Causal Learning under Distribution Shifts using Independent Change of Mechanisms

A dissertation submitted towards the degree

Doctor of Natural Sciences
(Dr. rer. nat.)

of the Faculty of Mathematics and Computer Science
of Saarland University

by
Sarah Mameche

Saarbrücken, 2026

UNIVERSITÄT DES SAARLANDES

Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form in einem Verfahren zur Erlangung eines akademischen Grades vorgelegt.

Declaration of original authorship

I hereby declare that this dissertation is my own original work except where otherwise indicated. All data or concepts drawn directly or indirectly from other sources have been correctly acknowledged. This dissertation has not been submitted in its present or similar form to any other academic institution either in Germany or abroad for the award of any other degree.

Saarbrücken, ***May 2026***

gez. / signed

Sarah Mameche

Abstract

Changes in data distribution can be a source of bias in statistical and machine learning, but they are common in most empirical sciences. In healthcare, predictive models trained on one population may not generalize to other subgroups with genetic or socio-economic differences; in genetics, gene expression measurements under baseline conditions differ from those after a genetic perturbation, e.g., with CRISPR–Cas editing; and in the environmental sciences, measurements often come from non-stationary processes changing over time. In these settings, the standard learning assumption of independently and identically distributed samples from the population may not be realistic.

In this thesis, we address such settings using assumptions from causality, specifically the notions of modularity and independent change of causal mechanisms. We first introduce a causal model to address heterogeneous data from multiple contexts. In this context we propose (i) a scoring criterion for score-based learning under Gaussian process additive noise models, (ii) a causal discovery algorithm in topological order, and (iii) a statistical test for latent confounding. Then, we move the ideas to (iv) non-stationary time series with unknown temporal changepoints, and finally consider (v) datasets that combine heterogeneous subgroups with distinct generating processes. As real-world applications of our methods, we consider learning cell signaling pathways in flow cytometry data and detecting regime shifts in data from hydrology and meteorology.

Together, these results motivate the use of ideas from causality towards achieving robust learning under changing data-generating conditions.

Zusammenfassung

In der Statistik und dem maschinellen Lernen gehen verzerrte oder voreingenommene Resultate häufig auf strukturelle Unterschiede in der Datenverteilung zurück, die allerdings in den meisten empirischen Wissenschaften zu finden sind. In der Medizin zum Beispiel können prädiktive Modelle, die auf einer bestimmten Population trainiert wurden, möglicherweise nicht auf andere Untergruppen mit genetischen oder sozioökonomischen Unterschieden verallgemeinert werden; in der Biologie unterscheiden sich Genexpressionsmessungen unter normalen Bedingungen von denen nach einer Intervention, etwa durch CRISPR-Cas-Editierung; und in den Umweltwissenschaften stammen Messungen von nicht-stationären Prozessen, die sich mit der Zeit ändern können. In solchen Situationen ist die Standardannahme des statistischen Lernens, dass Stichproben unabhängig und identisch verteilt sind, oft nicht realistisch.

In dieser Dissertation untersuchen wir solche Szenarien unter Verwendung von Annahmen aus der Kausalitätsforschung, insbesondere der Modularität kausaler Mechanismen und deren Veränderung unabhängig voneinander. Zunächst formulieren wir letzteres in einem kausalen Modell mit dem Ziel, heterogene Daten aus unterschiedlichen Kontexten zu modellieren. Um solche Modelle zu lernen, schlagen wir (i) Methoden zu lokalem Lernen vor, unter geeigneten Annahmen (Gaussian Process Additive Noise Models), (ii) damit einhergehend einen Lernalgorithmus für kausale Graphen, sowie (iii) eine Methode, unbekannte Störvariablen (Confounders) zu finden. Wir befassen uns außerdem mit (iv) temporalen Datensätzen mit zeitlichen Änderungszeitpunkten, und schließlich mit (v) Datensätzen, die heterogene Untergruppen mit unterschiedlichen generativen Prozessen kombinieren. Wir zeigen auch reale Anwendungen unserer Methoden in der Biologie und den Umweltwissenschaften.

Diese Beiträge motivieren eine Forschungsrichtung an der Schnittstelle von der Verteilungsverschiebung und dem kausalen Lernen, um robustes Lernen unter sich verändernden Bedingungen zu ermöglichen.

Acknowledgements

This is a preliminary short version of the acknowledgments.

I would like to thank my advisor, Jilles Vreeken, for his advice, scientific support, and feedback. I am very grateful to each of my coauthors for their collaboration and a productive and supportive environment. I would also like to thank the members of the reviewing committee for agreeing and taking the time to review this thesis.

Contents

1 | Informal Introduction

Case Study Population Paradox I

A case report of COVID-19 Delta variant cases in the UK (Public Health England, 2021) shows vaccination status (0/1) and case fatality (0/1) for diagnosed individuals, as well as the percentage of deaths among these cases (Case Fatality Rate (CFR), %). Overall, the fatality rate appears higher for vaccinated individuals than for the unvaccinated population (Figure 1.1, left), yet for both individuals younger than 50 and those aged 50 and above, vaccination lowers the fatality rate (Figure 1.1, middle/right).

Many statistical and machine learning methods focus on predictive problems under independently and identically distributed (i.i.d.) sampling, assuming observations from a single, homogeneous population. In practice, this assumption is often unrealistic, as data may be collected from heterogeneous populations, across distinct locations, or under multiple experimental conditions. Combining such data while assuming they are i.i.d. can lead to biased conclusions, for example through misleading correlations.

The case study above shows an instance of Simpson's paradox (Simpson, 1951) where vaccination appears to be associated with increased fatality in the aggregated data, while it is associated with reduced fatality within each age group (Figure 1.1). This discrepancy is due to differences in age composition across vaccination groups. Individuals aged 50 and above are more likely to be vaccinated, but at the same time also face higher baseline disease risk (e.g., Pezzullo, 2022). As a result, the

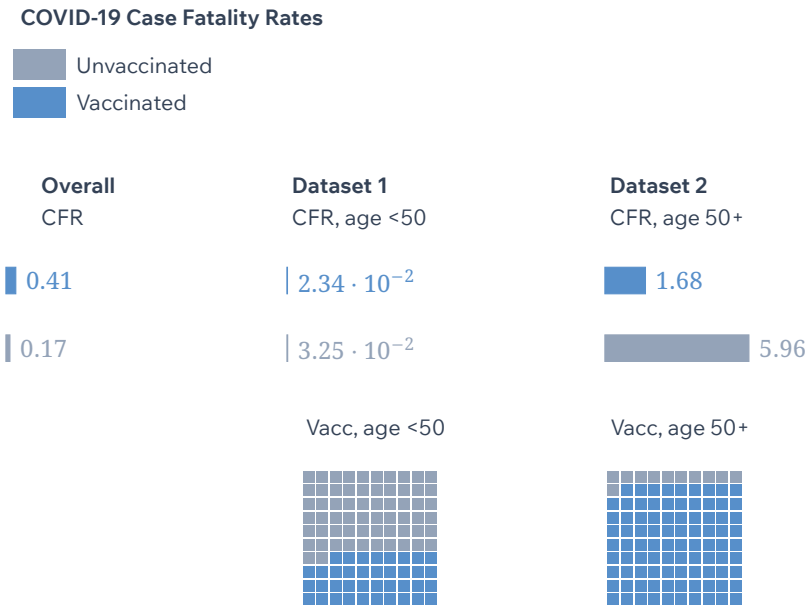


Figure 1.1 In a dataset over Vaccination (0/1) and Case Fatality (0/1) in COVID-19 (Public Health England, 2021), the Case Fatality Rate (%), i.e., proportion of deaths per diagnosed cases, appears higher among vaccinated individuals (left). Within two age groups, CFR is lower among vaccinated individuals (top right). The reversal occurs due to differences in vaccination rates across age groups (bottom).

combined data does not accurately reflect the true effect of vaccination.

This phenomenon is not specific to the previous example but arises more broadly as distribution shifts and heterogeneous sampling are common in practice. In observational studies and clinical trials, different populations may have genetic or socioeconomic differences. In economics, statistical relationships change over time due to market cycles, policy interventions, or geopolitical events. Similar challenges also arise in experimental sciences, such as single-cell biology, where measurements are taken under targeted interventions that modify system components, often with unknown effects (e.g., Dominguez et al., 2016). We use the term context to refer to datasets collected under different conditions.

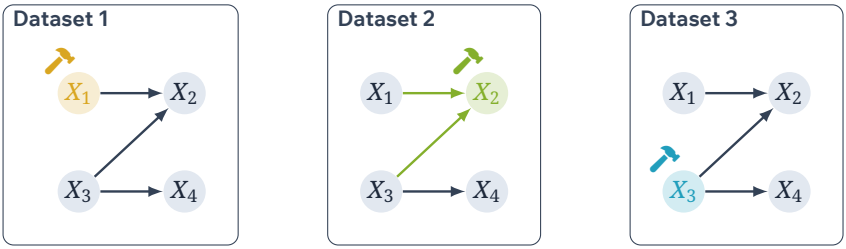


Figure 1.2 Example illustration of the multi-context setting that is the main focus of this thesis. Each observed dataset corresponds to a different context (boxes) in which distribution shifts of variables occur (hammers). In general, we assume these distribution shifts may target an unknown set of variables.

To illustrate this setting, we give a simplified depiction in Figure 1.2 showing multiple contexts that could correspond to multiple hospitals, measuring the same variables but serving different populations, following different procedures, or using different measurement technologies. While each dataset may individually satisfy the i.i.d. assumption, this no longer holds for their combination. At the same time, keeping the datasets separate while analysing them jointly under appropriate models can reveal structural insights about the underlying data-generating process to domain experts. For example, of particular interest to us will be the questions of which causal relationships exist (arrows) and where changes in their generating mechanisms occur (hammers).

Against this background, rather than treating data as arising from a single homogeneous process, we instead assume that different parts of the data may be generated by different mechanisms throughout this thesis. This requires statistical models that can encode such mechanisms and support reasoning about them, which is possible through causal models (Pearl, 2009).

1.1 Causal Models for Non-i.i.d. Data

To reason about systems such as the one illustrated in Figure 1.2, we consider statistical models that explicitly represent changes in the data-

Figure 1.3 A causal model for the motivating example (Population Paradox I) over the variables vaccination (Vacc) and Case Fatality (CF), with a third variable Risk that acts as a confounder in the combined data (left), and remains fixed when separating the population into two age groups with similar risk (middle, right).

generating process. Structural causal models (SCMs) do so by representing variables as outputs of underlying mechanisms such as physical laws, biological processes, or other generative relationships.

Causal Models. To illustrate, we revisit the motivating example, where we consider a causal model in the form of a directed acyclic graph with edges pointing from causes to effects, here for Vaccination (*Vacc*) to Case Fatality (*CF*) (Figure 1.3, left). In this case, the baseline disease risk of an individual (*Risk*) is an additional variable relevant to the analysis. Disease risk is likely related to vaccination status because individuals at higher risk may be prioritized in vaccination programs, and it also affects case fatality through higher susceptibility, thereby confounding the association between *Vacc* and *CF*. In graphical terms, *Risk* opens a back-door path (Pearl, 2009) between *Vacc* and *CF*.

Age is not itself the confounder in this simplified setting, but stratifying by age may act as a proxy adjustment if baseline risk is approximately homogeneous within each age group. Under this assumption, conditioning on age removes the confounding path (Figure 1.3, middle/right). In this example, stratification resolves the confounding effect because the confounder remains approximately fixed within each dataset. Such a setting is sometimes referred to as pseudo-confounding (Huang et al., 2020). Note that other forms of confounding exist, which we address in a later chapter (Chapter 5).

Causal Discovery. In general, it is often reasonable to assume that there exists an underlying causal graphical structure over the observed variables which remains the same across contexts (e.g., Mooij et al., 2020; Huang et al., 2020), while the associated causal mechanisms change across con-

texts. However, the preceding discussion assumes that the relevant causal graph is known. In general, the existence and direction of causal effects are often not given a priori; for example, regarding the effectiveness of a novel vaccine candidate.

Establishing causality typically relies on randomized controlled trials (RCTs), which can, however, be time-intensive, expensive, or infeasible due to ethical concerns.

Furthermore, in some settings, even when interventional data are available, their scale makes manual analysis impractical. In cell biology, for example, measurements under targeted perturbations can be collected using technologies such as flow cytometry (Sachs et al., 2005). Such interventional datasets can be large-scale, making direct inspection difficult.

This motivates the problem of learning causal models algorithmically from data, known as causal discovery (Squires & Uhler, 2022). At the same time, a fundamental challenge in causal discovery is that observational data generally does not identify a unique causal graph. Conditional independence testing, in particular, provides only partial causal information and is itself a difficult problem (Shah & Peters, 2020). Other approaches therefore impose additional assumptions on the functional forms of causal mechanisms and noise terms (Peters et al., 2014; Hoyer et al., 2009) that may be difficult to verify in practice.

The key idea we will use in the following chapters is that, perhaps surprisingly, by embracing the availability of multiple datasets with distribution shifts, we can uncover additional structure of the underlying causal model (e.g., Schölkopf et al., 2021).

1.2 Using Non-i.i.d. Data for Causal Learning

Besides modeling distribution shifts through causal models, we pursue a second goal of using distribution shifts to draw conclusions about cause-effect relationships. In particular, heterogeneous distributions over multiple contexts are often more informative for causal discovery than a single i.i.d. distribution.

Invariance and Independence of Mechanisms. A prominent idea in causality is that the mechanism generating an effect from its direct causes remains invariant across contexts or environments (Peters et al., 2016). In Figure 1.4, top, for example, the mechanism for X_4 and corresponding cause-effect conditional $X_4 \mid X_3$ remain invariant, even when the input

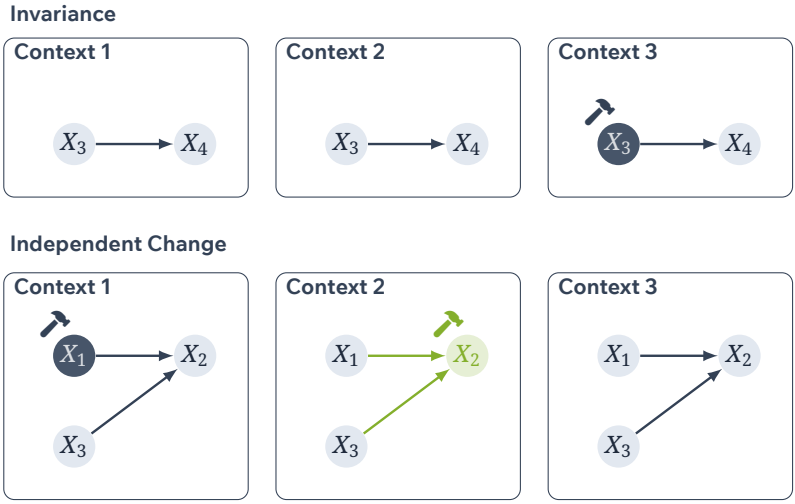


Figure 1.4 Illustration of causal invariance and independent changes in the example from Figure 1.2. *Top:* Causal mechanisms for one variable, such as X_4 , are invariant (gray arrows) to changes in other variables (gray hammers). *Bottom:* Only when a variable is directly manipulated, such as X_2 (colored hammer) does its causal mechanism change. In turn, such changes leave other mechanisms invariant, and are therefore called independent changes.

distribution of its cause X_3 changes in Context 3 (gray). However, interventions can induce changes in causal mechanisms, for example a change of X_2 in Context 2 (colored). In this work, we assume that changes in mechanisms are independent; a change in one mechanism does not induce changes in others.

We state this assumption informally as follows.

Assumption 1.1 (Independent Changes of Causal Mechanisms, informal). *The causal generative process over the set of variables X is composed of autonomous modules that neither inform nor influence each other. In particular, any two mechanisms f_i and f_j generating the variables X_i and X_j, respectively, change independently and without sharing information.*

The above is the main structural assumption we will use to relax the

i.i.d. assumption. More broadly, Independent Change (IC) is based on the principle of independence of causal mechanisms (ICM) (Schölkopf et al., 2021). While IC and related ideas have been explored previously (Schölkopf et al., 2021; Peters et al., 2017) including for causal structure learning (Perry et al., 2022; Guo et al., 2022), we will show that its use extends to many settings; from the detection of latent confounding, to time series where mechanism shifts occur at unknown changepoints, to mixtures of heterogeneous population groups where mechanism shifts occur for an unknown subpopulation.

Algorithmic Independence of Mechanisms. Throughout the chapters, we primarily use the second part of Assumption 1.1, namely that causal mechanisms change without sharing information. We will formulate this using the notion of algorithmic independent change (Janzing & Schölkopf, 2010). Intuitively, the idea is to view mechanisms as programs mapping input to output distributions, which allow concise descriptions of effects from their direct causes, without needing information from other parts of a system.

This raises the question of how this property is useful to address multi-context settings such as in Figure 1.2. In a nutshell, the correct causal model admits a representation in which mechanism changes remain sparse and mutually independent across contexts. In contrast, incorrect or anti-causal representations generally require more coordinated changes across variables. This idea underlies the score-based methods developed in Chapters 3-4 for causal discovery from multi-context data. We also adapt this approach to time series data in Chapter 6.

In later chapters, we revisit these assumptions and extend the analysis to other problem settings. One common assumption is causal sufficiency, stating that all relevant variables are directly observed. In Chapter 5 we relax this assumption, and in particular show that IC also has uses for the detection of latent confounding variables. Finally, another assumption we make up to and including Chapter 6 is that data is a priori separated into different datasets corresponding to contexts. While reasonable in settings where data stems from distinct sources such as hospitals, this is not always the case. For example, within a single population, unknown subpopulations may have a genetic difference or a different subtype of a given disease. Such a setting will be the focus of the final Chapter 7.

1.3 Thesis Overview

To summarize, the introduction has raised two problems, firstly statistical learning under violations of the i.i.d. assumption, and secondly learning causal rather than purely correlational structure. We also motivated studying the intersection of these two problems, since causal models support principled reasoning about distribution shifts, and conversely non-i.i.d. data can provide additional information for causal identifiability and learning. Therefore, the topic of this thesis is causal learning from non-i.i.d. data.

The main contributions toward this goal are the following.

Chapter 2 We continue with a formal introduction of the problem setting and concepts described intuitively in this chapter.

Chapter 3 Given non-i.i.d. data from multiple contexts, we introduce an approach to discover causal edge orientations as well as causal mechanism shifts. We propose a local scoring criterion and show that it identifies causal directions theoretically and empirically.

Chapter 4 In the same setting, we propose a causal discovery algorithm in topological order that exploits properties of certain information theoretic scores, including the local scoring criterion for multi-context data. We study under which conditions this algorithm discovers the underlying causal model, supported by experimental evaluations.

Chapter 5 We move to a setting where latent confounding variables exist, and introduce a statistical test for their detection under independent and sparse mechanism changes.

Chapter 6 We adapt the analysis to non-stationary time series, proposing an algorithm for joint discovery of causal graphs and changepoints, along with real-world case studies.

Chapter 7 Finally, we broaden the setting to mixtures of different data sources composed of unknown subgroups with distinct data-generating processes. We propose a causal model with unknown mixtures of structural causal mechanisms and study its inference using score-based learning.

We give a visual overview of these topics in Figure 1.5. All chapters are based on published articles (Table 1.1), and implementations of the main algorithms are available online (Table H.2).

Collaborations. All publications presented in this thesis were collaborative projects. As lead author of the publications in Chapter 3 and Chapter 5, I contributed to theory and writing and conceived the algorithms and evaluations. Dr. David Kaltenpoth provided supervision and contributed to the theoretical results in Sections 3.2 and 5.2. Chapter 4 is based on work with Sascha Xu. Both authors contributed to all aspects of the paper, where Sascha focused on the i.i.d. case and my focus was on the non-i.i.d. case. Chapter 6 is based on work supervised by Dr. Lénaïg Cornanguer and Dr. Urmi Ninad, together with whom I jointly developed the method. My contributions focused on the theory, presentation, and experiments, with feedback and supervision from the co-authors. The publication in Chapter 7 is a collaboration with Dr. Janis Kalofolias who developed the theoretical results in Section 7.2 as well as contributed to the ideation and manuscript writing. As the supervisor of this thesis, Prof. Dr. Jilles Vreeken contributed substantially to all ideas and research directions.

In addition, I contributed to work published as Mian et al. (2024), which considers the problem of causal discovery in episodic data, where the approach was conceived by Dr. Osman Mian and I contributed to the theory and presentation, which we omit from the presentation for topical cohesion.

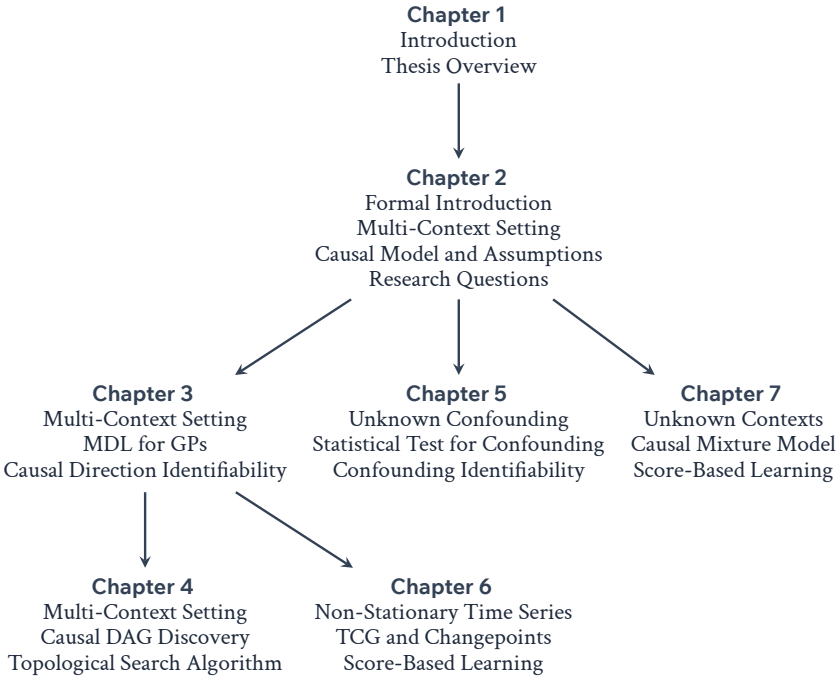


Figure 1.5 Overview of the chapters in this thesis and their reading dependency.

Chapter	Publication
Chapter 3	Mameche, S., Kaltenpoth, D., and Vreeken, J. Discovering Invariant and Changing Mechanisms from Data. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2022.
	Mameche, S., Kaltenpoth, D., and Vreeken, J. Learning Causal Models under Independent Changes. *Neural Information Processing Systems (NeurIPS)*, 2023.
Chapter 4	Xu, S.°, Mameche, S.°, and Vreeken, J. Information-Theoretic Causal Discovery in Topological Order. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2025.
Chapter 5	Mameche, S., Vreeken, J., and Kaltenpoth, D. Identifying Confounding from Causal Mechanism Shifts. In *International Conference on Artificial Intelligence and Statistics (AISTATS),* 2024.
Chapter 6	Mameche, S., Cornanguer, L., Ninad, U., and Vreeken, J. SpaceTime: Causal Discovery from Non-Stationary Time Series. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.
Chapter 7	Mameche, S., Kalofolias, J., and Vreeken, J. Causal Mixture Models: Characterization and Discovery. *Neural Information Processing Systems (NeurIPS)*, 2025.

Table 1.1 List of publications that each chapter is based on. ○ *marks shared first authorship.*

Chapter	Directory	Usage Examples
Ch. 3-4	`src/causalchange`	`demo/ch4_causal.ipynb`
Ch. 5	`src/coco`	`demo/ch5_confounding.ipynb`
Ch. 6	`src/stime`	`demo/ch6_time.ipynb`
Ch. 7	`src/causalchange`	`demo/ch7_cmm.ipynb`

Table 1.2 Python implementations of the main algorithms, maintained online at `github.com/srhmm/thesis`.

2 | Formal Introduction

In this chapter, we formalize the intuitions from Chapter 1. We introduce the general notation and terminology used throughout this dissertation and provide a chapter-wise overview in Appendix A.

