', \theta) - L(X, \theta)| \\ &+ \|\theta - \theta'\| \, |L(X, \theta') - L(X, \theta)| \quad ,\end{aligned}$$

we can use the linearization of L and get

$$\begin{aligned}|L(X, \theta) - L(X', \theta')| \leq &\underbrace{|L(x_k, \theta)|}_{\leq \log r} \\ &+ \|\theta - \theta'\| \underbrace{|L(X, \theta) - L(X, \theta')|}_{\propto n} \\ &+ \underbrace{\|\theta - \theta'\|}_{\leq 2M/n} |L(\theta) - L(\theta')| + \mathcal{O}\left(\frac{\log n}{n}\right) \\ \leq &\log r + 2M \log r + 2M \log r + \mathcal{O}\left(\frac{\log n}{n}\right) \\ = &(4M + 1) \log r + \mathcal{O}\left(\frac{\log n}{n}\right) \quad .\end{aligned}$$

It follows that the sensitivity is bounded by $(4M + 1) \log r + \mathcal{O}(\log n/n)$. Note that the assumption $|\theta - \theta'| \leq 2M/n$ for θ, θ' optimized on datasets that only differ in a single element holds for most learning algorithms, e.g., convex empirical risk minimization with finite VC-dimension or Rademacher complexity Von Luxburg and Schölkopf (2011). □

Proposition 5.6 *Assume that each local regret $\widehat{R}_i$ has sensitivity $\leq Q$. Then* PERI *with i.i.d. Laplace noise with scale $\lambda = Q/\epsilon$ added to each $\widehat{R}_i$ is ϵ-differentially private.*

PROOF: *[Peri Differentiable Privacy]The Laplace mechanism guarantees that adding noise with mean 0 and scale λ to a function f with sensitivity δf is $\delta f/\lambda$-differentially private. Since the regret has sensitivity $(4M+1)\log r + \mathcal{O}\left(\frac{\log n}{n}\right)$, choosing $\lambda = \epsilon^{-1}\left((4M+1)\log r + \mathcal{O}\left(\log n/n\right)\right)$ results in a sensitivity of*

$$\frac{\delta R}{\lambda} = \frac{(4M+1)\log r + \mathcal{O}\left(\frac{\log n}{n}\right)}{\epsilon^{-1}\left((4M+1)\log r + \mathcal{O}\left(\log n/n\right)\right)} = \epsilon \ .$$

□

Bibliography

H. Akaike. A new look at the statistical model identification. *IEEE TAC*, 19 (6):716–723, 1974. ISSN 0018-9286.

K. AN. Sulla determinazione empirica di una legge didistribuzione. *Giorn Dell'inst Ital Degli Att*, 4:89–91, 1933.

T. V. Anand, A. H. Ribeiro, J. Tian, and E. Bareinboim. Causal effect identification in cluster dags. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12172–12179, 2023.

C. K. Assaad, E. Devijver, and E. Gaussier. Survey and evaluation of causal discovery methods for time series. *Journal of Artificial Intelligence Research*, 73:767–819, 2022.

E. Bareinboim and J. Pearl. Controlling selection bias in causal inference. In *AISTATS*, volume 22, pages 100–108. PMLR, 2012.

E. Bareinboim, J. Tian, and J. Pearl. Recovering from selection bias in causal and statistical inference. *AAAI*, 28(1), 2014.

P. Blöbaum, D. Janzing, T. Washio, S. Shimizu, and B. Schölkopf. Cause-effect inference by comparing regression errors. In *International Conference on Artificial Intelligence and Statistics*, pages 900–909. PMLR, 2018a.

P. Blöbaum, D. Janzing, T. Washio, S. Shimizu, and B. Schölkopf. Cause-effect inference by comparing regression errors. In *International Conference on Artificial Intelligence and Statistics*, pages 900–909. PMLR, 2018b.