2.1 Problem Setting

We start by describing the multi-context problem setting that is the focus of Chapters 3-4 and forms the basis for later chapters.

We consider a set of observed random variables $\mathcal{X} = \{X_1, \dots, X_d\}$ with joint distribution $P_{\mathcal{X}}$. We assume that $P_{\mathcal{X}}$ admits a density $p_{\mathcal{X}}$ with respect to some product measure, denoted by p whenever clear from context. We write $P_{X|\mathcal{X}_S}$ to denote the conditional distribution for $X \in \mathcal{X}$ and $\mathcal{X}_S \subseteq \mathcal{X}$, also written as $P(X \mid \mathcal{X}_S)$ for readability, with density $p_{X|\mathcal{X}_S}$. We denote realizations of $\mathcal{X}$ by $\boldsymbol{x} = (x_1, \dots, x_d) \in \mathbb{R}^d$.

A common assumption in machine learning and causality is that observations are obtained i.i.d. from a fixed distribution $P_{\mathcal{X}}$. Here, we instead assume that observations may stem from different distributions $P^c_{\mathcal{X}}$ indexed by c. We assume there exist m such contexts $c = 1, \dots, m$, sometimes called environments in the literature. For each context c, we observe a dataset $\mathcal{D}^c = \left\{\boldsymbol{x}^{c,1}, \dots, \boldsymbol{x}^{c,n_c}\right\}$ consisting of n_c realizations $\boldsymbol{x}^{c,i} \in \mathbb{R}^d$, which are drawn i.i.d. according to the distribution $P^c_{\mathcal{X}}$ with density $p^c_{\mathcal{X}}$.

This results in a collection of datasets $\boldsymbol{\mathcal{D}} = \{\mathcal{D}^1, \dots, \mathcal{D}^m\}$.

We model the data-generating process in this setting through a shared

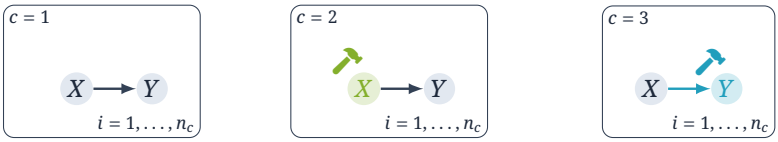


Figure 2.1 Example generating process for a cause X and effect Y observed in multiple contexts, in which the mechanism of each variable may change.

causal graph where causal mechanisms depend on the context. Figure 2.1 illustrates this through a bivariate example ($X \to Y$). While the causal structure remains constant, the structural mechanism f_Y undergoes a shift in one context, while the mechanism for X remains unaffected.

More generally, we assume there exists a common causal graphical structure (Lauritzen, 1996; Pearl, 2009) over the observed variables $\mathcal{X}$ in all contexts c, represented by a directed acyclic graph (DAG) $G = (\mathcal{X}, \mathcal{E})$. The nodes $\mathcal{X}$ correspond to the observed variables and the set of directed edges $\mathcal{E}$ encodes causal relationships, with $(X_k, X_j) \in \mathcal{E}$ whenever X_k is a direct cause of X_j. The set of direct parents of a node X_j in G is denoted by $Pa_j^G \subseteq \mathcal{X} \setminus \{X_j\}$, or simply Pa_j when the graph is clear from the context.

Under this graph, the structural mechanism associated with a variable X_j may change across contexts c. Accordingly, we assume there exist multiple functions

$$\{f_j^1, \ldots, f_j^{k_j}\},$$

and that in context c the structural equation for X_j is given by

$$X_j^c = f_j^{\pi_j^c}\left(Pa_j, N_j\right) \ , \tag{2.1}$$

where π_j^c indicates which causal function applies in context c. Above, N_j is an exogenous noise variable. As is standard, we assume that $N_1, \ldots, N_d$ are jointly independent and do not depend on the context.

More formally, π_j^c is a realization of a random variable Π_j. This latent variable takes values in

$$\Pi_j \in \{1, \ldots, k_j\}$$

to indicate which of the k_j possible mechanisms is active for X_j in each given context c.

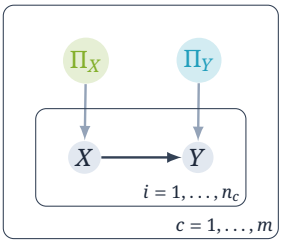


Figure 2.2 Example causal model for a cause X and effect Y observed in m contexts. In context c, the active causal mechanism f_Y^k for Y is indicated by $\pi_Y^c = k$ such that $Y = f_Y^k(X, N_Y)$, analogously for X.

Collectively, we denote these variables as $\Pi = \{\Pi_1, \ldots, \Pi_d\}$, and refer to them as mechanism assignments. To include Π in the graphical representation, we augment the graph G by including one latent node Π_j for each X_j in the graph, together with an edge $\Pi_j \to X_j$. This results in an augmented causal DAG $G_{\Pi,\mathcal{X}}$ over the variables $\mathcal{X} \cup \Pi$, where we assume that there are no edges in the subgraph over Π.

We denote the overall causal model as $\mathcal{M} = (G, \Pi)$. We collapse the data-generating process shown in the example of Figure 2.1 to the process shown in Figure 2.2. For each context c, we sample a mechanism assignment $\Pi \sim P_\Pi$, resulting in π_j^c for all j. Conditional on these assignments, we generate n_c i.i.d. observations from the context-specific SCM in Eq. (2.1).

As is common in causal discovery problems (Pearl, 2009), discovering the causal model from observations is only possible under additional assumptions, which we introduce next.

2.2 Assumptions

Below we introduce the main assumptions related to the multi-context causal model. A chapter-wise overview of the assumptions used later is given in Appendix C–Appendix G.

Throughout, we assume the data-generating process introduced above with structural equations in Eq. (2.1) and causal model $\mathcal{M} = (G, \Pi)$. We

make the common assumptions of the causal Markov property, faithfulness, and sufficiency, in our setting with the augmented graph over $\mathcal{X}$ and Π. The causal Markov property states that the joint density factorizes into conditional densities of each variable given its parents in the augmented causal graph.

Assumption 2.1 (Causal Markov Property)**.** *We assume the causal Markov property,*

$$p(\boldsymbol{x}, \boldsymbol{\pi}) = p(\boldsymbol{\pi}) \prod_{j=1}^{d} p(x_j \mid \boldsymbol{pa}_j, \pi_j) \ . \tag{2.2}$$

For each context c, conditioning on the mechanism assignment $\boldsymbol{\pi}^c$ induces the context-specific distribution

$$p^c(\boldsymbol{x}) = p(\boldsymbol{x} \mid \boldsymbol{\pi}^c) = \prod_{j=1}^{d} p\left(x_j \mid \boldsymbol{pa}_j, \pi_j^c\right) \ .$$

The dataset $\mathcal{D}^c$ then consists of n_c i.i.d. observations from p^c.

We further assume that the joint distribution over all variables is faithful to the augmented graph.

Assumption 2.2 (Causal Faithfulness)**.** *All conditional independences in the joint distribution over $(\mathcal{X}, \Pi)$ correspond to $d-$separations in $G_{\Pi,\mathcal{X}}$.*

The above uses the standard definition of $d-$separation which we defer to Appendix B.

Third, a common assumption is causal sufficiency which states that no unobserved confounding variables exist.

Assumption 2.3 (Causal Sufficiency)**.** *There are no unobserved confounders, i.e., there are no unobserved common causes over $(\mathcal{X}, \Pi)$ in $G_{\Pi,\mathcal{X}}$.*

We will relax the sufficiency assumption in Chapter 5.

In addition to these assumptions that are standard in the causal discovery literature (Pearl, 2009), our work relies on further structural assumptions about how causal mechanisms change across contexts. We first state the assumption that changes occur independently (Assumption 1.1) which underlies several of our results.

Assumption 2.4 (Independent Change). *The mechanism assignment variables are mutually independent,*

$$\Pi_j \perp\!\!\!\perp \Pi_k \qquad \text{for all } j \neq k \ .$$

For intuition, when sampling a context at random we expect the active causal mechanism of one variable in that context to not inform which is the active mechanism of another. Since the variables Π are exogenous source nodes in the augmented causal model, the causal Markov property implies the factorization

$$p(\boldsymbol{\pi}) = \prod_{j=1}^{d} p(\pi_j) \ .$$

We state this property explicitly in Assumption 2.4 since it is the central asymmetry used in later identifiability results.

As a counterpart to the causal faithfulness condition, we assume a notion of faithfulness regarding distribution shifts, called change faithfulness. This states the mechanism assignment variables encode all changes in the observed distributions.

Assumption 2.5 (Change Faithfulness under Π). *The variables Π correspond faithfully to changes in the underlying causal mechanisms,*

$$\pi_j^c \neq \pi_j^{c'} \iff P^c(X_j \mid Pa_j) \neq P^{c'}(X_j \mid Pa_j)$$

for any contexts c, c' and variable X_j.

The above ensures that changes in Π_j correspond exactly to changes in the conditional $X_j \mid Pa_j$ associated with the causal mechanism f_j^k. Similar assumptions have been made in related works (Squires et al., 2020).

Independence of mechanism shifts alone allows scenarios where the mechanisms of two variables X_j and X_k either always change across all contexts or never change at all. Such cases provide no information about the underlying causal structure. We therefore impose an additional sparsity assumption on mechanism shifts (Perry et al., 2022) which ensures that changes occur neither too rarely nor too frequently.

To formalize this notion, we quantify how often the conditional distribution of a variable changes across contexts. Specifically, we consider two independent random draws C, C' from an underlying (unspecified)

distribution over contexts to define the probability that the corresponding conditionals differ.

Assumption 2.6 (Sparse Change). *Let C, C' be contexts drawn i.i.d. from a distribution, and let*

$$p_j^{\text{chng}} = Pr\left(P^C(X_j \mid Pa_j) \neq P^{C'}(X_j \mid Pa_j)\right) ,$$

then we assume $0 < p_j^{\text{chng}} < \frac{1}{2}$ for each X_j.

This assumption ensures that, while each variable changes with non-zero probability (lower bound), the causal mechanisms remain invariant in the majority of context pairs (upper bound), thus the assumption can be viewed as a relaxation of Causal Invariance (Peters et al., 2016).

The preceding assumptions address the statistical frequency and independence of changes. A slightly different idea underlies the algorithmic perspective of causation (Janzing & Schölkopf, 2010; Lemeire & Janzing, 2013) which restricts the complexity of the causal mechanisms themselves. For example, if $X \to Y$ is the causal direction, then the shortest description of the mechanism generating X should not contain information that helps shorten the description of the mechanism generating Y from X.

More generally, the algorithmic framework of causation characterizes causal mechanisms as programs mapping inputs (causes) to outputs (effects), with their description length measured by the Kolmogorov complexity K (Kolmogorov, 1968). Given a binary string $x \in \{0,1\}^*$, its Kolmogorov complexity $K(x)$ is the length of the shortest program r that outputs x on a universal Turing machine $\mathcal{U}$ and halts. For distributions p, $K(P_Y)$ is the length of the shortest program that approximates P_Y to arbitrary precision q, $K(P_Y) = \min_{r \in \{0,1\}^*} \{|r| : \forall y, |\mathcal{U}(r, y, q) - P_Y(y)| \leq \frac{1}{q}\}$.

The algorithmic mutual information between objects x and y is defined as $I_A(x; y) = K(x) + K(y) - K(x, y)$, similarly for distributions. For multiple objects, we write $I_A(p_1; \ldots; p_s) \stackrel{+}{=} 0$ to denote pairwise algorithmic independence, i.e., $I_A(p_i; p_j) \stackrel{+}{=} 0$ for $i \neq j$, where independence holds up to an additive constant independent of the inputs, i.e., up to a program of size $O(1)$.

The algorithmic version of the causal Markov condition (Janzing & Schölkopf, 2010) states that this independence holds for the causal conditionals,

$$I_A\left(\{P(X_j \mid Pa_j, \Pi_j = k)\}_{j,k}\right) \stackrel{+}{=} 0 \qquad \text{for } j = 1, \ldots, d,\ k = 1, \ldots, k_j .$$

The following formulation expresses the same postulate in terms of the description length.

Assumption 2.7 (Algorithmic Independent Change). *A model $\mathcal{M} = (G, \Pi)$ is a valid causal hypothesis only if the shortest description of the observed distribution P_X decomposes into independent descriptions of each causal conditional,*

$$K(P(X)) \overset{+}{=} \sum_{j=1}^{d} \sum_{k=1}^{k_j} K\big(P(X_j \mid Pa_j^G, \Pi_j = k)\big) \ .$$

In this view, if mechanisms are truly independent, the shortest program to describe them jointly should simply be the concatenation of the shortest individual programs. This will be the main basis of our score-based causal discovery approaches (Chapters 3-4 and Chapter 6). As Kolmogorov complexity above is not computable in practice, we approximate $K(\cdot)$ using the description length under a specific functional model class (Assumptions 3.1-3.1).

2.3 Contributions

We conclude this chapter by stating the main research questions addressed in this dissertation, grouped into three main topics.

Our first research question addresses how to learn the above causal model from observations.

Question 1 (Multi-Context Causal Discovery). *Given $\mathcal{D} = \{\mathcal{D}^1, \dots, \mathcal{D}^m\}$ over m contexts, we are interested in discovering both the underlying causal graph and the causal mechanisms per context.*

The above states our main causal discovery problem. For comparison, existing work in causal discovery from multiple contexts (e.g., Mooij et al., 2020; Huang et al., 2020; Squires et al., 2020; Perry et al., 2022) often introduces constraint-based methods. We take a different perspective and study the idea of algorithmic independent change (Assumption 2.7) within a score-based approach, proposing a local scoring criterion for causal discovery. We first focus on discovering mechanism shifts and using these to identify causal directions between pairs of variables (Chapter 3), then extend these ideas to discovering causal DAGs (Chapter 4).

In Chapter 6, we move from tabular to time-series data, answering the question of how to jointly learn a causal DAG, mechanisms, and their temporal regimes. Most existing work either addresses regime detection in time series witt considering multiple contexts (Saggioro et al., 2020), or allows multi-context time series but no shifts over time (Günther et al., 2023b), and only recently addresses both aspects jointly (Rahmani & Frossard, 2025). We do so using a score-based discovery approach and an EM-style algorithm. In Chapter 6 we state the problem setting as well as the temporal version of the causal graph in more detail.

In Chapter 5 and Chapter 7, we relax or modify assumptions underlying the previous tasks. In Chapter 5 we challenge the assumption that no latent confounding variables exist (Assumption 2.3).

Question 2 (Unknown Confounders). *Given $\mathcal{D} = \{\mathcal{D}^1, \dots, \mathcal{D}^m\}$ over m contexts, where a set of unobserved confounding variables exists, we are interested in discovering the confounders, the causal graph, and any mechanism changes.*

Previous work on detecting confounders often relies on latent-variable modeling under restrictions of the model class (Kaltenpoth, 2024). Our contribution is to show that mechanism shifts alone allow identifying confounding without specific model class restrictions.

Finally, all aforementioned approaches assume observations from multiple contexts. This corresponds to applications where observations come from distinct sources, such as different medical centers. In other cases, such sources are not a priori separated from each other and distribution shifts can occur within a single dataset, such as a population where unknown population subgroups show a different disease subtype or treatment response. To address this, we consider the question of disentangling multiple causal mechanisms for each observed variable X_j.

Question 3 (Unknown Contexts). *Given a single dataset $\mathcal{D}$, we are interested in discovering a causal graph over $\mathcal{X}$, as well as mixture components associated with different causal mechanisms for each observed variable.*

In Chapter 7 we state this problem setting in more detail.

3 | Causal Direction Discovery from Multiple Contexts

Case Study Population Paradox (revisited)

Given patient data from medical centers located in three different countries (Figure 3.1, left), when combining observations across countries, vaccination rates with a novel vaccine for a given disease (X) appear to have a positive correlation with the rate of new infections with the disease (Y). When each country is considered separately, however, higher vaccination rates are associated with fewer infections (Figure 3.1, right).

In this chapter, we study causal discovery from multi-context data, following the framework introduced in the previous chapter.

The above case study illustrates once more the observation from the introductory chapter that aggregating data across contexts can result in misleading correlations. While Figure 3.1 shows a toy example, such effects arise in real-world settings. For the COVID-19 case fatality rates, for example, von Kügelgen et al. (2021) point out inconsistencies in cross-country comparisons due to differences in national age demographics.

Having introduced a causal model to address such settings in a more principled way (Chapter 2), we now address the problem of discovering this model from multi-context observations.

The chapter is based on work published as Mameche et al. (2022; 2023).

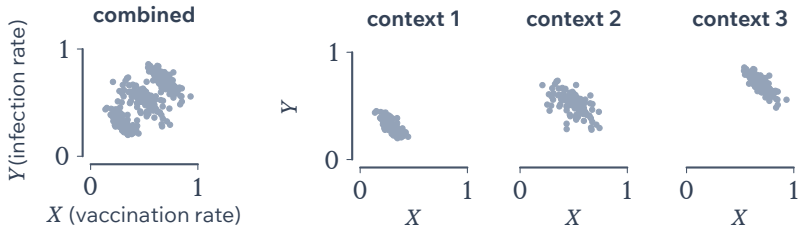


Figure 3.1 Given multi-context data over vaccination rate X and infection rate Y with a disease from different countries $c = 1, \ldots, 3$, vaccinations appear overall positively correlated with new infections (left), whereas per-context samples are negatively correlated (right). Note that the causal mechanisms are not invariant across the contexts in the example, but differ in bias terms.

We base our approach conceptually on Algorithmic Independent Change (Assumption 2.7) which we instantiate using the Minimum Description Length (MDL) principle together with Gaussian process (GP) regression models with additive noise. We introduce this scoring criterion in Section 3.1 alongside identifiability guarantees in Section 3.2 that we base on the IC (Assumption 2.4). For ease of exposition, in Section 3.3 we focus on discovering causal edge directions and causal mechanism shifts, rather than the full graph structure, and return to the latter problem in Chapter 4. We conclude with experimental evaluations in Section 3.5.

3.1 Gaussian Process-based Scoring Criterion

As the basis for our approach, we rely on the algorithmic version of independent change (Assumption 2.7), which states that a valid causal hypothesis admits a description that decomposes into independent descriptions of its causal mechanisms.

As Assumption 2.7 is formulated in terms of Kolmogorov complexity K, which is not computable due to the halting problem (Li & Vitányi, 2009), we instead consider a restricted model class $\mathcal{H}$ and the Minimum Description Length principle (Grünwald, 2007). MDL measures the number of bits required to encode data under a given model, balancing goodness of fit against model complexity within a fixed model class.

We model each causal mechanism nonparametrically using a GP addi-

tive noise model (Rasmussen & Williams, 2005). Concretely, for variable X_j and mechanism index k, we assume

$$X_j = f_j^k(Pa_j) + N_j, \qquad N_j \sim \mathcal{N}(0, \sigma^2), \tag{3.1}$$

where f_j^k is assumed to lie in a GP function class with mean function m and kernel κ. Thus, for all contexts c with $\pi_j^c = k$, observations of X_j are assumed to share the same f_j^k.

We consider the standard radial basis function (RBF) kernel,

$$\kappa(x_i, x_j) = \lambda \exp(-\tfrac{1}{2l} \left\| x_i - x_j \right\|_2^2)$$

with length-scale parameter l and $\lambda > 0$, as it induces a function class dense in the space of continuous functions (Rasmussen & Williams, 2005).

Given variable X_j and mechanism index k, let

$$\boldsymbol{x}_j^{(k)} = \left\{ x_j^{c,i} \,\middle|\, \pi_j^c = k \right\}, \qquad \boldsymbol{pa}_j^{(k)} = \left\{ \boldsymbol{pa}_j^{(c,i)} \,\middle|\, \pi_j^c = k \right\}$$

denote the pooled observations of X_j and its parents from all contexts c with $\pi_j^c = k$ for $c = 1, \ldots, m, i = 1, \ldots, n_c$.

The corresponding MDL score (Grünwald, 2007; Kakade et al., 2005) is given by

$$L\left(\boldsymbol{x}_j^{(k)} \mid \boldsymbol{pa}_j^{(k)}; \mathcal{H}\right) = \min_{f \in \mathcal{H}_\kappa} \Bigg(\underbrace{\|f\|_\kappa^2}_{\text{complexity}} \underbrace{- \log p\left(\boldsymbol{x}_j^{(k)} \mid \boldsymbol{pa}_j^{(k)}\right)}_{\text{fit}} \Bigg) + R\left(\boldsymbol{pa}_j^{(k)}\right). \tag{3.2}$$

where $p\left(\boldsymbol{x}_j^{(k)} \mid \boldsymbol{pa}_j^{(k)}\right)$ denotes the conditional likelihood of the pooled response observations given the pooled parent observations. The regret term

$$R\left(\boldsymbol{pa}_j^{(k)}\right) = \tfrac{1}{2} \log \det(\sigma^{-2} K_{\boldsymbol{pa}_j^{(k)}} + I)$$

is a penalty depending on the choice of the parent set, and $K_{\boldsymbol{pa}_j^{(k)}}$ denotes the Gram matrix of κ applied to the pooled parent observations $\boldsymbol{pa}_j^{(k)}$.

By the independence in Assumption 2.7, the overall score is

$$L(\mathcal{D}; \mathcal{M}) = L(\mathcal{D}; G, \Pi) = \sum_{j=1}^{d} \sum_{k=1}^{k_j} L\left(\boldsymbol{x}_j^{(k)} \mid \boldsymbol{pa}_j^{(k)}; \mathcal{H}\right). \tag{3.3}$$

The score above decomposes over variables as well as over the mechanism groups induced by Π. For each (j, k), the score term is computed from the pooled observations of X_j and its parents across all contexts c satisfying $\pi_j^c = k$.

Our objective is therefore to recover the causal model $\mathcal{M}$ consisting of a graph G and mechanism assignments Π that minimize the scoring criterion,

$$\mathcal{M} = (G, \Pi) \in \arg\min_{G,\Pi} L(\mathcal{D}; G, \Pi) \ . \tag{3.4}$$

Similar instantiations of the score L using regularized regression exist under other model classes, for example parametric models (Mameche et al., 2022), polynomial regression (Marx & Vreeken, 2019b), and spline regression (Mian et al., 2021). For these, identifiability guarantees are known in the i.i.d. (single-context) setting. Here, we build upon and extend these methods, focusing on the GP-based instantiation and basing identifiability on the multi-context setting in particular. While in the single-context setting, identifiability relies on additive noise asymmetries (Bühlmann et al., 2014; Marx & Vreeken, 2019b), in the multi-context setting we can obtain a different approach under independent change.

3.2 Identifiability Guarantees

Next we address under which conditions the minimizers (G, Π) of Eq. (3.4) correspond to the underlying causal model denoted by (G^*, Π^*).

Besides the generic assumptions in Chapter 2, we need the following assumptions regarding the functional model class.

Assumption 3.1 (Additive Noise Model)**.** *For each X_j we assume that the structural equation model in Eq. (2.1) is an additive noise model according to Eq. (3.1).*

Under the ANM, we assume that the GP model class is well-specified.

Assumption 3.2 (Sufficient Capacity)**.** *For each X_j and each structural causal mechanism f_j^k in Eq. (3.1), there exists a function in the model class $\mathcal{H}$ that can express f_j^k.*

We give an overview of the assumptions in Appendix C, where we also include all detailed theoretical results while we provide proof sketches in the main text.

3.2.1 Recovery up to Markov Equivalence

We begin by showing that score minimization recovers the Markov equivalence class (MEC) of the underlying DAG, whose formal definition we provide in Appendix B.

Theorem 3.1 (Identifiability of MECs and Mechanism Assignments)**.** *Let the causal Markov property (Assumption 2.1), the faithfulness assumptions (Assumptions 2.2-2.5) and sufficiency (Assumption 2.3) hold and assume an ANM (Assumption 3.1) and Sufficient Capacity (Assumption 3.2). Then for any minimizer G, Π of Eq. (3.4),*

$$Pr(G \sim G^*,\ \Pi = \Pi^*) \to 1$$

as the number of contexts m and observations n_c for all c tend to infinity.

Proof Sketch. The proof follows a common score-consistency argument (e.g., Brouillard et al., 2020). We compare the score in Eq. (3.3) of a candidate model (G, Π) to that of the true model (G^*, Π^*) and show that any incorrect graph or incorrect mechanism assignment is penalized either through data fit or model complexity.

First consider the graph. If $G \not\sim G^*$, then G is not Markov equivalent to the true DAG. By the Causal Markov property and Faithfulness (Assumption 2.1, Assumption 2.2), G does not encode the same conditional independence relations as G^*, and hence the true observational distribution cannot be represented exactly by a factorization according to G. Therefore the candidate incurs strictly positive asymptotic fit loss. By contrast, if $G \sim G^*$, then the same joint distribution admits an equivalent factorization under G, so no asymptotic fit penalty arises from the graph alone. Thus every score minimizer must lie in the MEC of G^*.

Second, consider the mechanism assignments. If Π merges two distinct true mechanisms, then one fitted conditional must represent two different true conditionals, resulting in a penalty in data fit under Change Faithfulness (Assumption 2.5). If Π instead splits a true mechanism un-

necessarily, the data fit does not improve while the description length increases. Hence every score minimizer must also satisfy $\Pi = \Pi^*$.

Therefore, every minimizer recovers the true mechanism assignments and a Markov-equivalent graph in the limit. □

In the case of a single observational distribution, the MEC of G consists of all DAGs that encode the same conditional independence relations (Appendix B). Therefore, recovery of the MEC is generally the strongest guarantee one can expect without exploiting additional structure beyond the observational distribution.

In the case of multiple contexts, we now show that we can resolve edge orientations within the MEC, using Independent Change (Assumption 2.4) which we have not yet explicitly used above.

3.2.2 Recovery of Edge Orientations

Our main result is that the presence of mechanism shifts identifies causal edge orientations. We show that if a true causal edge is oriented incorrectly in a graph, then the resulting inferred conditionals in this graph show additional distribution shifts across the contexts. In turn, this forces the model to introduce additional mechanism shifts which are penalized by the score.

Theorem 3.2 (Identifiability of Causal Directions). *Let the assumptions from Chapter 2 (Assumptions 2.1-2.6) hold and assume an ANM (Assumption 3.1) and Sufficient Capacity (Assumption 3.2). Then*

$$Pr(G = G^*,\ \Pi = \Pi^*) \to 1$$

as the number of contexts m and observations n_c for all c tend to infinity.

Proof Sketch. By Theorem 3.1, it remains to distinguish between DAGs in the MEC of G^*. Let $G \sim G^*$ with $G \neq G^*$, and consider a true edge $X_j \to X_k$ that is reversed in G.

Under Independent and Sparse Change (Assumptions 2.4-2.6), and as $m \to \infty$, with probability tending to one, there exist contexts c, c' such that

$$\pi_j^c \neq \pi_j^{c'} \qquad \text{but} \qquad \pi_k^c = \pi_k^{c'} .$$

In the true graph, X_j is a parent of X_k, so conditioning on $Pa_k^{G^*}$ blocks the path $\Pi_j \to X_j \to X_k$. Hence the conditional distribution of X_k depends only on its own mechanism, and therefore

$$P^c(X_k \mid Pa_k^{G^*}) = P^{c'}(X_k \mid Pa_k^{G^*}) \, .$$

Under the graph G with the reversed edge, the true cause X_j is omitted from the parent set of X_k. Then Π_j remains d-connected to X_k given Pa_k^G, so by Markov and Faithfulness,

$$P^c(X_k \mid Pa_k^G) \neq P^{c'}(X_k \mid Pa_k^G) \, .$$

Thus the true mechanism grouping for X_k is no longer sufficient under G, as the candidate model must either merge distinct conditionals or introduce additional mechanism groups, increasing data fit or model complexity, respectively. Hence G has strictly larger score than G^*.

We can repeat the argument for every incorrectly oriented edge. □

Combining Theorems 3.1 and 3.2, we obtain that the proposed score is consistent and recovers the true causal model in the infinite-sample limit. However, the search space over causal models is too large to minimize Eq. (3.4) exhaustively. We therefore propose a causal discovery algorithm in the following section.

3.3 Algorithm

The preceding results require optimization over candidate graphs G, over causal mechanism assignments Π, and repeated fitting of GP regressions, the latter of which is computationally demanding.

In this section, we introduce an algorithm addressing these problems. For simplicity we minimize the score over a set of candidate graphs $G_1, \ldots, G_{max}$ in the approach, and postpone the full search problem over directed acyclic graphs (DAGs) to Chapter 4.

Mechanism Assignments. We first address how to infer the latent variables Π from observations $\mathcal{D}$. Given each X_j under its parent set Pa_j in one of the graphs $G \in \{G_1, \ldots, G_{max}\}$, Π_j takes values in $\{1, \ldots, k_j\}$ for unknown kj and assigns each context c to a mechanism through π_j^c.

Therefore we can interpret Π_j as a partition of the context indices $\{1, \ldots, m\}$ into k_j clusters, and face essentially a clustering problem. Exhaustive search over such clusters scales as $O(d \cdot b_m)$ where b_m denotes

Algorithm 3.1: LINC

input : Dataset $\mathcal{D} = \{\mathcal{D}^1, \ldots, \mathcal{D}^m\}$ over $\mathcal{X}$ in m contexts
set of graph hypotheses $\mathcal{G} = \{G_1, \ldots, G_{max}\}$
output: causal model $\mathcal{M} = (G, \Pi)$

foreach graph $G \in \mathcal{G}$ **do**
  foreach variable X_j with parents Pa_j in G **do**
    foreach context c **do**
      Fit GP $f_j^c \sim \left(\boldsymbol{x}_j^{(c)} \mid \boldsymbol{pa}_j^c\right)$ *// per-context regression*
    foreach pair of contexts c, c' **do**
      Compute score difference $\Delta_{c,c'}$ between shared and separate GP models using f_j^c and $f_j^{c'}$
    Infer Π_j from all $\Delta_{c,c'}$ *// mechanism changes*
    foreach mechanism group k **do**
      Fit GP $f_j^k \sim \left(\boldsymbol{x}_j^{(k)} \mid \boldsymbol{pa}_j^{(k)}\right)$ *// per-mechanism regression*
      Compute score $L\left(\boldsymbol{x}_j^{(k)} \mid \boldsymbol{pa}_j^{(k)}; \mathcal{H}\right)$ in Eq. (3.2)
  Compute score $L(\mathcal{D}; \mathcal{M})$ in Eq. (3.3)
return causal model $\mathcal{M}$ with minimal score $L(\mathcal{D}; \mathcal{M})$

the Bell number in m (Knopfmacher & Mays, 2005), which makes exact search infeasible even for moderate numbers of contexts m.

As is common in clustering problems, we propose a heuristic based on pairwise comparisons. For each pair $c \neq c'$ of contexts, we compare two hypotheses, the first assuming that both contexts share a single GP $f^{c,c'}$; the second assigning distinct GPs f^c and $f^{c'}$ to each context. We denote the corresponding model hypotheses by G, Π for the shared-GP case and by G', Π' for the separate-GP case.

We compare these hypotheses under MDL by computing the compression gain

$$\Delta_{c,c'} = L(\mathcal{D}; G, \Pi) - L(\mathcal{D}; G', \Pi') \ .$$

Using the pairwise compression gains $\Delta_{c,c'}$, we then infer the underlying partition into k_j clusters using a clustering approach, resulting in the values of Π_j. This reduces the combinatorial search problem over mechanism assignments to a clustering task based on MDL score differences.

Score Significance. For later use we also consider a significance threshold for the compression gain $\Delta_{c,c'}$ (Marx & Vreeken, 2019b). The no-hypercompression inequality (Grünwald, 2007) additionally allows to define such a threshold. It states that the probability that an incorrect model achieves a compression gain of k bits over the true model is bounded by 2^{-k}. Consequently, for a significance threshold α, a gain of k bits is treated as insignificant whenever $2^{-k} > \alpha$, since such a gain is not unlikely under the null comparison.

RFF approximation. We next address the computational cost of GP regression, as standard GP inference scales cubically in the number of data points in the worst case. To mitigate this complexity, we employ an approximation based on Random Fourier Features (RFFs) (Rahimi & Recht, 2007). The RFF approach approximates a shift-invariant kernel through its Fourier representation, yielding an explicit finite-dimensional feature map. This leads to a linear regression model that approximates the GP. With n_{RFF} random features, the computational complexity reduces to $O(\min\{n^3, n_{\text{RFF}}^3\})$, for n observations, substantially improving scalability.

LINC. We summarize the algorithm for learning causal models under INdependent Changes (LINC) in Algorithm 3.1. Given multi-context data and a set of candidate graphs, the method proceeds as follows. For each candidate graph G, we fit a GP to each conditional $X_j \mid Pa_j^G$ within each context (lines 3–4). For each pair of contexts, we compute the compression gain under the shared-versus-separate GP hypotheses (lines 5–6). We then infer Π_j from these pairwise scores using clustering (line 7). For each variable X_j and inferred mechanism k, GP regressions are fitted on pooled observations across all contexts c satisfying $\pi_j^c = k$, as defined in Section 3.1, which allows to compute the overall score of this model (lines 8–10). We return the model $\mathcal{M}$ with best (smallest) description length $L(\mathcal{D}; \mathcal{M})$ (line 12).

3.4 Related Work

Below, we outline related research in causal discovery from multi-context data. We include a detailed discussion of specific algorithms for causal Directed Acyclic Graph (DAG) discovery to Chapter 4 which focuses on the DAG search problem in detail.

AMC-Based Causal Discovery. A related line of work studies causal discovery through the algorithmic Markov condition (AMC), which states that a causal model admits a description that decomposes into independent descriptions of its causal mechanisms (Janzing & Schölkopf, 2010; Schölkopf et al., 2021). Since Kolmogorov complexity is not computable, practical approaches approximate this principle using score-based criteria such as MDL. In the i.i.d. setting, this perspective underlies methods based on additive noise models, functional causal models, and regression-based compression scores (Hoyer et al., 2009; Peters et al., 2014; Marx & Vreeken, 2019b). Our approach extends this line of work to multi-context data by combining MDL-based scoring with the discovery of mechanism assignments that change across contexts.

Multi-Context Setting. Causal invariance plays a central role in the literature on multi-context data. It exploits an asymmetry between causal and anti-causal directions across different contexts or environments (Peters et al., 2016; Heinze-Deml et al., 2018; Arjovsky et al., 2019).

Subsequent work generalizes this setting by explicitly modeling changes in causal mechanisms (Huang et al., 2020; Mooij et al., 2020). These approaches treat the context as an observed discrete variable C that enters the augmented causal graph. This formulation enables constraint-based causal discovery jointly over X and C.

Interventional Setting. A related line of work studies interventional settings rather than mechanism changes. In this setting, interventions directly modify the value of a variable Y (e.g. Yang et al., 2018), such that each context c results from intervening on a subset of variables $\mathcal{I}_c \subseteq X$, where intervention targets may be known (Hauser & Bühlmann, 2012; Yang et al., 2018) or unknown (Squires et al., 2020). Examples include perfect interventions $Y := do(Y)$ in the sense of Pearl's do-calculus (Pearl, 2009), shift interventions (Rothenhäusler et al., 2021) or soft interventions (Eberhardt & Scheines, 2007). Interventions induce distribution shifts that similarly strengthen identifiability, here from the Markov Equivalence Class (MEC) to the smaller Interventional Markov Equivalence Class (IMEC) (Yang et al., 2018). This holds true even in case the intervention targets are unknown (Squires et al., 2020).

Identifiability from Changes. Our main theoretical results (Theorems 3.1-3.2) establish that independent and sparse changes enable identification of causal directions. This principle does not depend on the specific use of GPs or MDL. In principle, any functional model and scoring crite-

rion may be used, provided that Assumption 3.2 holds. For example, one could adopt alternative score-based approaches such as Brouillard et al. (2020).

This provides a score-based, information-theoretic complement to methods based on the Sparse Mechanism Shift (SMS) assumption, such as Perry et al. (2022). They show that counting mechanism changes suffices for causal structure identifiability, based on which they propose the Mechanism Shift Score (MSS). Ghassami et al. (2018) propose a related score (MC) for linear systems.

Our score implicitly performs such counting through regularization while simultaneously incorporating regression fit via the GP likelihood. MSS and MC do not account for model fit in this way. We compare our approach with MSS and MC in the following.

3.5 Experiments

We conclude with an evaluation of our score. The questions that guide the evaluation are the following.

(*i*) *Causal Direction Discovery.* To strengthen the result in Theorem 3.2, can we discover causal directions in multi-context data in practice?

(*ii*) *Ablation Studies.* To justify the instantiation choice using GPs and MDL, is the approach robust across different settings and under model misspecification, and is there a benefit of using the score-based approach over a mechanism shift counting score?

3.5.1 Experimental Setup

As we postpone causal DAG search in Chapter 4, we begin by studying whether the MDL score allows deciding between Markov-equivalent structures, where we provide the correct DAG skeleton and use L to orient causal edges. We compare the score against previous scoring criteria with the same goal, specifically the MC score (Ghassami et al., 2018), which utilizes the sparse mechanism shift property in linear systems, and the MSS score (Perry et al., 2022) for non-parametric settings, instantiated with the Kernelized Conditional Independence (KCI) test. We also include one baseline causal discovery algorithm, here the constraint-based CD-NOD (Huang et al., 2020).

We sample a ground truth G^* from an Erdös Rènyi DAG model, minimize our score over all graphs $G \sim G^*$, and in the best G evaluate edge directions for all node pairs, reporting F1 scores averaged over each pair X_j, X_k in $\mathcal{X}$ (F1-causal). For the synthetic experiments, we use the same data generation choices as in Perry et al. (2022); Huang et al. (2020), and generate data in context c via

$$X_j = \sum_{k \in Pa_j} \omega_{jk}^c f_{jk}(X_k) + \sigma_j^c N_j^c \ . \tag{3.5}$$

with weights $\omega_{jk}^c \sim \mathcal{U}(0.5, 2.5)$, where the noise N_j^c is either uniformly or normally distributed, and f chosen from $\{x^2, x^3, \tanh, \text{sinc}\}$ with equal probability. In each c, we choose a random subset $S \subseteq \{1, \ldots, , d\}$ of variable indices that change, where $n_S := |S|$ denotes their number. The change is implemented by resampling the corresponding mechanism parameters in Eq. (3.5). Additionally, we test noise scaling through $\sigma \sim \mathcal{U}(2, 5)$ and hard interventions via $w_j = 0$.

Besides the synthetic data, we also consider more realistic semi-synthetic data from the SERGIO simulator (Dibaeinia & Sinha, 2020). SERGIO simulates data from a generating process that resembles the real-world dynamics of gene transcription and regulation. Under a gene regulatory network G over d genes, it models the causal functions f using stochastic differential equations. As proposed by Hägele et al. (2023), we use the randomly sampled DAG G as the reference network and generate $m = d$ interventional datasets as our contexts. In each context c, the generator performs a knockout intervention on a different gene X_j.

3.5.2 Causal Direction Discovery

Figure 3.2 shows our main results on F1-causal, where the gray baseline illustrates random orientation of edge directions. On synthetic data, all shift-counting scores (MSS, MC) as well as our algorithm LINC (GP, RFF) show improved recovery of edge directions with more contexts, in line with our identifiability results; the constraint-based baseline CD-NOD does not benefit comparably from additional contexts. The linear score MC performs well despite non-linearities in this dataset, but requires a larger number of observed variables. While the confidence intervals for our method overlap with those of MSS and MC, it performs well already in cases with few variables or sparse changes. An important observation is

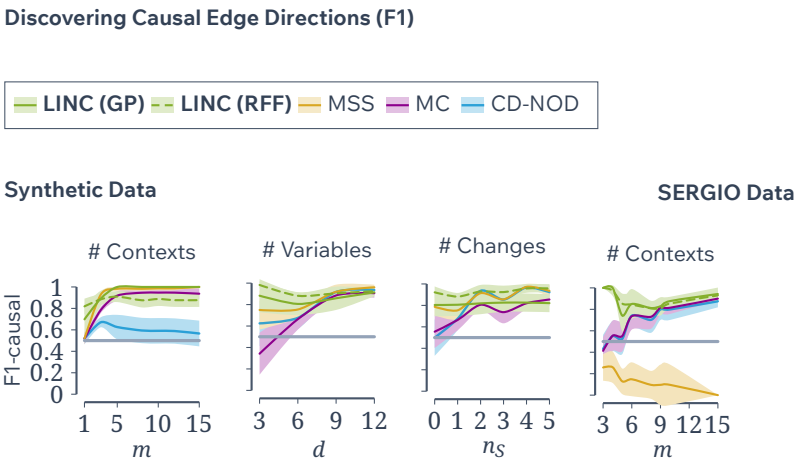


Figure 3.2 To evaluate the methods on discovering causal directions, we sample Erdős-Rényi DAGs G^*, use the methods to discover the best $G \sim G^*$, and report F1-scores over each edge direction in G (F1-causal). The first three plots show synthetic data, with $m = 5$, $d = 6$, and $s_X = 2$, varying one of these parameters at a time. In the rightmost plot, we simulate RNA-seq Data with SERGIO (Dibaeinia & Sinha, 2020) from a random network G over $m - 1$ genes, with a knockout intervention on one gene per context.

that LINC already performs well for $m = 1$ and for $n_S = 0$, corresponding to the i.i.d. case. This suggests that, beyond the IC-based asymmetry studied theoretically, the MDL score also captures asymmetries induced by the underlying regression model. We return to this point in an ablation study (Section 3.5.4).

Figure 3.2, right shows the results for the SERGIO semi-synthetic data. MSS performs poorly here, perhaps due to relying solely on mechanism shifts, while our score takes regression fits into account; meanwhile, CD-NOD shows better performance with an increasing number of conditions. Note that as we consider DAGs over at least $d = 2$ variables and in this experiment, the number of contexts $m = d + 1$, the x-axis range starts at 3. Only our score already performs well in $m = 3$ conditions, i.e., the bivariate case, where the constraint-based methods do not infer the causal direction better than the random baseline (gray).

Next, we perform additional experiments that vary the data generation setup to study violations of certain assumptions.

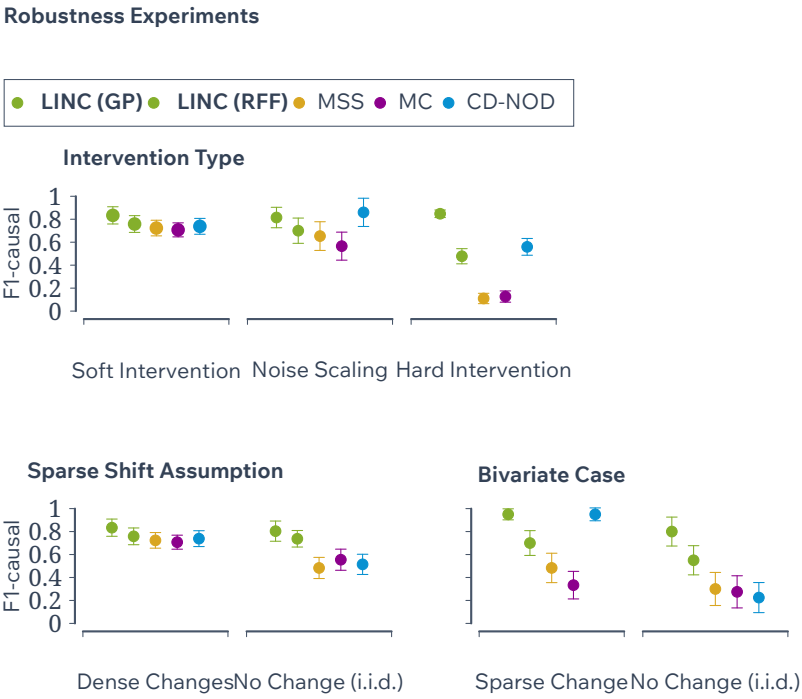


Figure 3.3 Showing the recovery of causal edge directions on synthetic data (F1-causal), we consider robustness under different intervention types (top), in settings with many, respectively, few changes (bottom left), and the bivariate case (right).

3.5.3 Robustness to Model Misspecification

Figure 3.3 shows experiments under other intervention types. We show the default setting with causal mechanism shifts, i.e., soft interventions (top left), noise scaling (middle), and hard interventions (right). CD-NOD performs well under noise scaling, where MSS and MC degrade slightly; only the GP variant appears robust to hard interventions.

Second, we consider the effects of Sparse Changes (Assumption 2.6) also used by MSS and MC. In Figure 3.3, bottom left, we simulate data where $n_S = d$ (left), respectively, $n_S = 0$ (right). The latter is essentially

an i.i.d. setting, in which shift counting (MSS, MC) performs no better than random guessing of causal directions. At the same time, our variants continue to perform well, consistent with the score also exploiting asymmetries of the underlying regression model beyond mechanism changes (Section 3.5.4).

Third, we also show the special case of a single variable pair in Figure 3.3, bottom right, where most methods struggle to identify the causal direction. CD-NOD achieves this when changes exist (left subplot) but no longer in an i.i.d. setting, where only the GP variant achieves reasonable F1 scores (right subplot).

The main conclusion matches the result in Perry et al. (2022), where shift counting identifies all causal directions as m grows (Figure 3.2). In contrast, our score appears to be robust to model misspecification, and notably performs well in bivariate and i.i.d. settings (Figure 3.3).

3.5.4 Score Behavior

The results in Figure 3.2 indicate that in practice, the score can also work well in the absence of causal mechanism changes ($s_X = 0$). This does not follow from our IC-based theory directly. Rather, the GP-based MDL score may still exploit asymmetries already present in the underlying additive-noise regression model, as in related work on MDL and additive-noise methods (Marx & Vreeken, 2019c; Mian et al., 2021). The following examples illustrate how this model-class asymmetry is present already in the i.i.d. case and becomes more robust when independent mechanism changes are available.

Example 3.1 Consider a causal chain $X \to Y \to Z$, where we generate data from

$$X = N_X, \qquad Y = \begin{cases} f_Y(X) + N_Y & \text{if } \Pi_Y = 1 \\ \tilde{f}_Y(X) + N_Y & \text{if } \Pi_Y = 2 \end{cases}, \qquad Z = f_Z(Y) + (N_Z),$$

with $\Pi_Y \in \{1, 2\}$ and $c = 1, \ldots, 5$, with each f drawn from a GP. We compare models where Y has cause X, respectively cause Z, and which can contain up to five different (k_Y many) GP regressions applied in the five contexts. Figure 3.4, left shows the MDL scores L of the best model at each k_Y. Models in the causal direction (green) consistently score better than in the anti-causal direction (gray) and, additionally, the score is minimized

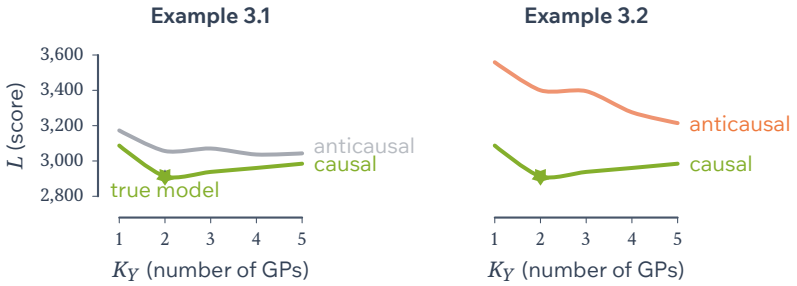


Figure 3.4 For a causal chain, we compare the MDL scores L of models in the causal (green) resp. anti-causal direction (gray, orange) with K_Y different GP regressions in five contexts. *Left:* The minimum (star) corresponds to the ground truth. *Right:* Description lengths in the causal direction remain independent of, and invariant under, changes in mechanism of the covariates.

when using two GPs (star). △

The example illustrates the regularization behavior of the MDL score, recovering the correct edge orientation and the true number of causal changes. Here, Y is generated from X with two different causal functions, while neither X nor Z changes across contexts. Thus, the observed asymmetry arises already from the underlying additive-noise regression model. In the next example, we show how independent mechanism changes further strengthen this asymmetry.