P. Boeken, N. de Kroon, M. de Jong, J. M. Mooij, and O. Zoeter. Correcting for selection bias and missing response in regression using privileged information. In *UAI*, pages 195–205. PMLR, 2023.

P. Brouillard, S. Lachapelle, A. Lacoste, S. Lacoste-Julien, and A. Drouin. Differentiable causal discovery from interventional data. *Advances in Neural Information Processing Systems*, 2020.

K. Budhathoki and J. Vreeken. Mdl for causal inference on discrete data. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 751–756. IEEE, 2017.

P. Bühlmann, J. Peters, J. Ernest, et al. Cam: Causal additive models, high-dimensional order search and penalized regression. *Annals Stat.*, 42(6):2526–2556, 2014.

Charlatan. Correlation, 2013. URL `https://www.reddit.com/r/geek/comments/1kjuxt/correlation/`.

D. M. Chickering. Optimal structure identification with greedy search. *JMLR*, 3:507–554, 2002.

M. Chickering, D. Heckerman, and C. Meek. Large-sample learning of bayesian networks is np-hard. *JMLR*, 5, 2004.

T. Chu, C. Glymour, and G. Ridgeway. Search for additive nonlinear time series causal models. *Journal of Machine Learning Research*, 9(5), 2008.

D. Colombo and M. H. Maathuis. Order-independent constraint-based causal structure learning. *JMLR*, 15(1):3741–3782, 2014.

S. Compton, M. Kocaoglu, K. Greenewald, and D. Katz. Entropic causal inference: Identifiability and finite sample results. *arXiv preprint arXiv:2101.03501*, 2021.

G. F. Cooper and C. Yoo. Causal discovery from a mixture of experimental and observational data. *UAI*, 1999.

J. Correa and E. Bareinboim. General transportability of soft interventions: Completeness results. *Advances in Neural Information Processing Systems*, 33:10902–10912, 2020a.

J. Correa and E. Bareinboim. A calculus for stochastic interventions: Causal effect identification and surrogate experiments. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 10093–10100, 2020b.

D. Deutsch. Quantum theory, the church-turing principle and the universal quantum computer. 400(1818):97–117, 1985.

D. Dua and C. Graff. Uci machine learning repository. 2017. URL `http://archive.ics.uci.edu/ml`.

C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

D. Eaton and K. Murphy. Exact bayesian structure learning from uncertain interventions. In *AISTATS*, pages 107–114. PMLR, 2007.

G. R. A. Faria, A. Martins, and M. A. Figueiredo. Differentiable causal discovery under latent interventions. In *Conference on Causal Learning and Reasoning*, pages 253–274. PMLR, 2022.

D. E. Farrar and R. R. Glauber. Multicollinearity in regression analysis: the problem revisited. *The Review of Economic and Statistics*, 1967.

J. H. Friedman. Multivariate adaptive regression splines. *The annals of statistics*, pages 1–67, 1991.

K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. *Advances in neural information processing systems*, 20, 2007.

E. Gao, J. Chen, L. Shen, T. Liu, M. Gong, and H. Bondell. Federated causal discovery. *arXiv preprint arXiv:2112.03555*, 2021.

E. Gao, I. Ng, M. Gong, L. Shen, W. Huang, T. Liu, K. Zhang, and H. Bondell. Missdag: Causal discovery in the presence of missing data with continuous additive noise models. *Advances in Neural Information Processing Systems*, 35:5024–5038, 2022.

J. Geiping, H. Bauermeister, H. Dröge, and M. Moeller. Inverting gradients - how easy is it to break privacy in federated learning? In *Advances in Neural Information Processing Systems*, volume 33, pages 16937–16947. Curran Associates, Inc., 2020.

M. Ghanbari, J. Lasserre, and M. Vingron. The distance precision matrix: computing networks from non-linear relationships. *Bioinformatics*, 35(6): 1009–1017, 08 2018.

A. Ghassami, S. Salehkaleybar, N. Kiyavash, and K. Zhang. Learning causal structures using regression invariance. *arXiv preprint arXiv:1705.09644*, 2017.