Example 3.2 Now we assume five distinct mechanisms for both X and Z,

$$X = \begin{cases} g_X^1(N_X) & \text{if } \Pi_X = 1 \\ \dots & \\ g_X^5(N_X) & \text{if } \Pi_X = 5 \ , \end{cases} \qquad Z = \begin{cases} g_Z^1(Y) + N_Z & \text{if } \Pi_Z = 1 \\ \dots & \\ g_Z^5(Y) + N_Z & \text{if } \Pi_Z = 5 \ , \end{cases}$$

while Y alternates between two functions (cf. Example 3.1). Figure 3.4, right shows the effect of this change. The description lengths in the causal direction remain unaffected (green), whereas description lengths of all anti-causal models increase noticeably (gray to orange). △

In the anti-causal direction, mechanisms for Y from Z must fine-tune

themselves to each change in Z. As a result, the best anti-causal model uses a distinct function in each context (orange, $x = 5$), whereas in the causal direction two functions still suffice (green, $x = 2$).

Taken together, Example 3.1 and Example 3.2 suggest that the score uses two sources of asymmetry, first a model-class asymmetry already present in additive-noise regression (Marx & Vreeken, 2019c; Mian et al., 2021), and second the asymmetry induced by independent mechanism changes. Our theory addresses the latter explicitly; understanding their combination more formally is an interesting direction for future work.

3.6 Conclusion

In summary, the goal of this chapter is causal learning from multiple contexts, specifically discovering causal edge orientations and independent causal mechanisms.

Given multi-context datasets $\mathcal{D} = \{\mathcal{D}^1, \ldots, \mathcal{D}^m\}$, we discover a causal model including independent latent variables Π representing causal mechanism shifts. In practice, we model each variable X_j as generated through a GP f_j^k with additive noise and minimize a decomposable scoring criterion L measuring the description lengths of these mechanisms.

One limitation of this approach is the graphical representation we assume (Chapter 2). In more complex settings, differences in the graph structure (Wang et al., 2018) or mixtures of DAGs (Saeed et al., 2020) may be a more realistic abstraction of real-world systems. Further, we model distribution shifts as changes in causal mechanisms, not addressing changes in noise distribution, perfect interventions, or latent variables here.

When the modeling assumptions hold, we can identify causal directions in the graph (Theorem 3.2). Existing results consider the identification of edge orientations using the independence of noises (IN) (Peters et al., 2014; Marx & Vreeken, 2019c). IN and IC need different sets of assumptions, where IN restricts the assumed model class, IC assumes that causal mechanism changes are independent (Assumption 2.4). As our scoring criterion similarly uses additive-noise models, it may combine aspects of both IN and IC. Empirically, we observe that it can perform well even in the single-context case or in the absence of mechanism shifts. A theoretical understanding of this combination is an interesting direction for future work.

Limitations related to the algorithm include the large search space over context groups that share the same causal mechanism, which we base on pairwise comparisons between contexts. Future work could explore providing approximation guarantees for this approach or consider alternative implementation choices. Regarding the implementation choices, the use of IC does not rely on specific model restrictions, and future research could test related scoring criteria or other regressors. The experimental setup so far is simplified in that we only consider the discovery of causal directions, assuming the correct MEC as a starting point. Therefore, we address causal DAG search in the following chapter.

Software. To make it easy for the reader to apply LINC, we make a Python implementation publicly available online (`github.com/srhmm/thesis`). To get started, we additionally provide an example usage in the form of a Jupyter notebook (`demo/ch4_causal.ipynb`).

4 | Causal DAG Discovery from Multiple Contexts

Case Study Pathway Problem

In genomics, constructing Gene Regulatory Networks (GRNs) can improve the understanding of the regulatory dynamics governing gene expression. Modern single-cell transcriptomics (Dibaeinia & Sinha, 2020) provide high-resolution snapshots of these profiles. Besides steady-state measurements, experimenters can perform targeted interventions by fixing the expression of specific genes to observe the response (Figure 4.1). Given steady-state observations (left) and measurements from such interventions (middle/right), a geneticist aims to reconstruct the underlying GRN, but without assuming prior knowledge of the specific targets or the downstream effects of those interventions (blue).

In this chapter, we address the problem of discovering fully directed causal networks, in the form of Directed Acyclic Graphs (DAGs), from multi-context data.

The preceding case study shows a common scenario in biology where contexts correspond to distinct experimental conditions with interventions. We focus on soft interventions, as we can model these through changes in causal mechanisms (Eberhardt & Scheines, 2007). Sufficient background knowledge on the effects of such interventions, as required

The chapter is based on work published as Xu et al. (2025).

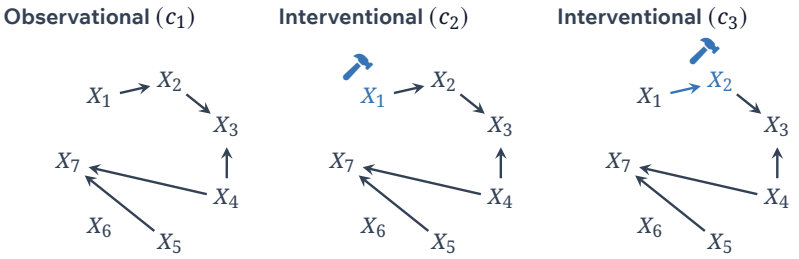


Figure 4.1 A gene regulatory network (GRN) in the form of a DAG describing the causal relationships over a set of genes. Targeted interventions affect one variable at a time (blue), resulting in changes to the causal mechanisms (blue edges) that are independent of the remaining mechanisms (gray edges).

for example in active intervention targeting (Hauser & Bühlmann, 2014), is not available in all applications; in gene editing, for example, intervention effects can arise at sites other than the intended target (Zhang et al., 2015), making it difficult to specify the exact set of variables affected by each intervention. Consequently, we treat mechanism changes as unknown, consistent with the methodology in the previous chapters.

To address such settings, we rely on the idea of Independent Change as before, where we expect interventions to change the causal mechanisms of their targets, while the remaining causal conditionals in the DAG remain invariant across contexts. To apply this intuition, we incorporate the local scoring criterion based on the Minimum Description Length (MDL) principle for Gaussian Processes (GPs) introduced in Chapter 3 into a score-based causal discovery algorithm, TOPIC.

In Section 4.1 we first introduce the algorithm TOPIC itself, which can also be used in a single-context dataset setting and with other local scoring criteria (Xu et al., 2025). Then we justify TOPIC in the multi-context setting in Section 4.1, along with an empirical evaluation in Section 4.5.

4.1 DAG Search in Topological Order

Given data $\mathcal{D}$ over d variables $\mathcal{X} = \{X_1, \ldots, X_d\}$ under a causal model $\mathcal{M}$, we assume a local scoring criterion that decomposes over variables as

$$L(\mathcal{D}; \mathcal{M}) = \sum_{j=1}^{d} L\left(X_j \mid Pa_j^G\right) .$$

Whenever we write a local score of the form $L(X \mid Pa_X^G)$, we refer to the multi-context GP-based MDL score from Chapter 3, including re-estimation of the corresponding mechanism assignments and pooled observations under the candidate parent set,

$$L\left(X_j \mid Pa_j^G\right) = \sum_{k=1}^{k_j} L\left(\boldsymbol{x}_j^{(k)} \mid \boldsymbol{pa}_j^{(k)}; \mathcal{H}\right) , \tag{4.1}$$

where each term denotes the score contribution of mechanism k for variable X_j, determined by the assignments Π_j, under the GP functional model class $\mathcal{H}$ with additive noise.

The goal of this chapter is to discover the graph G minimizing the above score. While we present the algorithm for this multi-context instantiation, similar scoring criteria can be used within the same approach in the single-context setting (Xu et al., 2025).

4.1.1 Topological Ordering

The key idea in our approach is to include edges in the DAG whenever they result in local improvements of the score in Eq. (4.1). In particular, with the same reasoning as before (cf. Theorem 3.2) including a missing cause $X_k \in Pa_j^{G^*}$ of a given X_j in the graph results in a score improvement.

Example 4.1 Consider the true DAG G^* in Figure 4.1. With $G = (\mathcal{X}, \varnothing)$ initially, say we consider the node X_2 with missing causal parent X_1. The intervention on X_1 results in a distribution change that propagates to X_2, $P^1(X_2) \neq P^2(X_2)$. In contrast, for the causal conditional, $P^1(X_2 \mid X_1) = P^2(X_2 \mid X_1)$ as the causal mechanism of X_2 remains invariant in contexts 1 and 2. Hence, adding the true cause X_1 of X_2 to G strictly reduces the number of mechanism distinctions needed in the model. △

In general, starting with an empty graph, we add edges in order of the score improvement they provide under the current model, proceeding along the topological ordering of the causal graph.

We specify a topological order T over the variables $\mathcal{X}$ by an injective function $T : \mathcal{X} \to \{1, \ldots, d\}$, where we call T a causal order when it agrees with G^*, i.e., for all edges $(X_j, X_k) \in \mathcal{E}^*$, we have $T(X_j) < T(X_k)$. We also define partial topological orders T^k where only k nodes are ordered, where T^k is a function $T^k : \mathcal{X} \to \{1, \ldots, k\} \cup \{-1\}$, where $T^k(X_j) = -1$ if X_j is not yet assigned a position, and otherwise $T^k(X_j) \neq T^k(X_k)$ for all $k \neq j$.

Of interest are also nodes that can be placed next under a partially constructed graph G and partial order T^{k-1}, called admissible nodes. A node X is admissible if all its true parents are already included in the current graph, i.e., $Pa_X^{G^*} \subseteq Pa_X^G$. Initially, when T is empty only source nodes of G^* are admissible.

First, suppose we have access to an oracle O which returns an admissible next node given a partial graph G and order T^{k-1}. We first describe how to infer G^* given O, and then show how to define O itself.

Let G be the empty graph with $\mathcal{V} = \mathcal{X}$ and $\mathcal{E} = \emptyset$, and T^0 an empty topological order with $T^0(X_j) = -1$ for all X_j.

For iterations $k = 1, \ldots, d$, we perform in order

1. Obtain the next node X as $X = O\left(G, T^{k-1}\right)$ and set $T^k(X) = k$.

2. Add all outgoing edges (X, Y) to $\mathcal{E}$ for which $T^k(Y) > k$ or $T^k(Y) = -1$ and

$$L\left(Y \mid Pa_Y^G \cup \{X\}\right) < L\left(Y \mid Pa_Y^G\right) . \tag{4.2}$$

3. Remove all incoming edges (Z, X) from $\mathcal{E}$ for which

$$L\left(X \mid Pa_X^G \setminus \{Z\}\right) = L\left(X \mid Pa_X^G\right) . \tag{4.3}$$

The remaining step is to define the oracle itself.

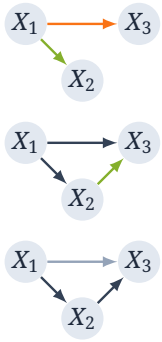


Figure 4.2 Example of $k = 3$ iterations of TOPIC, with edge additions (green if causal, red otherwise) and removal (light gray).

4.1.2 Source Node Identification

To define O, we use the distinguishing property of admissible next nodes, namely that all their causal parents are already included in G. Hence no unresolved node is a missing true parent of an admissible node.

Therefore, for each variable pair X, Y we consider the gain $\Delta^G_{X\to Y}$ in score of including an edge (X, Y) in G. This gain is defined as the score difference between two causal models $\mathcal{M}$ and $\mathcal{M}_{X\to Y}$, the latter differing from $\mathcal{M}$ only by the additional edge. That is, $\mathcal{M}$ consists of the graph G and Π which represents changes in conditionals under G; $\mathcal{M}_{X\to Y}$ consists of the graph $G_{X\to Y}$ with $\mathcal{E}_{X\to Y} = \mathcal{E} \cup \{(X, Y)\}$ and $\Pi_{X\to Y}$ representing a potentially different set of changes in conditionals under the new graph. Then the directional gain is

$$\Delta^G_{X\to Y} = L\left(Y \mid Pa^G_Y\right) - L\left(Y \mid Pa^G_Y \cup \{X\}\right) . \tag{4.4}$$

where we suppress the dependency on the graph G if clear from context. A positive gain indicates that X provides predictive information about Y not yet captured by G.

To quantify which edge orientation the score supports more strongly, we compare the gains δ in both directions via

$$\Delta_{X;Y} = \Delta_{X\to Y} - \Delta_{Y\to X} ,$$

where intuitively $\Delta_{X;Y} > 0$ indicates that, under the current graph G, adding X as a parent of Y is more beneficial than the reverse direction. This suggests that asymptotically $\Delta_{X;Y} \geq 0$ for an admissible node X relative to any other unresolved node Y, whereas a non-admissible node still has a missing parent for which $\Delta_{X;Y} < 0$.

We therefore define the oracle using a maximin criterion. For each candidate X, we identify its worst-case relationship, as the node Y that most strongly challenges the hypothesis that X is a source. We then select the node X that best maintains this source-node property across all Y,

$$O\left(G, T^{k-1}\right) = \underset{X \text{ unresolved}}{\arg\max} \left(\min_{Y \text{ unresolved}, Y\neq X} \Delta_{X;Y} \right) . \tag{4.5}$$

The oracle selects the unresolved node X that is most robustly supported as a source, relative to all other unresolved nodes Y.

We conclude by showing that the proposed strategy is consistent.

4.2 Consistency

We first state consistency of the topological search under access to a correct oracle.

Theorem 4.1 (Consistency of Topological Search)**.** *Let the assumptions of Theorem 3.2 hold. Assume the oracle O is correct (cf. Theorem 4.2). That is, at each iteration it returns an unresolved node whose true parents are already present in the current graph. Then TOPIC recovers the true DAG G^* as the number of contexts m and the sample sizes n_c tend to infinity.*

Proof Sketch. First, we show by induction over the iterations of TOPIC that after iteration k, T^k forms a causal topological order of G^*, and for all resolved nodes X with $T^k(X) \neq -1$, all outgoing true edges $(X, Y) \in \mathcal{E}^*$ are contained in $\mathcal{E}$.

To this end, by correctness of the oracle, the node $X_k = O\left(G, T^{k-1}\right)$ is an admissible next node, i.e., $Pa_X^{G^*} \subseteq Pa_X^G$, and assigning $T^k(X_k) = k$ preserves a valid topological order. Now consider any unresolved node Y, i.e., $T^{k-1}(Y) = -1$. If $(X_k, Y) \in \mathcal{E}^*$ and $(X_k, Y) \notin \mathcal{E}$, then $X_k \in Pa_Y^{G^*} \setminus Pa_Y^G$. By the same argument from Theorem 3.2 using Independent Change (Assumption 2.4), adding a missing true parent X_k strictly reduces the score of Y in Eq. (4.1), therefore TOPIC adds the edge (X_k, Y). Therefore, after resolving X_k, all outgoing true edges from X_k are included in $\mathcal{E}$.

By induction, the algorithm constructs a valid topological order and recovers all true edges. To show that it also prunes all spurious edges, consider any node Y. Once all true parents of Y are included in Pa_Y^G, any additional parent $Z \in Pa_Y^G \setminus Pa_Y^{G^*}$ is redundant. By the same score-consistency argument as in Theorem 3.1, removing such a spurious parent does not worsen the asymptotic likelihood term, while it improves the score through the model complexity penalty. In contrast, removing a true parent induces a strictly positive likelihood penalty. Hence the pruning step removes all and only spurious edges.

Overall, at the end of all iterations we obtain $G = G^*$. □

Finally we show that the oracle indeed selects an admissible next node. The identifiability result of Theorem 3.2 implies the local score asymmetry used below, where adding a missing true parent strictly decreases the score, whereas redundant parents do not.

Theorem 4.2 (Oracle Correctness). *Let the assumptions of Theorem 3.2 hold. Let T^{k-1} be a correct partial topological order and let G contain all outgoing true edges from resolved nodes. Then the oracle*

$$O\left(G, T^{k-1}\right) = \underset{X \text{ unresolved}}{arg\,max} \left(\underset{Y \text{ unresolved}, Y \neq X}{min} \Delta_{X \to Y} - \Delta_{Y \to X} \right)$$

returns an admissible node X such that $Pa_X^{G^*} \subseteq Pa_X^G$.

Proof Sketch. For each unresolved node X, define its worst-case directional asymmetry by

$$\Delta_{\min}(X) := \min_{Y \text{ unresolved}, Y \neq X} \left(\Delta_{X \to Y} - \Delta_{Y \to X}\right) .$$

We show that admissible nodes satisfy $\Delta_{\min}(X) \geq 0$, whereas non-admissible nodes satisfy $\Delta_{\min}(X) < 0$.

First, let X be admissible, so $Pa_X^{G^*} \subseteq Pa_X^G$. Then X has no missing true parent among the unresolved nodes. Hence for any unresolved $Y \neq X$, adding the edge $Y \to X$ cannot insert a missing true cause into the conditional of X, so by the same score-consistency argument as in Theorem 3.1, we have $\Delta_{Y \to X} \leq 0$. If Y is a descendant of X, then by the same mechanism-shift argument as in Theorem 3.2, conditioning on X reduces unresolved distribution shifts in the conditional of Y, so $\Delta_{X \to Y} \geq 0$. Therefore $\Delta_{X \to Y} - \Delta_{Y \to X} \geq 0$ for all unresolved $Y \neq X$, and thus $\Delta_{\min}(X) \geq 0$.

Now let X be non-admissible. Then there exists a missing true parent $Z \in Pa_X^{G^*} \setminus Pa_X^G$, which is unresolved because T^{k-1} is correct. By the same argument as in Theorem 3.2, adding the true edge $Z \to X$ strictly improves the local score, so $\Delta_{Z \to X} > 0$, whereas the reverse edge $X \to Z$ does not resolve a missing cause and therefore yields a strictly smaller gain. Hence $\Delta_{X \to Z} - \Delta_{Z \to X} < 0$. Since Z is among the unresolved nodes considered in the minimum, it follows that $\Delta_{\min}(X) < 0$.

Thus admissible nodes are characterized by nonnegative worst-case asymmetry, while non-admissible nodes have negative worst-case asymmetry. Since at least one admissible unresolved node always exists, the maximin oracle

$$O\left(G, T^{k-1}\right) = \underset{X \text{ unresolved}}{\arg\max}\ \Delta_{\min}(X)$$

selects an admissible node. Therefore $Pa_X^{G^*} \subseteq Pa_X^G$ holds. □

Algorithm 4.1: TOPIC

input : Dataset $\mathcal{D} = \{\mathcal{D}^1, \ldots, \mathcal{D}^m\}$ over $\mathcal{X}$ in m contexts
output: causal model $\mathcal{M} = (G, \Pi)$

1 **foreach** iteration $k = 1, \ldots, d$ **do**
2 Obtain an admissible unresolved X (Eq. (4.5)) *//next "source"*
3 Set $T^k(X) = k$
4 **foreach** Y with $T^k(Y) > k$ or $T^k(Y) = -1$ **do**
5 **if** $L\left(Y \mid Pa_Y^G \cup \{X\}\right) < L\left(Y \mid Pa_Y^G\right)$ **then**
6 Add (X, Y) to $\mathcal{E}$ *//edge additions*
7 **foreach** Z with $(Z, X) \in \mathcal{E}$ **do**
8 **if** $\left|L\left(X \mid Pa_X^G \setminus \{Z\}\right) - L\left(X \mid Pa_X^G\right)\right| < \epsilon$ **then**
9 Remove (Z, X) from $\mathcal{E}$ *//edge pruning*
10 **return** causal model $\mathcal{M}$ with minimal score $L(\mathcal{D}; \mathcal{M})$

The consistency arguments in this chapter focus on the multi-context setting, where causal mechanism shifts inform edge orientations. The algorithm TOPIC can also be instantiated in the single-context setting with an appropriate local scoring criterion. In that case, consistency can be established using a similar proof structure, relying on the assumption of additivity of the effects of causal parents in an additive-noise model, as detailed in Xu et al. (2025).

4.3 Algorithm

We summarize the resulting algorithm as Algorithm 4.1. Starting with an empty graph G and an empty topological order T, it iterates d times.

In each iteration, the oracle O identifies the next admissible node X using the maximin directional gain criterion (line 2). For the newly resolved node X, the algorithm evaluates all possible outgoing edges to unresolved nodes and adds them in case they improve the score (lines 3–5). Conversely, it re-evaluates all existing incoming edges for pruning (lines 6–8). Once all nodes are ordered, TOPIC returns the model.

Whenever we compute the score L for a relationship $X \mid Pa_X^G$, this

requires re-estimation of the variables Π reflecting changes in the conditional $P(X \mid Pa_X^{\mathcal{G}})$ for which we apply the same techniques as in the previous chapter (Section 3.3).

In practice, we replace the exact equality in the pruning step in Eq. (4.3) with a thresholded variant in Algorithm 4.1, where ϵ is a score significance threshold derived from the no-hypercompression inequality (cf. Section 3.3). This modification is motivated by the fact that, while in the theoretical analysis score differences for redundant edges tend to zero, finite-sample data can result in small non-zero score differences, such that we only retain edges whose improvement is above the threshold ϵ.

4.4 Related Work

Causal DAG discovery approaches are usually classified as either constraint-based or score-based, where the former rely on conditional independence constraints in the data, and the latter identify the causal model as the one that minimizes a local scoring criterion, often based on edge sparsity. Hybrid approaches combine conditional independence tests with a scoring criterion. We can also make a second classification into classical methods that address the i.i.d. case, learning a DAG from a single observational dataset, and extensions of these methods to address non-i.i.d. settings, here, with observations from multiple contexts with unknown interventions or mechanism changes. We refer to, e.g., Squires & Uhler (2022); Zanga et al. (2022) for surveys, and give a non-exhaustive overview.

Constraint-based Methods. Classical constraint-based methods include the well-known PC algorithm (Spirtes et al., 2000), FCI (Spirtes et al., 1999), which relaxes the causal sufficiency assumption, and the GSP family Solus et al. (2021) as a hybrid approach. Constraint-based approaches apply domain-specific conditional independence tests to address continuous (Zhang et al., 2011) or discrete (Marx & Vreeken, 2019a) settings, and typically return a Completed Partially Directed Acyclic Graph (CPDAG) representing the class of Markov equivalent graphs of the causal DAG.

Score-based Methods. Score-based methods such as GES (Chickering, 2002) search greedily over the space of MECs to minimize sparsity-based scoring criteria, such as the BIC score or kernelized scores (Huang et al., 2018). To recover edge directions beyond Markov equivalence, methods such as LiNGAM (Shimizu et al., 2006) and RESIT (Hoyer et al.,

2009) restrict the class of functional models to additive noise models and test for non-gaussianity, respectively, independence, of residual distributions under these models. Some approaches use MDL formulations, such as the GLOBE algorithm (Mian et al., 2021). Recently, methods based on continuous, neural network-based optimization have gained popularity. NOTEARS (Zheng et al., 2018) formulates causal discovery as a continuous-optimization task to learn DAGs, while Yu et al. (2019) propose to use graph neural networks to learn the causal structure.

Ordering-based Algorithms. Several algorithms first recover a topological order and then obtain the fully oriented DAG. CAM (Bühlmann et al., 2014), SCORE (Rolland et al., 2022) and later extensions such as NOGAM (Montagna et al., 2023) apply topological order estimators for an additive SCM, then add respectively prune edges using Lasso regression. Reisach et al. (2023) point out var-sortability issues with synthetic datasets, and propose to use variance/R^2 respectively to determine the causal order.

Multi-Context Data. Extensions of constraint-based approaches to the multi-context setting include, as the most well-known examples, the Joint Causal Inference (JCI) framework (Mooij et al., 2020) which uses independence testing in an augmented graphical model, and Causal Discovery from non-stationary/heterogeneous data (CD-NOD) (Huang et al., 2020) which extends the PC algorithm. Meanwhile, unknown-target interventional GSP (UT-IGSP) (Squires et al., 2020) extends the previous IGSP and GSP algorithms (Solus et al., 2021) to a setting with unknown intervention targets, a hybrid score- and constraint-based approach. Although able to provide tighter guarantees in the presence of interventions, the methods only identify equivalence classes; in the case of UT-IGSP, this corresponds to the interventional MEC. The scores leveraging the sparse shift principle discussed previously discover additional causal directions (Perry et al., 2022), but these works do not address scalable DAG search.

4.5 Experiments

We conclude with an evaluation of causal discovery, with a similar setup as in the previous chapter, addressing

(*i*) *Synthetic Data*, sampled from a multi-context setting, and

(*ii*) *Real-World Data*, here on causal cell signaling pathways (Sachs et al., 2005).

4.5.1 Experimental Setup

We briefly recall our data generation from Chapter 3, where we sample Erdös-Rényi DAGs and sample data from $X_j = \sum_{k \in Pa_j} \omega_{jk}^{(c)} f_{jk}(X_k) + \sigma_j^{(c)} N_j^{(c)}$ with weights $\omega_{ij}^{(c)} \sim \mathcal{U}(0.5, 2.5)$, either uniform or Gaussian noise, and functions f sampled from $\{x^2, x^3, \tanh, \text{sinc}\}$. Compared to before, we do not provide the correct skeleton, instead discovering the full DAG.

Among causal discovery approaches for a non-linear, multi-context setting with unknown causal mechanism shifts are primarily CD-NOD Huang et al. (2020); JCI (Mooij et al., 2020) which we here instantiate with FCI (Spirtes et al., 2000), referred to as JCI-FCI; and UT-IGSP (Squires et al., 2020). We use the kernel CI test (Zhang et al., 2011). To evaluate the discovered DAGs against the ground truth, we consider the recently proposed separation distance (SD) of causal graphs, which can be used to evaluate DAGs or CPDAGs (Wahl & Runge, 2025). To contrast the ability of TOPIC in directional edge recovery compared to the constraint-based methods, we also show true positive rates (TPR), respectively, FPR, over directed edges in the DAG.

We compare the induced DAG G against the discovered DAG or partially directed acyclic graph (PAG) G'. To give intuitive insight into correct vs. incorrect edge orientations, we show F1 scores over directed edge counts $\mathcal{E}$ in G' (called F1-dir), where we note that in the case of CPDAGs, we only count edges that are directed with certainty. As this is a simplistic score mainly included for illustrative purposes, we also consider classical distance metrics. A common distance measure for two DAGs or CPDAGs G, G' is the Structural Hamming Distance (SHD). More suitable for graphs G, G' with a causal interpretation, such as the Structural Intervention Distance (SID) Peters & Bühlmann (2015) and other separation distances; we consider the scores proposed by Wahl & Runge (2025). The *s/c-metrics* (S/C) are based on counting separation statements and comparing their validity in G, G'. Scalable variants thereof are the *separation distances* (SD) that associate each pair of separable nodes in G' with a separation set $\mathcal{S}$ under a given separation strategy, and validate whether $\mathcal{S}$ remains separating in G. We report the SC (without randomization) and the SD where, depending on the method output, we use the DAG or a CPDAG variant of the metrics.

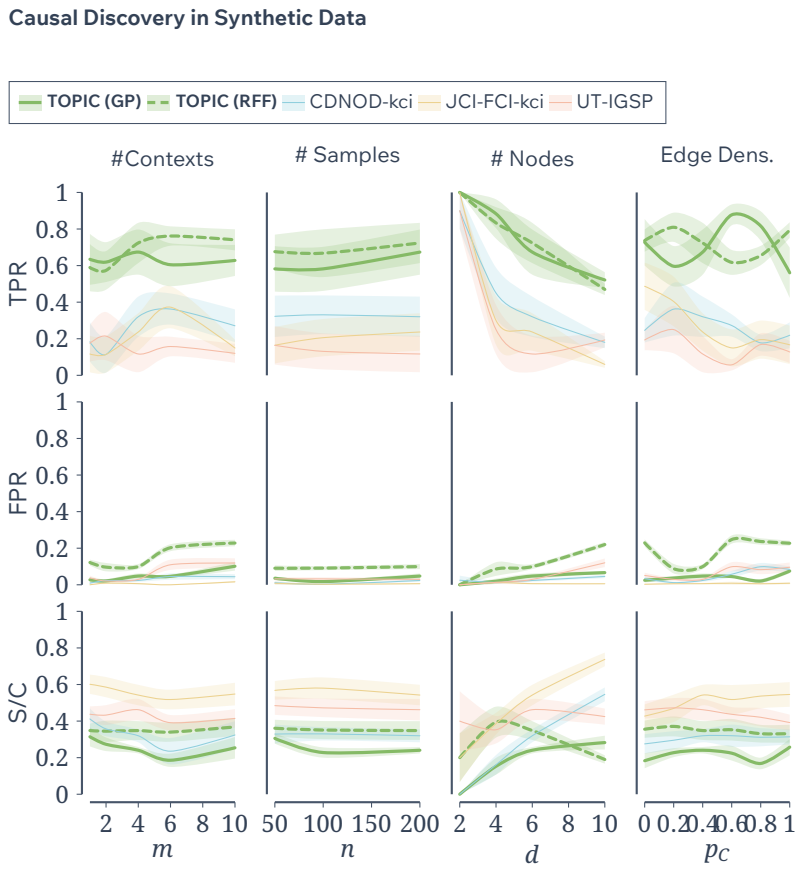


Figure 4.3 To evaluate the methods on discovering causal DAGs on synthetic data, we sample Erdös Rènyi DAGs G^*, use the methods to discover the best G (DAG or CPDAG), and evaluate each edge in G (TPR, FPR) as well as the structural separation distance (SD, lower is better) between G and G^*.

4.5.2 Synthetic Data

Figure 4.3 shows the results on synthetically generated DAGs. While all methods, except for the RFF variant in certain cases, have low false positive rates and rarely misdirect edges (middle, lower is better) we can

Metric	TOPIC
TP	2
FP	16
FN	8
TN	95
SD	0.24
S/C	0.24
SHD	0.75

Table 4.1 DAG evaluation on flow cytometry data (Sachs et al., 2005).

also see that TOPIC has stronger performance in recovering causal edge directions (top, higher is better). The latter also performs well in terms of the separation distances (SD and S/C, lower is better).

4.5.3 Real-world Data

Finally, we consider the real-world flow cytometry dataset curated by Sachs et al. (2005) on causal cell signaling pathways over 11 variables. The dataset originates from single-cell protein-signalling samples of the human immune response system, where each of the 11 observed variables corresponds to the activity of one compound of interest, either a protein or a phospholipid. The dataset contains multiple experimental conditions, in each of which a particular molecular modifier has been applied to the cells, with the intended primary effect of perturbing one of the 11 compounds of interest, resulting in a measurable change in the corresponding activity. A consensus ground truth network is available, although the orientation of the causal relationships is subject to discussion (Mooij et al., 2020). Similar to previous work, we do not utilize background knowledge regarding intervention types or targets. Here, TOPIC recovers only two of the true edges and relatively many false positive edges, as Table 4.1 shows, suggesting that it is challenging to recover the consensus network from the given data (cf. Mooij et al., 2020; Squires et al., 2020); see also Chapter 5 for a study of latent confounding and feedback in this data.

4.6 Conclusion

In this chapter, we address causal DAG discovery from multi-context data. Building on our previous ideas of using IC as a causal asymmetry, we show that we can use it to discover the DAG in topological order.

The central assumption we leverage is that distribution shifts are detectable, exist, and affect nodes independently. Recent works study the detectability of distributions using faithfulness notions (Mazaheri et al., 2025b), which is an important direction of further study.

Regarding the second point, note that if the available contexts do not contain the causal mechanism shifts along the relevant causal paths for a variable pair X and Y, then the score consistency arguments cannot be applied to distinguish their direction. On a positive note, similar guarantees can be given in the absence of changes, assuming a Causal Additive Model (CAM), i.e., that causal functions are additive with respect to their parents Bühlmann et al. (2014); Xu et al. (2025). In that case, we can express the model as a post-nonlinear model in the causal direction, which is known to be identifiable.

Score-based algorithms have been proposed under different regressors, regularizers, and topological pruning (e.g., Bühlmann et al., 2014; Rolland et al., 2022; Mian et al., 2021). Therefore, future work could explore cross-combinations of algorithms, such as GES, CAM, and TOPIC, and scoring criteria, potentially broadening the scope also to discrete and mixed settings or types of interventions other than soft interventions corresponding to causal mechanism shifts. We also do not address richer DAG models here, such as in the setting of causal representation learning (CRL) Schölkopf et al. (2012), where causal models are inferred from high-dimensional or unstructured data, such as that arising from images.

To relax some of our assumptions, we address violations of causal sufficiency in the next chapter.

Software. To make it easy to apply TOPIC, we provide a Python implementation at `github.com/srhmm/causalchange`. As a usage example, we include the Jupyter notebook `demo/ch4_causal.ipynb`.

5 | Identifying Latent Confounding in Multiple Contexts

Case Study Confounding Case Study

In flow cytometry data over multiple proteins and phospholipids (Sachs et al., 2005), the consensus causal network includes a directed edge from Raf to Mek (Figure 5.1, left). Measurements of these variables are available from multiple experimental conditions, one of which includes the reagent U0126. Biological knowledge indicates that U0126 specifically inhibits Mek without directly affecting Raf or other upstream variables. However, in the experiments, U0126 causes a noticeable change in both Mek and Raf (Figure 5.1, right). This contradiction suggests that the signaling system cannot be fully described by a simple DAG.

The goal of this chapter is to relax the common assumption of causal sufficiency by accounting for latent confounding. For motivation, the case study above illustrates that modeling biological systems as acyclic graphs without latent effects is not always realistic.

In the example, prior knowledge suggests a causal edge from Raf to Mek and a intervention with U0126 targeting only Mek, suggesting an independent mechanism shift of Mek in this experimental condition; the actual measurements however show that both variables change jointly

The chapter is based on work published as Mameche et al. (2024).

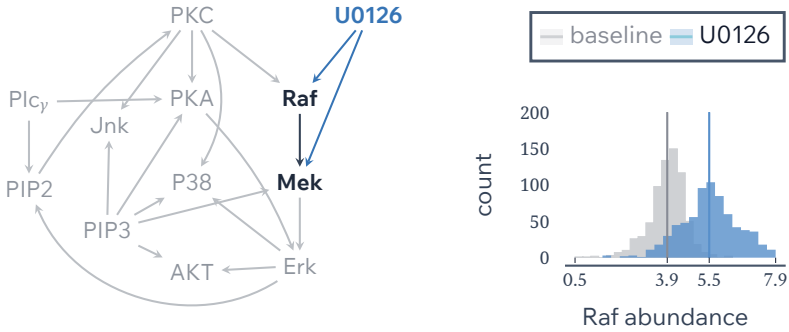


Figure 5.1 The consensus network for the Sachs data, including the Raf-to-Mek edge (left). One experimental condition includes the compound U0126. According to background knowledge, it only affects Mek, but in practice, we also observe an effect on Raf (right), suggesting that the acyclic model without latent effects may be misspecified.

in this experimental context. This inconsistency is documented in the previous literature (Mooij et al., 2020). Possible explanations include the presence of feedback loops in the MAPK pathway (Kolch, 2005) or confounding due to unmeasured components.

While Chapter 3 addresses interventions with unknown targets, it maintains the assumption of causal sufficiency within a Directed Acyclic Graph. However, as seen in the Raf-Mek example, real-world biological responses can manifest differently, for example the intervention on a supposed effect (Mek) induces a change in its supposed cause (Raf). In this chapter, we focus on latent confounding, the existence of unobserved common causes. We focus specifically on detecting latent confounders by measuring the dependence of distribution changes across contexts. Our approach, described in Section 5.1, uses Mutual Information and Total Correlation to measure such dependencies. We provide a theoretical justification for this approach under both known and unknown causal graphs in Section 5.2, introduce a practical algorithm in Section 5.3, and perform an empirical evaluation in Section 5.5.

5.1 Latent Confounding in Multiple Contexts

Besides the observed variables $\mathcal{X} = \{X_1, \dots, X_{d_x}\}$, we consider a set of unobserved random variables $\mathcal{L} = \{L_1, \dots, L_{d_l}\}$. The combined variable set is denoted as $\mathcal{V} = \mathcal{X} \cup \mathcal{L}$. Observations of $\mathcal{X}$ are collected across m different contexts, resulting in the multi-context dataset $\mathcal{D} = \{\mathcal{D}^1, \dots, \mathcal{D}^m\}$.

We extend the causal DAG $G_{\mathcal{V}} = (\mathcal{V}, \mathcal{E}_{\mathcal{V}})$ to include these latent variables. The full graph induces a subgraph $G_{\mathcal{X}} = (\mathcal{X}, \mathcal{E}_{\mathcal{X}})$ over the observed variables only. For simplicity, we assume that confounders are exogenous sources without dependencies among themselves.

Assumption 5.1 (Exogeneity of Confounders). *We assume the latent variables $\mathcal{L}$ are exogenous confounders, i.e., there are no edges $X_j \to L_i$ nor edges $L_i \to L_{i'}$ for any $L_i, L_{i'}$ and X_j.*

To identify these confounders, we put special focus on causal mechanism shifts under similar causal assumptions as in Chapter 2.

We assume causal sufficiency holds over the full variable set $\mathcal{V}$, thereby allowing for violations of sufficiency over the observed variables $\mathcal{X}$. In addition, we assume independent and sparse changes (Assumption 2.4, Assumption 2.6), faithfulness (Assumption 2.2) and change faithfulness (Assumption 2.5) as before. We restate the Markov Property for ease of access below and provide the full set of assumptions in Appendix **??**.

Assumption 5.2 (Causal Markov Property over $\mathcal{V}$). *We assume the causal Markov property over $\mathcal{V}$, namely*

$$p(\boldsymbol{v}, \boldsymbol{\pi}) = p(\boldsymbol{\pi}) \prod_{j=1}^{d_v} p\left(v_j \mid \boldsymbol{pa}_j^{G_{\mathcal{V}}}, \pi_j\right) ,$$

Above, the Π^* includes the selection nodes for both observed and latent variables. Central to our approach is the observation that, while the joint distribution $P_{\mathcal{V}}$ factorizes according to the Markov property in Assumption 5.2, the marginal observed distribution $P_{\mathcal{X}}$ generally does not. This has implications on the observed distribution shifts, as follows.

5.1.1 Effects of Confounding on Observed Changes

Our main observation is that distribution shifts inferred from observed data need not correspond to independent causal mechanism changes, as shifts induced by latent confounders propagate across multiple variables.

In particular, distribution shifts induced by confounders propagate to their observed descendants. To illustrate this effect, consider a bivariate example.

Example 5.1 Assume a variable pair X, Y that are generated in the cth context as

$$\begin{aligned} L &\sim N(0, \sigma_l^2(c)) \\ X &= \alpha L + \epsilon_x \\ Y &= \beta X + \gamma L + \epsilon_y \,, \end{aligned}$$

where only the variance $\sigma_l^2(c)$ changes across contexts. If we regress Y on X, we obtain

$$\begin{aligned} X &\sim N(0, \sigma_x^2 + \alpha^2 \sigma_l^2(c)) \\ \widehat{\beta}_{Y|X} &= \frac{\operatorname{cov}(X,Y)}{\operatorname{var}(X)} = \frac{\beta \sigma_x^2 + \alpha\gamma\sigma_l^2(c)}{(\sigma_x^2 + \alpha^2 \sigma_l^2(c))} \,, \end{aligned} \tag{5.1}$$

Therefore, if $\sigma_l^2(c) \neq \sigma_l^2(c')$ for contexts c and c', both the marginal distribution $P(X)$ and the conditional distribution $P(Y \mid X)$ change. △

Example 5.1 shows the distribution shift of L is propagated through the functional mechanisms of its children. As in Chapter 2, we consider variables Π_j as indexing groups of contexts that share the same conditional distribution of X_j given its parents. Then, the inferred mechanism selection variables Π_j and Π_k in Example 5.1 share statistical dependence, even when the true causal mechanisms are independent.

Formally, to distinguish between the causal ground truth and the case where $\mathcal{L}$ remains unobserved, we denote by Π^* the mechanism-selection variables in the full causal graph $G_{\mathcal{V}}$. In contrast, Π denotes the random variables corresponding to the observed conditional distributions under $G_{\mathcal{X}}$, which we infer from data. These notations are summarized in Table 5.1.

Our identification results will rely on showing that while Π_j^* and Π_k^*

Variable	Meaning	Property
Π^* (causal)	Changes in $P(V_j \mid Pa_j^{G_V})$	Independent (Assumption 2.4)
Π (inferred)	Changes in $P(X_j \mid Pa_j^{G_X})$	Dependent if confounded

Table 5.1 Comparison between true mechanism selection nodes and observed distribution shifts across contexts.

are independent (by IC), the estimated Π_j and Π_k will generally inherit dependencies if a shared latent parent exists.

In the following, we address how to measure such dependencies.

5.1.2 Measuring Dependence of Changes

Consider a variable pair X_1 and X_2 in a graph G, with their observed distribution changes across contexts given by the variables Π_1 and Π_2.

To detect confounding, we measure the dependence between Π_1 and Π_2. These variables can be interpreted as clusterings of the m contexts into I index sets π_1^i and J index sets π_2^j, respectively. To quantify the overlap in these clusters, we consider the Mutual Information (MI), a standard measure for comparing clusterings (Vinh et al., 2009).

The MI is calculated using a contingency table $\mathcal{M} = (n_{ij})$ where each entry n_{ij} counts the number of contexts in which the i-th clustering of X and the j-th clustering of Y overlap. That is, each $n_{ij} (i = 1, \dots, I, j = 1, \dots, J)$ counts the number of contexts in $\pi_1^i \cap \pi_2^j$ for each pair of sets π_1^i, π_2^j, with row margins $u_i = |\pi_1^i|$ and column margins $v_j = |\pi_2^j|$ counting their elements.

The marginal entropy of Π_1 and the joint entropy of Π_1, Π_2 are given by

$$H(\Pi_1) = -\sum_i \frac{u_i}{m} \log \frac{u_i}{m}, \qquad \text{and} \qquad H(\Pi_1, \Pi_2) = -\sum_{ij} \frac{n_{ij}}{m} \log \frac{n_{ij}}{m},$$

with $H(\Pi_2)$ defined similarly. Then, the MI is given by

$$\mathrm{I}(\Pi_1, \Pi_2) = H(\Pi_1) + H(\Pi_2) - H(\Pi_1, \Pi_2) = \sum_{ij} \frac{n_{ij}}{m} \log \frac{n_{ij} m}{u_i v_j}.$$

Because the raw MI estimate is positively biased for finite m (Vinh et al., 2009) we standardize the score by comparing it to the Expected Mutual Information (EMI) under a null hypothesis of independent clusterings. Assume independent Π_1', Π_2' with contingency table $\mathcal{M}$ with row margins u and column margins v. To characterize the mutual information under independence, the literature adopts a hypergeometric model of randomness (Vinh et al., 2009; 2010). That is, given the marginal counts u and v, the joint count N_{ij} is assumed to follow a hypergeometric distribution $N_{ij} \sim \mathcal{H}(u, v, m)$, with probability mass function

$$\Pr(N_{ij} = n_{ij} \mid a_i, b_j, m) = \frac{\binom{a_i}{n_{ij}}\binom{m-a_i}{b_j-n_{ij}}}{\binom{m}{b_j}}.$$

The expected mutual information between the independent clusterings is

$$\mathbb{E}\left[\mathrm{MI}(\Pi_1', \Pi_2')\right] = \sum_{\mathcal{M}} \mathrm{MI}(\mathcal{M}) \Pr(\mathcal{M}) = \sum_{ij} \sum_{n_{ij}} \mathrm{MI}(n_{ij}) \Pr(n_{ij} \mid u, v, m) \tag{5.2}$$

where $\mathrm{MI}(n_{ij}) = \frac{n_{ij}}{m} \log \frac{n_{ij} m}{u_i v_j}$ and $n_{ij} \in \left[\max\{0, u_i + v_j - m\}, \min\{u_i, v_j\}\right]$. By replacing the term $\mathrm{MI}(n_{ij})$ by $\mathrm{MI}(n_{ij})^2$, one can similarly obtain the second moment, and thus the variance

$$\mathrm{Var}(\mathrm{MI}(\Pi_1', \Pi_2')) = \mathbb{E}\left[\mathrm{MI}(\Pi_1', \Pi_2')^2\right] - \mathbb{E}\left[\mathrm{MI}(\Pi_1', \Pi_2')\right]^2 .$$

With this, we can compute the standardized score of our observed mutual information $\mathrm{I}(\Pi_1, \Pi_2)$

$$t = \frac{\mathrm{I}(\Pi_1, \Pi_2) - \mathbb{E}\left[\mathrm{MI}(\Pi_1', \Pi_2')\right]}{\sqrt{\mathrm{Var}(\mathrm{MI}(\Pi_1', \Pi_2'))}} . \tag{5.3}$$

To determine whether a set of variables shares a common confounder, we extend our score to sets of nodes $X_1, \ldots, X_s$ where again $\Pi_1, \ldots, \Pi_s$ show their changes in each context. To measure dependencies among the set of clusterings, we use the total correlation T, a natural multivariate extension of mutual information (Watanabe, 1960), defined as

$$T(\Pi_1, \ldots, \Pi_s) = \sum_i H(\Pi_i) - H(\Pi_1, \ldots, \Pi_s) = \sum_i I(\Pi_i, \Pi_{>i} \mid \Pi_{<i})$$

where $\Pi_{<i} = \{\Pi_1, \ldots, \Pi_{i-1}\}$ and similarly for $\Pi_{>i}$. This score can be corrected analogously to the pairwise MI case. As both corrected and uncorrected scores are asymptotically equivalent, we will consider T as is in our theoretical analysis.

The mutual information and total correlation thus measure the dependence of mechanism shifts among pairs or a set of variables, which we use to identify joint confounding as follows.

5.2 Identifying Confounding Using IC

We now show that these dependence measures are theoretically sufficient to identify confounding. All assumptions and proofs are detailed in Appendix E.

5.2.1 Pairwise Confounding

For now, assume that only the causal graph is known and the confounders are not, i.e. given is the subgraph $G_{\mathcal{X}}$ over $\mathcal{X}$ and observed distribution shifts Π. We show that confounding induces a dependence in Π. Following Example 5.1, we consider the bivariate case first.

Lemma 5.1 (Detecting Confounded Variable Pairs)**.** *Let X_j, X_k be a pair of variables that is either confounded through $L \in \mathcal{L}$, or unconfounded with causal relationship $X_j \to X_k$. Let Π_j and Π_k be the variables showing changes of X_j and X_k under $G_{\mathcal{X}}$, and t the score of $\mathrm{I}(\Pi_j, \Pi_k)$ in (5.3). Then we have*

$$\lim_{m\to\infty} Pr(t > q_{1-\alpha}) = \begin{cases} 1, & X_j, X_k \text{ confounded} \\ \alpha, & \text{otherwise}, \end{cases}$$

for any $\alpha > 0$, where $q_{1-\alpha}$ is the $1-\alpha$-quantile of the standard normal distribution.

Proof Sketch. If X_j and X_k are confounded, the observed Π_j and Π_k inherit all distribution shifts from Π^*_L, making them statistically dependent such that $\mathbb{E}[\mathrm{I}(\Pi_j, \Pi_k)] \geq \frac{m}{2} H(p) > 0$. If they are unconfounded (e.g., $X_j \to X_k$), Independent Change (Assumption 2.4) ensures that the true Π^*_j and Π^*_k are independent. In this case, the standardized t-score is derived from

a hypergeometric model of randomness, and as $m \to \infty$, t converges to a standard normal distribution by the properties of mutual information under independence (Vinh et al., 2009). □

Intuitively, this inequality holds when the information shared between X_j and X_l is already contained within X_k . If it were a single joint confounder, the total correlation would be higher because all three variables would be directly coupled to a single source. We extend this result beyond the bivariate case in the following.