C. Glymour, K. Zhang, and P. Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 2019.

A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *ALT*. Springer, 2005.

A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, and B. Schölkopf. Covariate Shift by Kernel Mean Matching. In *Dataset Shift in Machine Learning*. The MIT Press, 12 2008. ISBN 9780262255103.

P. Grünwald. *The Minimum Description Length Principle*. MIT Press, 2007.

D. M. Haughton. On the choice of a model to fit data from an exponential family. *Annals Math. Stat.*, 16(1):342–355, 1988.

A. Hauser and P. Bühlmann. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *JMLR*, 13 (1):2409–2464, 2012.

A. Hauser and P. Bühlmann. Jointly interventional and observational data: Estimation of interventional markov equivalence classes of directed acyclic graphs. *J. R. Statist. Soc. B*, 77, 03 2013.

A. Hauser and P. Bühlmann. Two optimal strategies for active learning of causal models from interventional data. *International Journal of Approximate Reasoning*, 55(4):926–939, jun 2014.

P. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In *NeurIPS*, volume 21. Curran, 2009a.

P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In *NIPS*, pages 689–696, 2009b.

S. Hu, Z. Chen, V. Partovi Nia, L. CHAN, and Y. Geng. Causal inference and mechanism clustering of a mixture of additive noise models. In *NeurIPS*, 2018.

B. Huang, K. Zhang, Y. Lin, B. Schölkopf, and C. Glymour. Generalized score functions for causal discovery. In *KDD*. ACM, 2018.

A. Jaber, M. Kocaoglu, K. Shanmugam, and E. Bareinboim. Causal discovery from soft interventions with unknown targets: Characterization and learning. *Advances in neural information processing systems*, 33, 2020.

D. Janzing and B. Schölkopf. Causal inference using the algorithmic markov condition. *IEEE TIT*, 56(10):5168–5194, 2010a.

D. Janzing and B. Schölkopf. Causal inference using the algorithmic markov condition. *IEEE TIT*, 56(10):5168–5194, 2010b.

D. Kalainathan and O. Goudet. Causal discovery toolbox: Uncover causal relationships in python. *arXiv preprint arXiv:1903.02278*, 2019.

M. Kalisch and P. Bühlmann. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *JMLR*, 8(Mar):613–636, 2007.

D. Kaltenpoth and J. Vreeken. We are not your real parents: Telling causal from confounded using MDL. In *SDM*, pages 199–207. SIAM, 2019.

D. Kaltenpoth and J. Vreeken. Causal discovery with hidden confounders using the algorithmic markov condition. In *Uncertainty in Artificial Intelligence*, pages 1016–1026. PMLR, 2023a.

D. Kaltenpoth and J. Vreeken. Nonlinear causal discovery with latent confounders. In *International Conference on Machine Learning*, pages 15639–15654. PMLR, 2023b.

D. Kaltenpoth and J. Vreeken. Identifying selection bias from observational data. *AAAI*, pages 8177–8185, 2023c.

N. K. Kitson, A. C. Constantinou, Z. Guo, Y. Liu, and K. Chobtham. A survey of bayesian network structure learning. *Artificial Intelligence Review*, 56(8): 8721–8814, 2023.

M. Kocaoglu, K. Shanmugam, A. Jaber, and E. Bareinboim. Characterization and learning of causal graphs with latent variables from soft interventions. *Advances in neural information processing systems*, 2019.

A. Kolmogorov. Three approaches to the quantitative definition of information. *Problemy Peredachi Informatsii*, 1(1):3–11, 1965.

L. G. Krafft. *A device for quantizing, grouping, and coding amplitude-modulated pulses*. PhD thesis, Massachusetts Institute of Technology, 1949.

T. Kyono, Y. Zhang, A. Bellot, and M. van der Schaar. Miracle: Causally-aware imputation via learning missing data mechanisms. *Advances in Neural Information Processing Systems*, 34:23806–23817, 2021.