5.2.2 Joint Confounding and Minimality

When considering three or more variables, we face an additional task of distinguishing between joint confounding, where a single latent affects all variables, and multiple pairwise latents. To ensure identifiability of these cases, we make the following minimality assumption.

Assumption 5.3 (Confounder Minimality). *For every subset $\mathcal{X}_S$ of size at least $|\mathcal{X}_S| \geq 4$, at most $2\,|\mathcal{X}_S|$ edges enter $\mathcal{X}_S$ from latent confounders L_i with at least three children in $\mathcal{X}_S$.*

This assumption prevents pathological structures where many overlapping pairwise confounders exactly mimic the behavior of a single joint confounder. Under this assumption, we can identify joint confounding.

Theorem 5.1 (Detecting Confounded Variable Groups). *Let $\mathcal{X}_S$ be a set of variables such that every pair $X_j, X_k \in \mathcal{X}_S$ is confounded, and let Π denote the mechanism-assignment variables induced by the observed distribution shifts under G_X. Then $\mathcal{X}_S$ is jointly confounded by the same latent variable L if and only if*

$$\lim_{m\to\infty} Pr\big(\mathrm{T}(\Pi_j, \Pi_k, \Pi_l) < \mathrm{I}(\Pi_j, \Pi_k) + \mathrm{I}(\Pi_k, \Pi_l)\big) = \begin{cases} 1, & X_j, X_k, X_l \text{ jointly confounded} \\ 0, & \text{otherwise}, \end{cases}$$

which holds for each triple X_j, X_k, X_l in $\mathcal{X}_S$.

Proof Sketch. The condition $T(\Pi_j, \Pi_k, \Pi_l) < \mathrm{I}(\Pi_j, \Pi_k) + \mathrm{I}(\Pi_k, \Pi_l)$ is mathematically equivalent to the conditional independence of partitions $\mathrm{I}(\Pi_j, \Pi_l \mid$

$\Pi_k) < \mathrm{I}(\Pi_j, \Pi_l)$. This "v-structure" in the partition space implies that the observed dependencies between mechanisms are not merely chained (e.g., $X_j \leftarrow L \rightarrow X_k \leftarrow L' \rightarrow X_l$) but arise from a shared source common to all three variables. Under the Confounder Minimality (Assumption 5.3), any configuration of multiple pairwise confounders designed to mimic this triplet-wise dependency would require strictly more than $2|X_S|$ edges entering the set, violating the edge-sparsity constraint. Thus, as $m \rightarrow \infty$, the total correlation T correctly distinguishes a single joint L from a collection of pairwise $\mathcal{L}$. □

Therefore, both the number of latent confounders as well as the sets of jointly confounded nodes associated with each latent variable are uniquely identifiable from the total correlation.

However, note that we assumed access to the true causal relationships up to this point, since access to the observed subgraph G_X is required to infer Π for each node X given its parents in G_X. As the last step, we extend the results to the case of an unknown graph.

5.2.3 Confounding with Unknown Causal Graph

In practice, the causal DAG over observed variables is often unknown. We now show that minimizing the Total Correlation is a consistent strategy for both discovering the graph and the confounders simultaneously.

First, consider a variable pair X_j, X_k with causal direction $X_j \rightarrow X_k$ applies. Then we know from Theorem 3.2 that misdirecting this edge adds additional dependencies between Π_j, Π_k, and from Lemma 5.1 that confounding does the same. Combining these ideas, we show that the MI

Lemma 5.2 (Pairwise Consistency). *Let X_j, X_k be a variable pair confounded by a latent L, with true mechanism shifts Π_j^*, Π_k^* and distribution shifts Π_j, Π_k under the misdirected edge. Then there exists some $\rho > 0$ such that*

$$Pr\Big(\mathrm{I}\left(\Pi_j^*, \Pi_k^*\right) < \mathrm{I}\left(\Pi_j, \Pi_k\right)\Big) = 1 - O\left(e^{-\rho n_c}\right) .$$

Proof Sketch. An incorrect edge orientation or unmodeled confounder introduces spurious dependencies between observed mechanism shifts. We model the failure to detect such a coupling across pairs of contexts as a Bernoulli trial. The probability that a shift in Π_L^* fails to manifest

as a detectable dependency in the observed conditionals is bounded by a constant strictly less than 1 (specifically $p + (1 - p)(1 - p + p^2)$). As the number of contexts m increases, these independent opportunities to observe the coupling accumulate. Consequently, the mutual information of the true, independent mechanisms remains near zero, while the I of any misdirected or confounded observed mechanism grows at a rate of $O(m)$, making the gap certain in the limit $1 - O(e^{-\rho m})$. □

Finally, consider larger graphs. Here, a common problem is that confounding introduces a number of spurious dependencies, for example the Markov Equivalence Class (MEC) over the observed $\mathcal{X}$ is known to contain additional spurious edges (Elidan et al., 2000; Kaltenpoth & Vreeken, 2023a). Therefore, for our results we need to bound the number of spurious siblings of a given target X_j obtained in the MEC. When they are not too large, we can show that we can recover the causal parents of each variable as follows.

Theorem 5.2 (Consistency). *Let $G_{\mathcal{V}}$ be the true graph over $\mathcal{V}$ and $G_{\mathcal{X}}$ be the corresponding subgraph over $\mathcal{X}$. Assume further that for each X_j, the number of latent confounders plus spurious siblings in the MEC of the observed distribution $P_{\mathcal{X}}$ is bounded by $\frac{log(0.5)}{log(1-p)}$.*

Then with high probability, $G_{\mathcal{X}}$ is the unique minimum of total correlation,

$$Pr\Big(\operatorname*{arg\,min}_{G,\Pi \text{ under } G} T(\Pi_1, \ldots, \Pi_d) = \{G_{\mathcal{X}}\}\Big) = 1 - O(e^{-\rho m}) \ ,$$

where the minimization ranges over graphs G and the corresponding variables Π under this graph, i.e., $\Pi = \{\Pi_1, \ldots, \Pi_d\}$ show distribution shifts of each conditional $P(X_j \mid Pa_j)$ in G.

Proof Sketch. By the definition of Total Correlation,

$$T(\Pi_1, \ldots, \Pi_d) = \sum_j \mathrm{I}(\Pi_j, \{\Pi_k : k \in Pa_j\}) \ .$$

Any graph structure G other than the true subgraph $G_{\mathcal{X}}$ will either omit a true parent or include a spurious child, both of which introduce joint shifts into the observed conditionals. The bound on spurious siblings ensures that the probability of shifts in Π_j^* being obscured remains below 0.5, preserving the power of the T score. By applying a union bound

over the candidate graphs in the Markov Equivalence Class, the true structure $G_{\mathcal{X}}$ emerges as the unique global minimizer of total correlation as $m \to \infty$. □

The result shows that, when our assumptions hold, minimizing the total correlation T among Π is sufficient to discover the true graph and its confounders.

In practice, however, this requires a two-stage procedure where for each G, one needs to apply the confounding tests to subsets $\mathcal{X}_S$, which may induce a large number of tests and leaves open how to choose the subsets efficiently. The next section therefore proposes a practical instantiation based on pairwise tests combined with spectral clustering.

5.3 Algorithm

We now introduce the practical implementation of our framework, which we call COCO for discovering confounding in multiple contexts.

Whereas our theoretical results in Section 5.2 establish asymptotic consistency of MI and TC over all possible graphs and variable subsets, a global search over these components is computationally infeasible. The main idea in COCO is therefore to provide a tractable approximation using clustering of variables under the mutual information, instead of direct statistical testing. We detail the algorithm below.

Detecting Mechanism Changes. Assume first that the causal graph G is known. The first phase of the algorithm has a similar goal as Chapters 3-4, namely detecting mechanism changes for each X_j given its parents in G.

To this end, for every pair of contexts c, c', we test the null hypothesis

$$H_0 : \; P^c\left(X_j \mid Pa_j^G\right) = P^{c'}\left(X_j \mid Pa_j^G\right) \; .$$

In practice, this hypothesis is tested via a conditional independence test between X_j and the context index given its parents, which is equivalent to testing invariance of the conditional distribution under standard assumptions. We use the Kernel Conditional Independence (KCI) test (Zhang et al., 2011) for this purpose, and a marginal test based on Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) when a variable has no parents in G.

Given the results of these $O(m^2)$ tests, we estimate the variable Π_j as a clustering of the m contexts. In practice these pairwise tests could give intransitive results due to finite-sample noise, for example assigning pairs c, c' and c', c'' to the same mechanism but c, c'' to a different one. We resolve these by favoring the detection of changes to maintain high sensitivity to detecting confounding, but other choices are possible.

Discovering Confounding. As we show in Theorem 5.1, confounding of a set of variables results in positive and typically increasing Mutual Information (MI) over mechanism changes across all variable pairs in the set. However, a naive approach that tests every possible subset of such variables would need 2^d tests and furthermore create a multiple-testing problem. We therefore frame the search for confounded sets as a global clustering problem.

To this end, we define an affinity matrix $\mathcal{M} \in \mathbb{R}^{d \times d}$ where each entry $\mathcal{M}_{jk}$ is the standardized t-score of the mutual information between Π_j and Π_k,

$$\mathcal{M}_{jk} = \frac{\mathrm{I}(\Pi_j, \Pi_k) - \mathbb{E}[\mathrm{I}]}{\sqrt{\mathrm{Var}(\mathrm{I})}} \,.$$

We then apply Spectral Clustering (Donath & Hoffman, 1972) to $\mathcal{M}$. This choice is theoretically motivated from Theorem 5.1, as spectral clustering identifies groups of variables with high pairwise MI. According to our minimality assumption, variables sharing a latent L will form a dense subgraph in the affinity matrix, therefore appearing as clusters. While our theoretical results are formulated in terms of total correlation, the use of pairwise mutual information provides a tractable surrogate that preserves the clustering structure induced by shared confounders in the asymptotic regime.

Discovery with Unknown Graphs. Finally, the description above assumes a fixed parent set Pa_j under the causal graph G. In cases where the DAG is unknown, Theorem 5.2 guarantees that the true causal structure is the global minimizer of the Total Correlation T. Therefore, we can use CoCo in a search-and-score procedure similarly to previous proposals (Perry et al., 2022).

We begin with the Markov Equivalence Class (MEC) of the graph, for example, obtained from a causal discovery algorithm. Then, for each candidate DAG in the MEC, we estimate the partitions Π and the resulting Total Correlation T. Importantly, because the mechanism partitions Π_j

Algorithm 5.1: COCO

input : Dataset $\mathcal{D} = \{\mathcal{D}^1, \ldots, \mathcal{D}^m\}$ over $\mathcal{X}$ in m contexts
set of graph hypotheses $\mathcal{G} = \{G_1, \ldots, G_{max}\}$
output: confounded variable sets $\mathcal{X}_S^1, \ldots, \mathcal{X}_S^{d_l}$ and graph G

1 **foreach** graph $G \in \mathcal{G}$ **do**
2 **foreach** variable X_j **do**
3 **foreach** pair of contexts c, c' **do**
4 $p_{c,c'} = \mathrm{KCI}\left(X_j^c \mid Pa_j^G, X_j^{c'} \mid Pa_j^G\right)$
5 Infer Π_j from $p_{c,c'}$ *// mechanism changes*
6 **foreach** pair of variables X_j, X_k **do**
7 $\mathcal{M}_{jk} = \frac{\mathrm{I}(\Pi_j, \Pi_k) - \mathbb{E}[\mathrm{I}]}{\sqrt{\mathrm{Var}(\mathrm{I})}}$ *// pairwise confounding*
8 $\{\mathcal{X}_S^1 \ldots \mathcal{X}_S^{d_l}\} = \mathrm{SpectralClustering}(\mathcal{M})$ *// setwise confounding*
9 **return** graph G and confounded subsets minimizing T

depend on the conditional distribution $P^c(X_j \mid Pa_j^G)$ (Table 5.1), we need to re-estimate these whenever the candidate parent set changes. Thus for each candidate DAG in the MEC, we re-partition the contexts, compute the resulting Total Correlation T, and select the graph that maximizes independence of the mechanism variables. Following the consistency result, we select the structure that minimizes T.

CoCo. Algorithm 5.1 summarizes the pseudocode for CoCo. In the first phase, we test all pairs of contexts for mechanism shifts (l. 1–4), and repeat this for each variable to obtain its partition. In the second phase, we test all pairs of variables for confounding (l. 5–6). Finally, we cluster the variables into subsets that are affected by the same confounder (l. 7–8).

5.4 Related Work

Causal discovery is traditionally limited by the causal sufficiency assumption, the requirement that no latent confounders exist (Pearl, 2009), as well as the i.i.d. assumption. The previous literature generally addresses the challenge of relaxing either but not both assumptions.

Relaxing Causal Sufficiency. Constraint-based algorithms like the FCI family (Spirtes et al., 2000) can detect potential confounding, but often yield partial results because they identify variables as possibly confounded. More recent approaches like Nested Markov Models (NMMs) utilize Verma constraints for identifiability (Shpitser et al., 2018; Richardson et al., 2023). Approaches like DCD (Bhattacharya et al., 2021) and those by Kaltenpoth & Vreeken (2023a) use differentiable constraints or model latents directly, but often require strong parametric assumptions.

Relaxing i.i.d.-ness. As discussed in more depth in the preceding chapters, the recent literature leverages data from different contexts to improve identifiability (Huang et al., 2020; Mooij et al., 2020), where the approach by Perry et al. (2022) shares the most similarity to our approach as it also relies on the independent and sparse shift principle. In comparison to their statistical testing procedure using p-values, we rely on the soft dependence measure of Mutual Information as well as a clustering idea to discover confounders.

Relaxing Both Assumptions. Related work at the intersection of these two assumption relaxations is relatively sparse. The most closely related framework is Joint Causal Inference (JCI) (Mooij et al., 2020), which can be combined with FCI to allow for latent variables in multi-context settings. Unlike JCI-FCI, which relies on standard conditional independence constraints, COCO explicitly uses the Independent Change principle to detect confounders that may be active across or within specific contexts. In the following section, we provide a direct empirical comparison between CoCo and the JCI-FCI approach.

5.5 Experiments

To conclude, we evaluate COCO on synthetic and real-world data with two primary goals,

(*i*) *Detecting Confounding*, where we evaluate the identified confounded variable sets, and

(*ii*) *Discovering Causal Directions*, where we assess the identification of directed edges in the presence of confounders.

5.5.1 Experimental Setup

As in previous chapters, we generate data under an Erdös-Rényi model using $V_j = \sum_{k \in Pa_j} \omega_{jk}^{(c)} f_{jk}(V_k) + \sigma_j^{(c)} N_j^{(c)}$ with weights $\omega_{jk}^{(c)} \sim \mathcal{U}(0.5, 2.5)$, either uniform or Gaussian noise, and f sampled from $\{x^2, x^3, \tanh, \text{sinc}\}$. New to this setting, we additionally keep some causal parents $\mathcal{L} \subseteq Pa_j$ unobserved. For simplicity, we assume that each confounder L_i is a source node with edges to a random subset of at least two variables. We again simulate changes in causal mechanisms by resampling functional coefficients.

To separate the effects of discovering latent variables, mechanism changes, and causal directions, we include different oracle variants. To study our confounding test in isolation, we consider an oracle for the true partitions, named COCO-Π^*. To study testing for mechanism shifts, but under a known graph, we use COCO-G^*, which uses the causal structure G^* as background knowledge and then applies Algorithm 5.1. Third, we also evaluate causal direction discovery. For this, we combine our approach with MSS (Perry et al., 2022) using the kernelized conditional independence test (KCI) (Zhang et al., 2011) to discover a fully directed DAG G from the true MEC, and then apply Algorithm 5.1 to this DAG.

Our main point of comparison is JCI (Mooij et al., 2020) instantiated with the FCI algorithm (Spirtes et al., 2000), referred to as JCI-FCI. It applies FCI to an augmented causal model including the context variable and appropriate edge constraints (Mooij et al., 2020), and returns, for each variable pair, whether the relation is causal, confounded, potentially confounded, or none of the above. We also apply FCI to the pooled data from all contexts, FCI(C), and to the data of each context individually, reporting the best of these results as FCI(c^*).

5.5.2 Synthetic Data

In our main experiment, we evaluate whether the methods correctly discover confounding in a multi-context DAG G. As the FCI variants can only determine potential confounding for node pairs, we evaluate confounding decisions for each node pair X_j, X_k in G (F1-confounded). We show the results for different parameters, including the number of contexts (m), the number of observed (d) and latent (d_l) variables, and the number of observed (s_X) and latent (s_L) mechanism shifts. We start

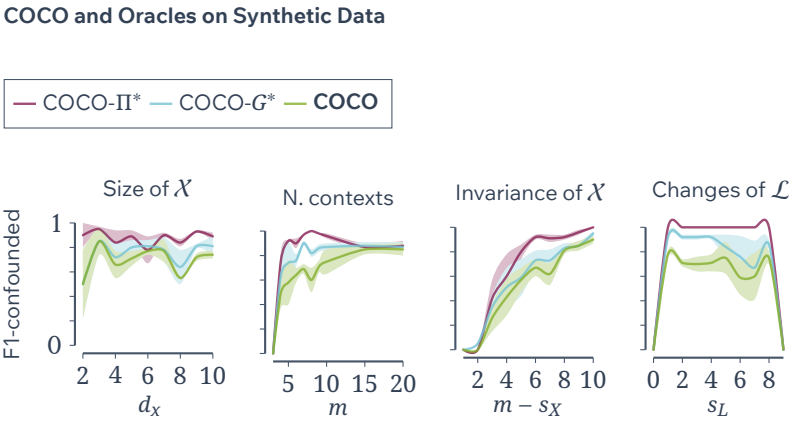


Figure 5.2 On synthetic data, we show COCO with and without oracles, with F1 scores evaluating pairwise confounding (F1-confounded).

from the parameters ($m = 10, d = 10, d_l = 1, s_X = 1, s_L = 2$) and change one parameter at a time.

We show the oracle results in Figure 5.2 and the full results in Figure 5.3. As expected from the theoretical analysis, the confounding detection with COCO improves with more contexts (subplot over m). The improvement is especially strong when we increase the number of those contexts where observed variables do *not* change while keeping confounder changes (s_L) fixed (subplot over $m - s_X$); this is expected because the contexts where confounded variable pairs X, Y do not change are those that allow us to detect joint shifts caused by L. Conversely, when changes in X, Y, or L are too dense (e.g., $m - s_X = 1$ and $s_L = 1$), or when changes in L are too sparse, we cannot reliably discover confounding. Aside from these cases, ours (green) clearly outperforms JCI-FCI (yellow) by a large margin in discovering confounders, while JCI-FCI in turn has a slight advantage over FCI applied either to the best-scoring single context or to pooled data (blue).

The gap between the oracle versions and full ours remains small in practice, supporting the results with unknown causal graph (Theorem 5.2). Finally, we also show how many causal edges are correctly oriented in Figure 5.3, bottom. As expected, we perform well under sparse shifts

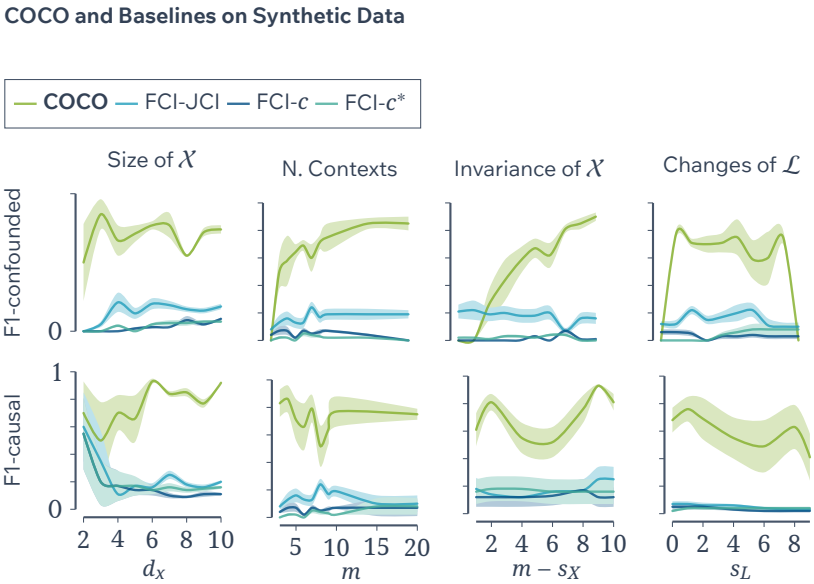


Figure 5.3 On synthetic data, we show COCO and baselines, with F1 scores evaluating pairwise confounding (F1-confounded) as well as on pairwise causal edge directions (F1-causal).

and with more contexts, while the FCI variants generally only discover a few causal edges.

We also perform an ablation study considering the robustness to the assumption of sparse change.

Sparse Changes (Assumption 2.6). The key assumption in our analysis is the sparse mechanism shift assumption. To study its effect, we vary the number of changes for the observed (s_X) and latent variables (s_L) in a fixed set of contexts, here $m = 20$. We generate data as in our main experiments and test with $COCO(\Pi^*)$ for confounding between all node pairs in a causal DAG. To evaluate the empirical power of our confounding test, we report the true positive rate (TPR-confd) over these decisions in Figure 5.4.

In Figure 5.4, left, we show the effect of increasing s_X. We run the experiment for each s_L, and color plots red if latent shifts are sparse

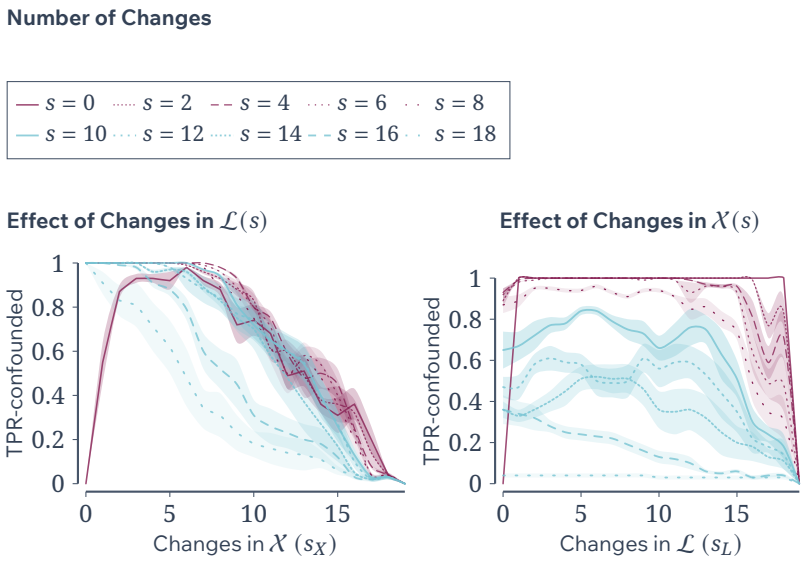


Figure 5.4 We show the power of our confounding test (the true positive rate of decisions over node pairs, higher is better) depending on the number of observed mechanism changes, $s := s_X$ (left) respectively of latent variables, $s := s_L$ (right) in $m = 20$ contexts. We can identify confounding when observed mechanism shifts are sparse ($s_X < 10$, red plots on the right) unless the confounder changes in almost every context ($s_L > 15$, blue plots on the left) or remains unchanged ($s_L = 0$).

($s_L < 10$) and blue otherwise. We observe a tipping point at $s_X = 9$ where the observed nodes have partitions consisting of 10 distinct groups among the 20 contexts. That is, exactly when mechanism shifts are no longer sparse, the power of our test decreases. For $s_X < 10$, we have perfect power in most cases. We note that in the special case where all variables have the same distribution in all contexts, $s_L = 0, s_X = 0$, the confounding effect is not measurable using our method.

In Figure 5.4, right, we show the same result when we increase s_L instead. Sparse shifts $s_X < 10$ are now colored red, dense shifts blue. We can see a clear separation of the two cases, confirming our observations above. In particular, under sparse shifts of s_X, we can tolerate up to $s_L = 15$ shifts of the confounder.

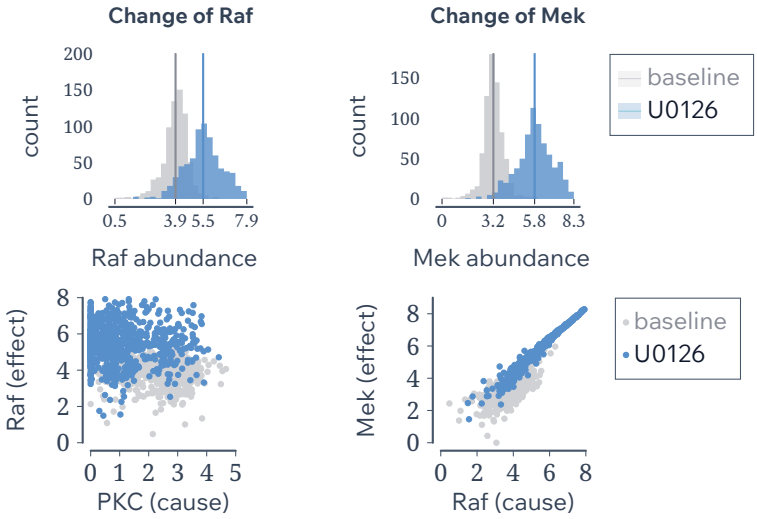


Figure 5.5 The joint change in causal conditionals for Mek (left) and Raf (right) given their causes under the consensus network. Both variables change jointly under the reagent U0126 (blue).

We conclude that our approach works best in settings where the sparse shift assumption holds for the observed variables, while we can handle more shifts for the latent variables.

5.5.3 Case Study: Pathway Problem II (Revisited)

To conclude, we return to the example in our motivating case study. The observation that Raf changes under the interventional condition U0126, contrary to background knowledge, was first noted by Mooij et al. (2020). We obtain the same result when applying COCO to the data under the consensus network.

For illustration, we show the marginal distributions in Figure 5.5 (top). Then, COCO compares the distributions of $X_j \mid Pa_j$ for Raf resp. Mek given their parents in the consensus network (Figure 5.1, left). The result shows that both differ in the condition c = U0126 from the observa-

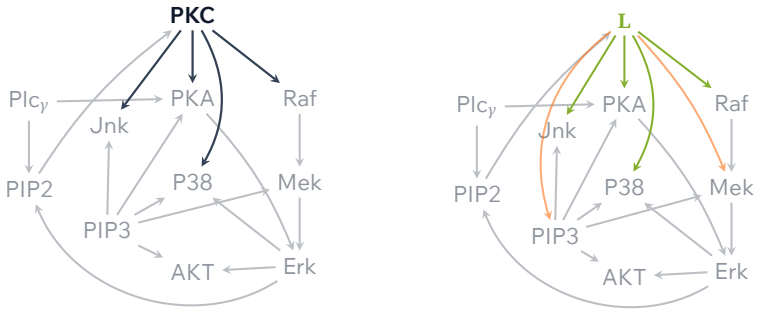


Figure 5.6 On the Sachs et al. (2005) dataset, COCO discovers the confounding effects of PKC. Solid green edges are correctly discovered as confounded, and orange edges are spurious.

tional baseline. We show the respective conditionals in the bottom of the illustration, where we can also visually see the distribution shift. Hence, such an observation could suggest to domain experts that the background knowledge represented in the consensus network and interventions is incomplete or misspecified, and suggest potential feedback between the variables.

We end with a larger case study on the flow cytometry dataset by Sachs et al. (2005). To study confounding effects in this dataset, we follow the study design of Kaltenpoth & Vreeken (2023b). Starting from the consensus causal network in Figure 5.6, we keep PKC hidden and provide the data over the remaining variables in nine contexts to COCO. As we illustrate in Figure 5.6, COCO correctly discovers a latent L and all of its outgoing edges (green) as well as two spurious ones (dashed). To ensure that COCO does not return spurious confounding, we repeat the experiment while keeping each other node in turn hidden. Notably, in all of these cases, we discovered Raf and Mek to be confounded, further suggesting the possibility of unmeasured confounding or feedback. Other than that, however, ours returns only one more false positive edge, and correctly rejects confounding in all other cases.

5.6 Conclusion

In this chapter, in addition to relaxing i.i.d.-ness, we also relax the assumption of causal sufficiency. That is, besides considering multi-context data with unknown mechanism changes (causal edges $\Pi_j \to X_j$), we also discover latent confounding (causal edges $L_i \to X_j$). Perhaps surprisingly, the former makes confounding easier to detect, since we can identify confounding from correlation patterns in the observed variables.

Specifically, we interpret causal mechanism changes as clusterings Π_j of the m contexts into subsets with invariant causal mechanism, and we base our test on the mutual information (MI) $\mathrm{I}(\Pi_j, \Pi_k)$ to detect confounded node pairs, or more generally total correlation $T(\Pi_1, \ldots, \Pi_s)$ to detect confounded node sets.

Comparing this approach to Chapter 3, the identification results are similar (Theorem 3.2; Theorem 5.2), showing that independent changes can be used both to identify cause from effect and, in the present setting, to distinguish the causal case $X \to Y$ from the confounded case $X \leftarrow L \to Y$. The main difference is that Chapter 3 explores a score-based functional modeling approach, where IC implicitly enters through the regularization, whereas here we use a correlation-based approach where we explicitly test for IC.

There are also more complex cases of confounding such as hierarchical confounding, such as latent chains ($L_1 \to L_2$) or overlapping influences. COCO identifies the minimal sufficient set of latent sources required to explain the observed dependencies, and as such could work in such cases but would show a single confounder (L_{12}). While a latent chain may be mathematically indistinguishable from a single effective confounder, our use of Total Correlation ensures that overlapping confounders are not collapsed into a single group.

A fundamental limitation of COCO is that it discovers confounding from distribution changes and, while this circumvents specific assumptions on the generating process, it is not applicable in the absence of changes such as in i.i.d. observational data, or when too few contexts are observed. Therefore, it may be of interest for future work to combine this approach with existing methods for confounding discovery, such as using rank constraints (e.g., Huang et al., 2022).

Software. To allow the reader to use the code, a Python implementation at `github.com/srhmm/causalchange` is available. To get started, the Jupyter notebook `demo/ch5_confounding.ipynb` shows an example usage.

6 | Causal Graph Discovery from Non-Stationary Time Series

Case Study Time Trial

In hydrology, researchers monitor water levels and meteorological variables at sites, called catchments, where surface water drains into a common outlet (Cornes et al., 2018), for example, monitoring variables such as temperature T, precipitation P, and river runoff Q over time and at different sites in Europe (Figure 6.1). In addition to differing local characteristics (such as area, terrain, and elevation), temporal changes (including hydrometeorological processes like snowmelt and climate events like droughts) occur at each site, leading to non-stationary time series, illustrated here for Q (Figure 6.1, left).

In this chapter, we extend the previous approaches to time series settings, where measurements are obtained repeatedly over time.

In this setting, observations originate from distinct geographical locations and are monitored as discrete-time processes. Due to temporal events, seasonality, or other time-dependent patterns changes in the generating process occur in time at unknown changepoints, hence we refer to non-stationary time series.

This needs a framework capable of both representing these causal interactions and identifying the latent changepoints. Here, we address this

The chapter is based on work published as Mameche et al. (2025a).

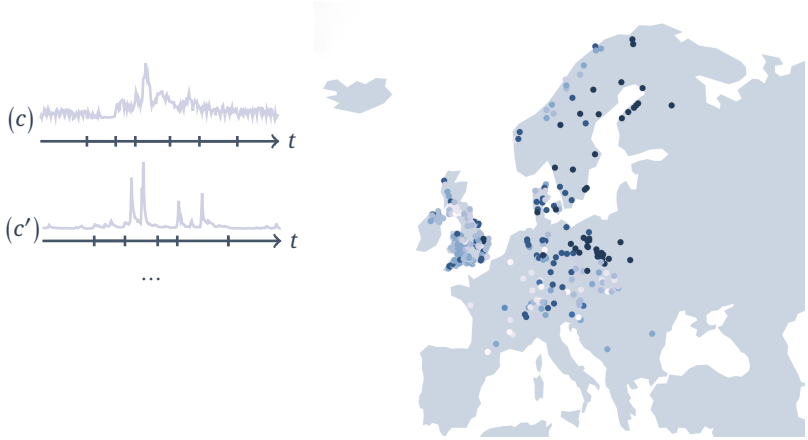


Figure 6.1 The dataset by Cornes et al. (2018) ranges over European catchments, sites where water drains into an outlet. Specifically, time series data on river discharge spanning several years are available for each location.

question through temporal causal modeling, where causal mechanisms change both across contexts and over time. For causal discovery, we build on the previous score-based approach and justify its use in the temporal setting with unknown changepoints. We also introduce a causal discovery algorithm and conclude with evaluations on real-world case studies.

6.1 Non-Stationary Time Series

Novel to this chapter, we assume that the variables $X = \{X_1, \dots, X_d\}$ are observed over time at indices $t = 1, \dots, n$. In addition, observations again come from multiple contexts $c = 1, \dots, m$.

In each context c, we observe a multivariate time series

$$\mathcal{D}^c = \left(\boldsymbol{x}^c(1), \dots, \boldsymbol{x}^c(n)\right), \qquad \boldsymbol{x}^c(t) \in \mathbb{R}^d,$$

resulting in the multi-context time series $\boldsymbol{\mathcal{D}} = \{\mathcal{D}^1, \dots, \mathcal{D}^m\}$.

We formalize the problem under a temporal causal model.

6.1.1 Temporal Causal Model

We model the system by a temporal causal graph (TCG) (Assaad et al., 2022) over nodes $X^c_{j(t)}$, where each node denotes variable X_j in context c at time t.

Definition 6.1 (Temporal Causal Graph (TCG)). *A Temporal Causal Graph (TCG) $G_{(t)} = (\mathcal{V}_{(t)}, \mathcal{E}_{(t)})$ is a directed acyclic graph (DAG) $G_{(t)}$ where the set $\mathcal{V}_{(t)}$ includes nodes $X^c_{j(t)}$ for each j, time t, and context c, and the edge set $\mathcal{E}_{(t)}$ includes*

- *instantaneous links $X^c_{j(t)} \to X^c_{k(t)}$ whenever X_j causes X_k contemporaneously within context c at time t, where $j \neq k$, and*
- *lag-specific links $X^c_{j(t-\tau)} \to X^c_{k(t)}$ pointing forward in time whenever $X^c_{j(t-\tau)}$ causes $X^c_{k(t)}$ with a time lag of $\tau \in \{1, \ldots, \tau_{max}\}$.*

Whenever clear from context, we write the graph $G_{(t)}$ as G. We use $Pa^c_{j(t)}$ to denote the set of contemporaneous or lagged parents of $X^c_{j(t)}$ in G. Consistent with our previous assumptions, the causal parent set is invariant across contexts c; we therefore suppress the context index for brevity. In addition, we assume that causal edges remain constant in time in the following sense.

Assumption 6.1 (Graph Consistency). *The graph $G_{(t)}$ is invariant across time and contexts, such that whenever $X^c_{j(t)} \to X^c_{k(t')} \in \mathcal{E}_{(t)}$, we also have*

$$X^{c'}_{j(t+\delta)} \to X^{c'}_{k(t'+\delta)} \in \mathcal{E}_{(t)}$$

for all time shifts $\delta \in \mathbb{Z} \geq 0$ and contexts c' for which the shifted nodes are defined.

This property is sometimes referred to as repeating edge property. In practice, we further assume that causal effects occur only up to a finite maximum lag $\tau_{\max}$. Under Graph Consistency, this allows us to represent $G_{(t)}$ equivalently by a finite window causal graph over time indices $t - \tau_{\max}, \ldots, t$. Whenever convenient, we therefore work with this finite-lag representation rather than the full unrolled graph.

The corresponding summary graph is the directed graph over $\mathcal{X}$ with

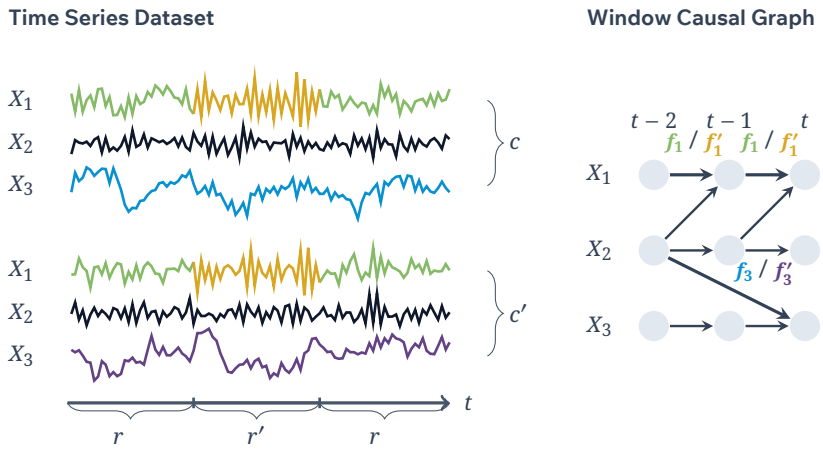


Figure 6.2 Given a non-stationary time series (left) in multiple contexts (top, bottom), we discover changepoints over time (braces), causal mechanism changes and repeating mechanisms (colors) and a temporal causal graph (right).

an edge (X_j, X_k) whenever there exists a lag $\tau \in \{0, \dots, \tau_{\max}\}$ such that

$$X^c_{j(t-\tau)} \to X^c_{k(t)} \in \mathcal{E}_{(t)} \ .$$

We assume that this summary graph is acyclic. While TCGs allow for cycles in the summary graph if lags are present, we restrict our focus to the acyclic case to maintain consistency with score-based DAG search.

While the graph remains fixed, we allow changes in causal mechanisms over both contexts and time. The latter changes in time occur at $n_{\mathcal{T}}$ changepoints

$$\mathcal{T} = \{1 = \tau_0 < \tau_1 < \cdots < \tau_{n_{\mathcal{T}}} < \tau_{n_{\mathcal{T}}+1} = n + 1\} \ .$$

Accordingly, we partition the time domain into $n_{\mathcal{T}} + 1$ segments, called regimes, each given by $I_r = \{\tau_{r-1}, \dots, \tau_r - 1\}$, where I_r denotes the set of time indices belonging to regime r. We assume global changepoints and regimes across all contexts, to model seasonality or external forces such as climate which affects an entire region. As a final step, we can define our structural causal model in similar fashion as in Chapter 3, where mechanism selections occur for each context and each time segment. For

each variable X_j, context c, and regime r, we introduce

$$\Pi_j \in \{1, \ldots, k_j\}$$

to denote a mechanism-selection variable, with realizations $\pi_j^{(c,r)}$ indicating the causal mechanism choice in the cth context and rth regime. Since mechanism assignments are defined relative to the segmentation induced by the changepoints, they implicitly depend on $\mathcal{T}$. We omit this dependence in the notation whenever clear from context. We assume that the same set of mechanisms $\{f_j^1, \ldots, f_j^{k_j}\}$ is shared across all contexts and segments. Then, the structural equation is

$$X_{j(t)}^c = f_j^k\left(Pa_{j(t)}, N_{j(t)}\right) \tag{6.1}$$

where $k = \pi_j^{(c,r)}$, for all $t \in I_r$. We assume that the noise variables $N_{j(t)}$ are jointly independent. Thus, mechanisms change across context-regime pairs but remain constant within each pair.

Given the conceptual similarity of this model to the previous multi-context setting, it is natural to ask whether the same score-based principle can recover not only the graph and mechanism assignments, but now also the unknown changepoints.

6.1.2 Scoring Criterion

We extend the score to the temporal setting by pooling observations not only across contexts, but also across the temporal regimes induced by the changepoints. For each context c, we observe the time series

$$\mathcal{D}^c = \left(\boldsymbol{x}^c(1), \ldots, \boldsymbol{x}^c(n)\right), \qquad \boldsymbol{x}^c(t) \in \mathbb{R}^d.$$

Given variable X_j and mechanism index k, let

$$\boldsymbol{x}_j^{(k)} = \left\{ x_j^c(t) \,\middle|\, \exists r \text{ such that } t \in I_r,\ \pi_j^{(c,r)} = k \right\},$$

$$\boldsymbol{pa}_j^{(k)} = \left\{ \boldsymbol{pa}_j^c(t) \,\middle|\, \exists r \text{ such that } t \in I_r,\ \pi_j^{(c,r)} = k \right\},$$

for $c = 1, \ldots, m$ and all time points t for which the parent values are defined.

These denote the pooled observations of X_j and its temporal parents from all context-regime pairs (c, r) such that mechanism k is active.

The corresponding local MDL score is defined analogously to Eq. (3.2), but using the pooled temporal observations $\boldsymbol{x}_j^{(k)}$ and parent observations $\boldsymbol{pa}_j^{(k)}$.

The overall MDL score is

$$L\left(\mathcal{D}; G, \mathcal{T}, \Pi\right) = \sum_{j=1}^{d} \sum_{k=1}^{k_j} L\left(\boldsymbol{x}_j^{(k)} \mid \boldsymbol{pa}_j^{(k)} ; \mathcal{H}\right) . \tag{6.2}$$

Thus, as in the multi-context setting, we pool all observations sharing the same mechanism. In particular, recurring mechanisms across different contexts and time windows are pooled together and encoded only once, favoring models in which the same causal mechanisms persist or recur across time.

In the temporal case, this relies on the assumption that mechanisms remain constant within each regime and that each mechanism is observed sufficiently often, which we formalize in the next section. An ensuing question is whether similar identification guarantees apply in this setting.

6.2 Score-based TCG and Changepoint Discovery

To address this question, we next formulate the assumptions under which the score is consistent in the temporal setting. Relative to Chapter 3, the new ingredient is that the changepoints $\mathcal{T}$ are now unknown and must be recovered jointly with the graph and mechanism assignments.

We adapt the assumptions from Chapter 3 to the temporal setting (Appendix F), replacing contexts by context-regime pairs. In particular, we assume the temporal analogues of the causal Markov property, faithfulness, sufficient capacity, independent change, sparse change, and change faithfulness. We again assume that the functional model is well specified.

Assumption 6.2 (Additive Noise Model)**.** *For each X_j we assume that the structural equation model is a temporal additive noise model according to Eq. (6.1).*

Under the ANM, we add a coverage assumption to ensure that each

true mechanism and each true regime receives asymptotically many observations.

Assumption 6.3 (Sufficient Regime Coverage). *Each true mechanism receives asymptotically many observations, i.e., for all j and k,*

$$\left|\left\{(c, r, t) \mid t \in I_r,\ \pi_j^{(c,r)} = k\right\}\right| \to \infty \quad \text{as } m, n \to \infty,$$

and each true regime I_r satisfies $|I_r| \to \infty$.

Thus, as $m, n \to \infty$, the pooled sample size for every true mechanism tends to infinity. This lets us give a consistency guarantee of the score.

Theorem 6.1 (Consistency). *Let the assumptions from Section 6.2 hold together with the temporal analogues of the assumptions from Chapter 3. Then for any minimizers $(G, \mathcal{T}, \Pi)$ of Eq. (6.2),*

$$Pr(G = G^*,\ \mathcal{T} = \mathcal{T}^*,\ \Pi = \Pi^*) \to 1 \qquad \text{as } m \to \infty \text{ and } n \to \infty.$$

Proof Sketch. For a fixed segmentation $\mathcal{T}$, the score in Eq. (6.2) has the same structure as the multi-context score in Chapter 3, with context-regime pairs taking the role of contexts. Hence the same score-consistency arguments as in Theorem 3.1 and Theorem 3.2 imply that, for fixed $\mathcal{T}$, the true temporal graph G^* and mechanism assignments Π^* uniquely minimize the score asymptotically.

The new ingredient is the optimization over changepoints. If a candidate segmentation $\mathcal{T}$ omits a true changepoint, then at least one estimated regime merges observations generated by two distinct true mechanisms. By Change Faithfulness, the corresponding conditionals differ, so fitting a single pooled mechanism incurs a strictly positive asymptotic fit penalty.

Conversely, if $\mathcal{T}$ introduces a spurious changepoint, then it splits a homogeneous true regime. In this case, the asymptotic fit does not improve, but the description length increases because the model introduces unnecessary mechanism distinctions.

Therefore, the true changepoints $\mathcal{T}^*$ uniquely minimize the score among all segmentations. Combining this with the fixed-segmentation argument yields

$$\Pr(G = G^*,\ \mathcal{T} = \mathcal{T}^*,\ \Pi = \Pi^*) \to 1$$

as $m, n \to \infty$. □

The result builds on the fact that observations sharing the same mechanism can be pooled across context-regime pairs to estimate a common conditional distribution. This is justified by the temporal model together with the assumption that each mechanism receives enough samples (Assumption 6.3).

The main observation is that the score performs changepoint detection implicitly, where changepoints are introduced exactly when they improve the pooled encoding of the data by separating distinct causal mechanisms. Thus, an exhaustive search over all segmentations would recover the true model asymptotically. In practice, however, such an exhaustive search is clearly computationally infeasible, which motivates the iterative algorithm we introduce next.

6.3 Algorithm

To conclude the framework, we introduce SPACETIME for causal discovery jointly over contexts and time. The objective is to minimize the score $L(\mathcal{D}; G, \mathcal{T}, \Pi)$ in Eq. (6.2) over three components, the temporal graph $G_{(\tau_{\max})}$, the changepoints $\mathcal{T}$, and the mechanism assignments Π. The components depend on each other, as changepoint detection requires a graph to compute residuals, whereas graph search and mechanism assignment require a regime segmentation induced by $\mathcal{T}$. We therefore optimize them through an iterative procedure.

Causal Changepoint Detection. Unlike contexts, which are typically predefined, changepoints $\mathcal{T}$ are unknown in our setting. Crucially, we focus on detecting causal changepoints, i.e., shifts in the conditional distributions of an effect given its causes, rather than shifts in marginal distributions.

For a candidate graph G, we fit GP regression models for each variable X_j on its current parent set and compute residuals

$$\epsilon_{j,c,t} = X^c_{j(t)} - \hat{f}_j\big(Pa^c_{j(t)}\big) \ .$$

Under a correct graph and segmentation, these residuals should be approximately stable within each regime and change at a changepoint. We then stack them across variables and contexts into a multivariate residual vector $\boldsymbol{\varepsilon}_t$ and use this to detect global changepoints.

To this end, we apply the Pruned Exact Linear Time (PELT) algorithm with an RBF kernel. The kernelized formulation enables detection of

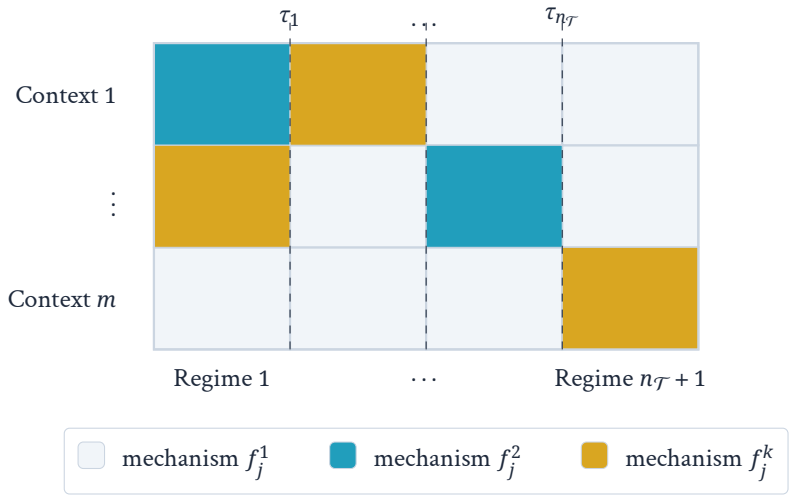


Figure 6.3 Illustration of the mechanism assignment step over contexts and regimes under changepoints $\mathcal{T} = \{\tau_1, \ldots, \tau_{n_\mathcal{T}}\}$. Context-regime pairs assigned to the same color share the same mechanism.

general non-linear changes in the residual distribution without restrictive parametric assumptions. Through its penalized objective, PELT also selects the number of changepoints automatically.