S. L. Lauritzen. Propagation of probabilities, means, and variances in mixed graphical association models. *Journal of the American Statistical Association*, 87(420):1098–1108, 1992.

S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(2):157–194, 1988.

S.-Y. Lee and K.-L. Tsui. Covariance structure analysis in several populations. *Psychometrika*, 47, 1982.

M. Li and P. Vitányi. *An Introduction to Kolmogorov Complexity and its Applications*. Springer, 2009.

R. Little and D. Rubin. Statistical analysis with missing data, third edition. 04 2019. doi: 10.1002/9781119482260.

N. Y. Lu, K. Zhang, and C. Yuan. Improving causal discovery by optimal bayesian network learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8741–8748, 2021.

L. Lyu and C. Chen. A novel attribute reconstruction attack in federated learning. *arXiv preprint arXiv:2108.06910*, 2021.

C. Ma and C. Zhang. Identifiable generative models for missing not at random data imputation. *Advances in Neural Information Processing Systems*, 34: 27645–27658, 2021.

C. Ma, J. Li, M. Ding, H. H. Yang, F. Shu, T. Q. Quek, and H. V. Poor. On safeguarding privacy and security in the framework of federated learning. *IEEE network*, 34(4):242–248, 2020.

S. Magliacane, T. van Ommen, T. Claassen, S. Bongers, P. Versteeg, and J. M. Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. In *NIPS*, volume 31, 2018.

S. Mameche, D. Kaltenpoth, and J. Vreeken. Discovering invariant and changing mechanisms from data. In *KDD*, page 12421252. ACM, 2022.

S. Mameche, D. Kaltenpoth, and J. Vreeken. Learning causal mechanisms under independent changes. In *NeurIPS*, 2023.

D. Margaritis and S. Thrun. Bayesian network induction via local neighborhoods. In *NIPS*, pages 505–511, 2000.

A. Marx and J. Vreeken. Telling Cause from Effect using MDL-based Local and Global Regression. In *ICDM*, pages 307–316. IEEE, 2017.

A. Marx and J. Vreeken. Causal inference on multivariate and mixed-type data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 655–671. Springer, 2018.

A. Marx and J. Vreeken. Identifiability of cause and effect using regularized regression. In *KDD*. ACM, 2019.

A. Marx and J. Vreeken. Formally justifying mdl-based inference of cause and effect. *arXiv preprint arXiv:2105.01902*, 2021.

C. Meek. Causal inference and causal explanation with background knowledge. In *UAI*, pages 403–410. Morgan Kaufmann Publishers Inc., 1995.

O. Mian and S. Mameche. An information theoretic framework for continual learning of causal networks. PMLR, 2024.

O. Mian, A. Marx, and J. Vreeken. Discovering fully oriented causal networks. In *AAAI*, 2021.

O. Mian, D. Kaltenpoth, and M. Kamp. Regret-based federated causal discovery. In *The KDD'22 Workshop on Causal Discovery*, pages 61–69. PMLR, 2022.

O. Mian, D. Kaltenpoth, M. Kamp, and J. Vreeken. Nothing but regretsprivacy-preserving federated causal discovery. In *International Conference on Artificial Intelligence and Statistics*, pages 8263–8278. PMLR, 2023a.

O. Mian, M. Kamp, and J. Vreeken. Information-theoretic causal discovery and intervention detection over multiple environments. In *Proceedings of the AAAI Conference on Artificial Intelligence, AAAI-23*, 2023b.

O. Mian, S. Mameche, and J. Vreeken. Learning causal networks from episodic data. In *KDD*. ACM, 2024.

K. Mohan, J. Pearl, and J. Tian. Graphical models for inference with missing data. *Advances in neural information processing systems*, 26, 2013.

J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Scholkopf. Distinguishing cause from effect using observational data: Methods and benchmarks. *ArXiv*, abs/1412.3773, 2014.

J. M. Mooij, S. Magliacane, and T. Claassen. Joint causal inference from multiple contexts. *JMLR*, 21, 2016.

M. Nauta, D. Bucur, and C. Seifert. Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction*, 1(1):19, 2019.

I. Ng and K. Zhang. Towards federated bayesian network structure learning with continuous optimization. In *International Conference on Artificial Intelligence and Statistics*, 2022.

A. Papana, C. Kyrtsou, D. Kugiumtzis, and C. Diks. Detecting causality in non-stationary time series using partial symbolic transfer entropy: Evidence in financial data. *Computational economics*, 47:341–365, 2016.

J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition, 2009.

J. Pearl. A solution to a class of selection bias problems. 2012.

J. Pearl, T. Verma, et al. A theory of inferred causation. *KR*, 91:441–452, 1991.

R. Perry, J. Von Kügelgen, and B. Schölkopf. Causal discovery in heterogeneous environments under the sparse mechanism shift hypothesis. *Advances in Neural Information Processing Systems*, 35:10904–10917, 2022.

J. Peters and P. Bühlmann. Structural intervention distance for evaluating causal graphs. *Neural computation*, 27(3):771–799, 2015.

J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. Causal discovery with continuous additive noise models. *JMLR*, 15, 2014.

J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *J. R. Statist. Soc. B*, pages 947–1012, 2016.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT Press, 2017.

J. Ramsey, M. Glymour, R. Sanchez-Romero, and C. Glymour. A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *J. Data Sci. Anal.*, 2017.

A. Reisach, C. Seiler, and S. Weichwald. Beware of the simulated dag! causal discovery benchmarks may be easy to game. *Advances in Neural Information Processing Systems*, 34, 2021.

J. Rissanen. Modeling by shortest data description. *Automatica*, 14(1):465–471, 1978.

J. Rissanen. A universal prior for integers and estimation by minimum description length. *Annals Stat.*, 11(2):416–431, 1983.

D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

J. Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Conference on Uncertainty in Artificial Intelligence*, pages 1388–1397. Pmlr, 2020.

K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

Y. Sakamoto, M. Ishiguro, and G. Kitagawa. Akaike information criterion statistics. *Dordrecht, The Netherlands: D. Reidel*, 81(10.5555):26853, 1986.

G. Schwarz. Estimating the dimension of a model. *Annals Stat.*, 6(2):461–464, 1978.

S. Shimizu. Joint estimation of linear non-gaussian acyclic models. *Neurocomputing*, 81, 2012.

S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen. A linear non-gaussian acyclic model for causal discovery. *JMLR*, 7, 2006.

R. Shokri, M. Stronati, C. Song, and V. Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2017.

K. Singhal, H. Sidahmed, Z. Garrett, S. Wu, J. Rush, and S. Prakash. Federated reconstruction: Partially local federated learning. *Advances in Neural Information Processing Systems*, 34, 2021.

N. Smirnov. Table for estimating the goodness of fit of empirical distributions. *The annals of mathematical statistics*, 19(2):279–281, 1948.

L. Solus, Y. Wang, L. Matejovicova, and C. Uhler. Consistency guarantees for permutation-based causal inference algorithms. *arXiv preprint arXiv:1702.03530*, 2017.

P. Spirtes, C. Meek, and T. Richardson. An algorithm for causal inference in the presence of latent variables and selection bias. *Computation, causation, and discovery*, 21, 1999.

P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman. *Causation, prediction, and search*. MIT Press, 2000a.

P. Spirtes, C. N. Glymour, R. Scheines, D. Heckerman, C. Meek, G. Cooper, and T. Richardson. *Causation, prediction, and search*. MIT press, 2000b.

C. Squires, Y. Wang, and C. Uhler. Permutation-based causal structure learning with unknown intervention targets. In *Conference on Uncertainty in Artificial Intelligence*, pages 1039–1048. PMLR, 2020.

A. Statnikov, S. Ma, M. Henaff, N. Lytkin, E. Efstathiadis, E. R. Peskin, and C. F. Aliferis. Ultra-scalable and efficient methods for hybrid observational and experimental local causal pathway discovery. *JMLR*, 16:3219–3267, 2015.

A. L. Steiner, A. J. Davis, S. Sillman, R. C. Owen, A. M. Michalak, and A. M. Fiore. Observed suppression of ozone formation at extremely high temperatures due to chemical and biophysical feedbacks. *Proceedings of the National Academy of Sciences*.

M. Sugiyama, M. Krauledat, and K.-R. Müller. Covariate shift adaptation by importance weighted cross validation. *J. Mach. Learn. Res.*, 8:9851005, dec 2007. ISSN 1532-4435.

Z. Sun, P. Kairouz, A. T. Suresh, and H. B. McMahan. Can you really backdoor federated learning? *arXiv preprint arXiv:1911.07963*, 2019.

R. E. Tillman. Structure learning with independent non-identically distributed data. In *ICML*, pages 1041–1048, 2009.

S. Triantafillou and I. Tsamardinos. Constraint-based causal discovery from multiple interventions over overlapping variable sets. *JMLR*, 16(1):2147–2205, 2015.

I. Tsamardinos, L. E. Brown, and C. F. Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78, 2006.

R. Tu, C. Zhang, P. Ackermann, K. Mohan, H. Kjellström, and K. Zhang. Causal discovery in the presence of missing data. In K. Chaudhuri and M. Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1762–1770. PMLR, 16–18 Apr 2019. URL `https://proceedings.mlr.press/v89/tu19a.html`.

U. Von Luxburg and B. Schölkopf. Statistical learning theory: models, concepts, and results. In *Inductive Logic*, volume 10 of *Handbook of the History of Logic*, pages 651–706. Elsevier, 2011.

L. Wang, Q. Pang, and D. Song. Towards practical differentially private causal graph discovery. *Advances in Neural Information Processing Systems*, 33: 5516–5526, 2020.

R. Xiong, A. Koenecke, M. Powell, Z. Shen, J. T. Vogelstein, and S. Athey. Federated causal inference in heterogeneous observational data. *arXiv preprint arXiv:2107.11732*, 2021.

S. Xu, O. A. Mian, A. Marx, and J. Vreeken. Inferring cause and effect in the presence of heteroscedastic noise. In *International Conference on Machine Learning*, pages 24615–24630. PMLR, 2022.

K. Yang, A. Katcoff, and C. Uhler. Characterizing and learning equivalence classes of causal dags under interventions. In *ICML*, pages 5541–5550. PMLR, 2018.

Q. Ye, A. A. Amini, and Q. Zhou. Distributed learning of generalized linear causal networks. *arXiv preprint arXiv:2201.09194*, 2022.

I.-C. Yeh and T.-K. Hsu. Building real estate valuation models with comparative approach through case-based reasoning. *Applied Soft Computing*, 65: 260–271, 2018.

K. Yu, L. Liu, J. Li, W. Ding, and T. D. Le. Multi-source causal feature selection. *IEEE TPAMI*, 2019a.

Y. Yu, J. Chen, T. Gao, and M. Yu. Dag-gnn: Dag structure learning with graph neural networks. In *International conference on machine learning*, pages 7154–7163. PMLR, 2019b.

K. Zhang, J. Peters, D. Janzing, and B. Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *UAI*, 2014.

K. Zhang, B. Huang, J. Zhang, C. Glymour, and B. Schölkopf. Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination. In *IJCAI*, 2017.

X. Zheng, B. Aragam, P. K. Ravikumar, and E. P. Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, 31, 2018a.

X. Zheng, B. Aragam, P. K. Ravikumar, and E. P. Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018b.

L. Zhu and S. Han. Deep leakage from gradients. In *Federated learning*, pages 17–31. Springer, 2020.