Mechanism Assignment. Once the timeline is segmented into regimes $\mathcal{T}$, we infer the latent mechanism assignments Π. This involves grouping context-regime pairs (c, r) that share the same generating process. For each variable X_j, we perform pairwise similarity tests between all context-regime pairs using Kernel-based Conditional Independence (KCI) tests (Zhang et al., 2011). These tests evaluate whether the conditional distributions $P\left(X_j \mid Pa_{j(t)}\right)$ are identical across pairs of context-regime assignments. We then apply connected component analysis to the resulting similarity matrix to cluster context-regime pairs (c, r) into discrete mechanism groups. We depict the idea in Figure 6.3.

Temporal Graph Search. With $\mathcal{T}$ and Π fixed, the problem reduces to finding the graph $G_{(\tau_{\max})}$ that minimizes the score. The search space includes both the contemporaneous parents and lagged parents with

Algorithm 6.1: SPACETIME

input : Multivariate time series $\mathcal{D} = \{\mathcal{D}^1, \dots, \mathcal{D}^m\}$
minimum changepoint distance τ_w,
maximum time lag $\tau_{\max}$, max. number of iterations I
output: window causal graph $G_{(\tau_{\max})}$, changepoints $\mathcal{T}$,
mechanisms Π

1 Initialize $G_{(\tau_{\max})} = (\mathcal{V}_{(\tau_{\max})}, \mathcal{E}_{(\tau_{\max})})$ with $\mathcal{E}_{(\tau_{\max})} = \emptyset$
2 **while** not converged and iteration I not reached **do**
3 Infer $\mathcal{T} = \text{CausalChangepointDetection}(\mathcal{D}, G_{(\tau_{\max})}, \Pi; \tau_w)$
4 Infer $\Pi = \text{MechanismAssignment}(\mathcal{D}, \mathcal{T}, G_{(\tau_{\max})})$
5 Update $G_{(\tau_{\max})} = \text{TemporalGraphSearch}(\mathcal{D}, \mathcal{T}, \Pi; \tau_{\max})$
6 **return** causal model $G_{(\tau_{\max})}, \mathcal{T}, \Pi$

$\tau \in \{1, \dots, \tau_{\max}\}$.

We instantiate this step using a greedy score-based search. Starting from an initial graph, we iteratively add edges that maximize the reduction in description length. For a candidate edge, we define its directed score gain as the decrease in the objective when adding that edge to the current graph, with the remaining components evaluated under the current iteration. Thus,

$$\Delta_{X^c_{j(t)} \to X^c_{k(t')}} = L(\mathcal{D}; G, \mathcal{T}, \Pi) - L(\mathcal{D}; G', \mathcal{T}, \Pi) \,,$$

where G' differs from G only by the addition of the edge. After a forward phase that adds edges, we perform a backward pruning phase to remove redundant edges and refine parent sets.

SpaceTime. We summarize the procedure in Algorithm 6.1. Each iteration consists of detecting changepoints $\mathcal{T}$ from changes in the residual distributions under the current graph (l. 3), estimating mechanism assignments Π given the current segmentation (l. 4), and then updating the temporal graph $G_{(\tau_{\max})}$ by minimizing the score $L(\mathcal{D}; G, \mathcal{T}, \Pi)$ (l. 5).

These steps are repeated until convergence, defined as no change in the changepoints $\mathcal{T}$ and the graph structure G between successive iterations, or until a maximum number of iterations is reached.

6.4 Related Work

Causal discovery in time series has traditionally been dominated by Granger-causality (Tjøstheim, 1981), which focuses on predicting whether past values of one variable can help forecast future values of another, hence often conflating prediction with causation. Our work aligns more closely with methods seeking structural relationships through functional models.

Stationary Methods. Constraint-based methods like PCMCI+ (Runge, 2020) extend the PC algorithm to time series using momentary conditional independence. Score-based approaches include TiMINo (Peters et al., 2013) which extends results for restricted functional models from the continuous setting to time series. The LiNGAM and NOTEARS algorithms, respectively, have the time series equivalents VARLiNGAM (Hyvärinen et al., 2010) and DYNOTEARS (Pamfil et al., 2020). A common limitation of these approaches is the assumption of temporal stationarity and the restriction to a single time series dataset.

Non-Stationarity. Recent work has begun addressing non-stationarity, studied from different perspectives. CD-NOD (Huang et al., 2020) incorporates time as an exogenous surrogate variable to model smooth shifts. While Gao et al. (2025) also explore causal changepoint detection, their framework is restricted to the discrete setting, whereas SPACETIME handles continuous variables via GP regression. The approach R-PCMCI (Saggioro et al., 2020) and the concurrent work CASTOR (Rahmani & Frossard, 2025), like us, assume the existence of recurrent regimes within which the data is stationary. PCMCI$_\Omega$ (Gao et al., 2024) only considers semi-stationary time series where the causal effects occur periodically. Unlike the aforementioned approaches, J-PCMCI+ (Günther et al., 2023b) can handle multiple heterogeneous time series datasets as well as non-stationarity, but does not perform regime detection or context partitioning. SPACETIME unifies these perspectives by treating time and context symmetrically.

6.5 Experiments

In this section, we evaluate our approach on synthetic data and on real-world datasets in hydrology and meteorology. In the experiments, we evaluate the synergy between the modular components of SPACETIME.

Specifically, we assess

(*i*) *Causal Discovery*: Can we recover the TCG across time and contexts?

(*ii*) *Changepoint Detection*: Can we detect temporal changepoints?

(*iii*) *Regime Partitioning*: Can we discover repeating temporal regimes with similar causal mechanisms?

6.5.1 Experimental Setup

To simulate time series datasets, we generate WCGs G with maximum time lag $\tau_{\max} = 2$, ensuring that the summary DAG G, excluding autocorrelations, is acyclic. We sample regime changepoints $\mathcal{T}$ uniformly at random using a pre-set minimal duration, by default $\tau_w = 30$. In each context and regime, we re-sample the edge weights of a fraction of all edges, and finally sample data using the same data generators as in previous works (Günther et al., 2023b), here with nonlinear functional form and Gaussian noise.

We compare SPACETIME primarily against J-PCMCI+ (Günther et al., 2023b), which handles multi-context data, and R-PCMCI (Saggioro et al., 2020), which focuses on changepoint detection, while our method infers both; we also consider CD-NOD by providing the time index as the context variable (Huang et al., 2020), but note that it only discovers an untimed DAG; as well as include a choice of temporal causal discovery baselines, here PCMCI+(Runge, 2020), VARLiNGAM (Hyvärinen et al., 2010) and DYNOTEARS (Pamfil et al., 2020).

6.5.2 Synthetic Data

Figure 6.4 shows the result on synthetic data, reporting F1 scores over $\mathcal{T}$ (top) for changepoint detection, F1 scores over directed edges in the summary DAG G (middle) and the temporal $G_{(t)}$ (bottom) for causal discovery, and the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) for regime partitioning, that is, assignment of each time window to correct causal mechanism. For comparison, we also report changepoint detection under a known G resp. causal discovery under a known $\mathcal{T}$ (hatched); for the remaining causal discovery methods,

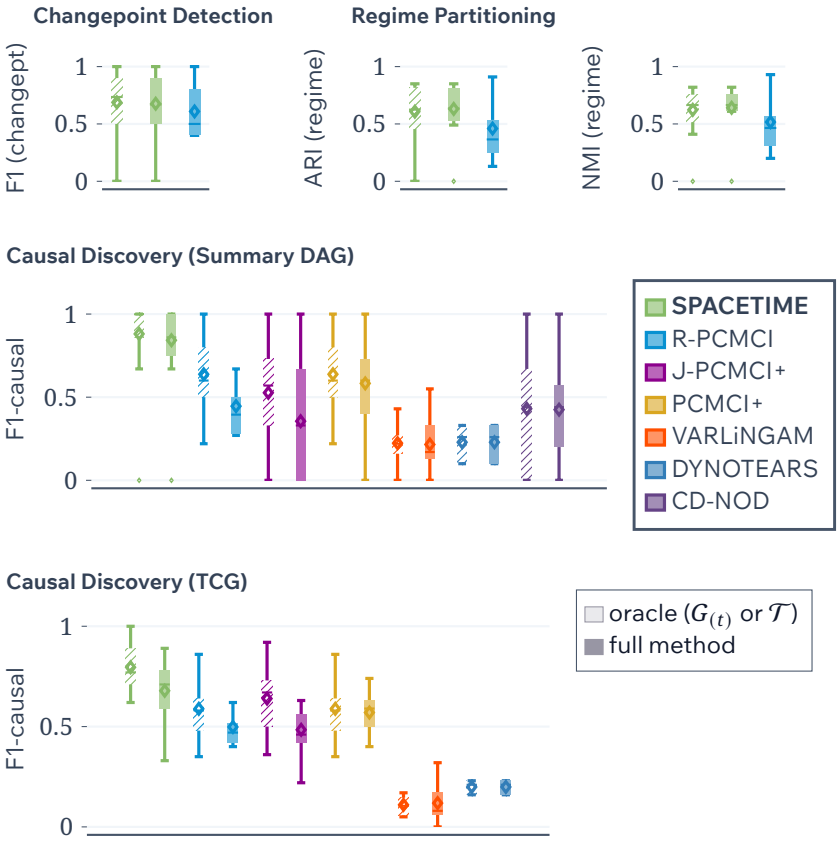


Figure 6.4 To demonstrate causal discovery, changepoint detection and regime partitioning in synthetic data, we evaluate the methods on multiple time series datasets with causal mechanism shifts across time and contexts, with nonlinear functional form, Gaussian noise, and parameters $d = 5, n = 200, m = 2, n_{\mathcal{T}} = 2$.

we also include a version where the correct changepoints $\mathcal{T}$ were given (hatched) alongside the full method (solid).

Causal Discovery. Regarding causal discovery, the methods DYNOTEARS and VARLiNGAM make linearity assumptions and the latter in addition assumes that the noises in the model are non-Gaussian, therefore it is

not surprising that they show sub-par performance here. Even though they as well as PCMCI+ assume stationarity, i.e. the absence of regime shifts, the effect of known (hatched) compared to unknown changepoints (solid) is not overly pronounced. This could be due to a potential sample-size advantage gained by pooling data across regimes, which may mask the violations of stationarity for these baseline methods, as well as that changes in edge strength are not as drastic as changes in the graph. While CD-NOD can address temporal shifts, it only discovers a summary DAG G. As all aforementioned methods assume a single time series, we favor the methods by providing the correct dataset indices. Only J-PCMCI+ can consider multiple contexts, but does not achieve a significant performance gap here. This may be due to mechanism shifts being too subtle, given that J-PCMCI+ relies on changing graphs across contexts, or because it may need a larger number of different contexts to effectively outperform single-context methods, in which case conditional independence tests however become increasingly expensive due to a high dimensional one-hot encoding.

Changepoint and Regime Identification. In the changepoint detection and partitioning experiments (Figure 6.4, top), we can see that our method is robust to a small number of mistakes in the causal graph as the full results (solid) are close to the oracle version (hatched). This also matches our observation that the interleaving steps in Alg. Algorithm 6.1 converge quickly, typically within three iterations. Reasons that we outperform R-PCMCI (Figure 6.4) may include our explicit modeling of regime persistence, whereas R-PCMCI requires the number of regime switches as a prior constraint, expecting a given number of regime-switches which we here provide as background knowledge.

Overall, SPACETIME reliably recovers *i)* graphical structure, *ii)* changepoints and *ii)* repeating regimes.

6.5.3 Case Study Revisited: Identifying Drivers of River Discharge

Returning to the real-world motivating example, our first real-world case study investigates the effect of meteorological drivers on river discharge over different gauged catchments across Europe, given data derived from the Global Runoff Data Centre (GRDC) datasets (Cornes et al., 2018). It includes multiple datasets from diverse geographical locations, as shown in Figure 6.1, which we pre-process as in Günther et al. (2023a) to obtain

$m = 307$ locations, where the respective time series with daily time resolution span multiple years. We set the time span to a duration of one year (2006) during DAG discovery given that the available time spans for different locations are non-overlapping. We provide all m datasets to our method, and use for simplicity fixed hyperparameters $\tau_w = 30$, $\tau_{\max} = 2$ for our method.

Causal Discovery over Catchments and Time. To explore the effects of meteorological variables on river discharge, we investigate the interactions between temperature (T), precipitation (P), and river runoff (Q). While a previous study of this data using PCMCI+ was limited in considering each location and month separately (Günther et al., 2023a), we are interested in a unified analysis of all catchments. As we show in Figure 6.5, we identify a lagged causal influence $P_{(t-1)} \to Q_{(t)}$, which aligns with the physical expectation of the hydrological response time within a catchment.

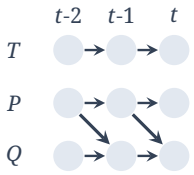


Figure 6.5 The WCG over temperature T, precipitation P, and river runoff Q.

Changepoints and Regimes over Time. To illustrate the changepoint detection, Figure 6.6, left shows the changepoints that our method returns for selected catchments c_1 (CH), c_2 (PL), and c_3 (GB), along with temporal regimes with similar causal relationship shown by color.

Causal Edge Strength across Catchments. Besides the common causal graph and variation across time, we are also interested in variation over space, that is, in how the causal mechanism of $P_{(t-1)} \to Q_{(t)}$ differs among catchment sites.

To better illustrate the regional differences, we consider the causal strength of this edge, as defined by the directional edge gain $\delta = \delta_{P_{(t-1)} \to Q_{(t)}}$ in Eq. (4.4). This gain is conceptually similar to the measures of causal arrow strength given by Janzing et al. (2012), who propose to quantify causal influences through a hypothetical intervention that modifies the joint distribution by removing the corresponding edge; here, we consider the difference in MDL scores with and without the edge. Günther et al. (2023a) also follow a similar approach to compare causal edge strengths in this dataset.

We visualize $\delta = \delta_{P_{(t-1)} \to Q_{(t)}}$ for each of the 307 locations in the year 2010 in Figure 6.6, right. To better illustrate, we applied a simple clus-

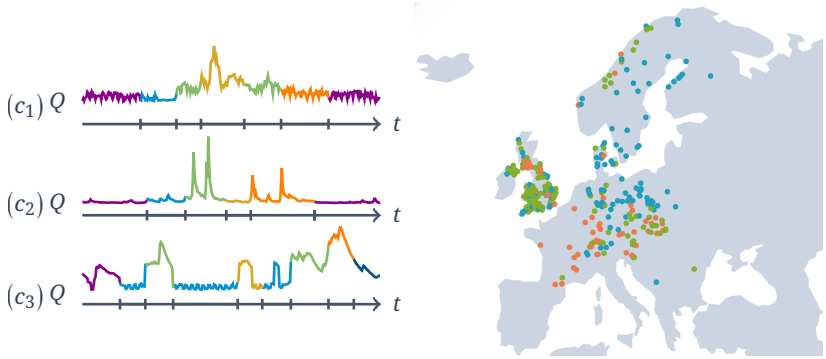


Figure 6.6 For the causal relationship $P \to Q$, we show the regime change-points $\mathcal{T}$ at example locations c_1 (Martinsbruck, CH), c_2 (Nowy Sacz, PL), and c_3 (Linnbrane, GB), where colors indicate regimes with similar causal mechanism. We discover these for each European catchment shown on the right, where we also show the causal edge strength $\delta = \delta_{P_{(t-1)} \to Q_{(t)}}$ per location by color (blue is $\bar{\delta} = -1.9$, green $\bar{\delta} = 1.3$, and orange $\bar{\delta} = 4.8$).

tering procedure over the directional edge strengths δ at each site and show three clusters of similar mean δ in color. Certain regional similarities exist, but also variation especially in central Europe, likely due to differences in topology and climate as well as distinct characteristics of each catchment such as area, altitude, and vegetation, as more closely examined in the study by Günther et al. (2023a).

6.5.4 Reconstructing Gradients of Biosphere-Atmosphere Interactions

Secondly, we consider biosphere–atmosphere fluxes from the FLUXNET database (Baldocchi, 2014). Here, we study interactions of meteorological and atmospheric variables, with primary focus on air temperature (T), global radiation (R_g), net ecosystem exchange (NEE), sensible heat (H), latent heat flux (LE) and vapour pressure deficit (VPD), over $m = 63$ locations and multiple years.

Distinct Trends over Time. With our approach, we test for statistical differences in causal mechanism, and, for example, can observe how the

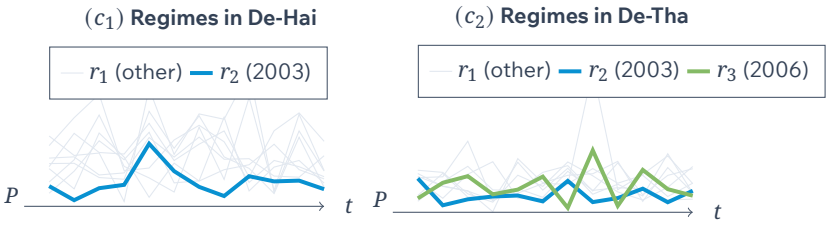


Figure 6.7 The regime partitions that our approach discovers reveal abnormal conditions, such as the effects of a European heatwave (2003) on the FLUXNET ecosystems Hainich (De-Hai) and Tharandt (De-Tha) (2000-2009).

causal mechanisms at a given location differ across years. To illustrate, we show in Figure 6.7 for the locations Hainich (DeHai) and Tharandth (DeThai) the regime partitions over multiple years (here 2000-2009). We discover a common causal mechanism in most years (greyed out) and distinct regimes in 2003, resp. 2006 (colored). These regime shifts correspond to documented extreme meteorological events, most notably the 2003 European heatwave and subsequent drought, which we see reflected in the monthly precipitation P for 2003 (Figure 6.7).

Causal Strengths across Datasets. Choosing the year of 2010 for the analysis, we discover a causal graph jointly over all locations and months (Figure 6.8).

As there is no known ground truth causal graph, and given the vast number of locations and months, we follow the qualitative study by Krich et al. (2021) to interpret the results. The aim is to see whether the causal strengths across locations and time correspond to plausible interactions between the atmospheric and meteorological variables. To this end, we similarly as before consider the directional edge gain δ for each of the seven edges in G. Collecting these for each location c and each month m, we project the resulting vectors to two dimensions using

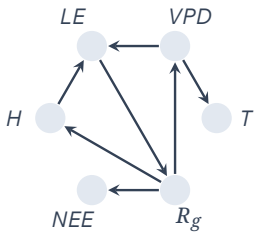


Figure 6.8 The DAG over atmospheric variables in the Fluxnet data (Baldocchi, 2014).

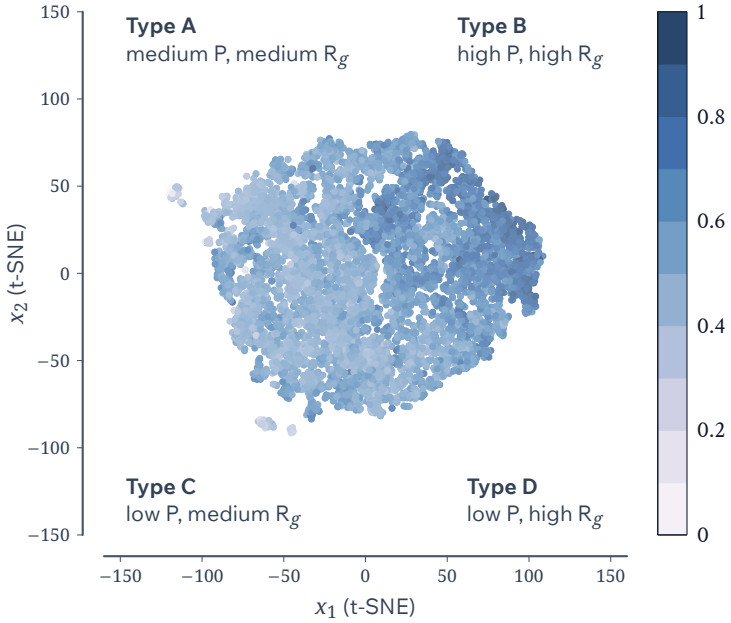


Figure 6.9 Given atmospheric variables observed in different ecosystems over time in the FLUXNET database (Baldocchi, 2014) and the discovered DAG G (Figure 6.8), we show the causal edge strengths in G per each month m and location c, projected to two dimensions using t-SNE. Thus, we can see where different ecosystems at different times of year (m, c) are positioned in terms of causal interaction strengths in the biosphere and atmosphere (x_1, x_2). The resulting four regions match different productivity states of the ecosystems at different times of year reported in a previous study (Krich et al., 2021).

the t-distributed stochastic neighbor embedding (t-SNE).

The resulting embedding is shown in Figure 6.9. Each sample point (x_1, x_2) represents a specific location c during a month m, where the position in the embedding space is jointly determined by all δ. To give further insight, we consider four regions of this space (Type A-D) obtained by optical clustering over the t-SNE space. We also inspected the actual geographical location of each ecosystem, as well as the respective distributions over the system variables. To illustrate the latter, we show

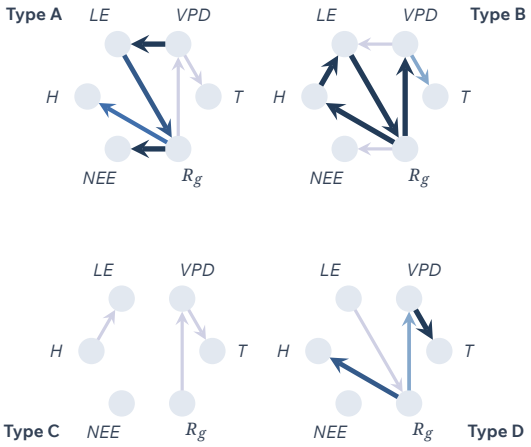


Figure 6.10 For the t-SNE dimensions shown in Figure 6.9, we apply clustering to identify four main regions (Type A-Type D) and for each show a representative causal DAG. In each DAG, we color edges $X \to Y$ by average strength, that is, the mean directed edge gain $\delta_{X \to Y}$ across the locations and months in the respective region (dark blue for strongest, gray for weaker, omitted if very weak).

precipitation P and radiation R_g and annotate each region accordingly. Finally, we show four representative causal graphs (Type A-D) in Figure 6.10, obtained by considering the mean edge strengths per region. For readability, we show edges with high strength in dark blue, and prune those that do not exceed a threshold.

While the t-SNE projection (Figure 6.9) exhibits a relatively continuous distribution rather than isolated clusters, the emergent gradients align closely with environmental drivers such as radiation and water availability. For example, strong edge connections in G (Type B) are associated to high radiation R_g and high precipitation P, which corresponds to optimal growing conditions in the respective ecosystems and months (cf. Krich et al., 2021). In contrast, weak edge connections in G (Type C) correspond to a low-energy output state that is, typical, for example, for winter months. In particular, the grouping by causal strength is not strongly related to the actual geographic location of the ecosystems, but

rather correlates with water availability and global radiation. The findings support the hypothesis that different ecosystems can traverse very similar meteorological states over time (Kraemer et al., 2020).

6.6 Conclusion

In this chapter, we consider time series from multiple contexts, where an additional challenge is detecting unknown changepoints in causal mechanisms over time, besides across datasets. To this end, we adapt the score-based approaches from Chapters 3-4 to the time series setting using autoregression with GPs and MDL regularization, as well as error-based changepoint detection.

Given the modular setup of the method, different choices can be made for each component of the algorithm, both for the score-based graph search (for example, using the topological-ordering based search from Chapter 4), the causal mechanism shift search (for example, using MDL score gains instead of KCI p-values) and finally the changepoint detection technique (such as using a different changepoint detection algorithm). For ease of exposition, we restrict ourselves to the implementation choices given in the original work (Mameche et al., 2025a).

A primary limitation of the current framework is the assumption of global changepoints across all variables and contexts. While this is meant to model seasonal or other global patterns, one could also consider settings where the temporal axes per context are not aligned.

Our algorithm also applies interleaved search in an EM fashion, which is not guaranteed to converge to an optimum, and is capped at a maximum number of iterations. Future extensions could integrate changepoint detection directly into the structure search, where we detect changepoints locally for each candidate parent set. This only requires a single search iteration to discover the TCG and changepoints; However, it in turn loses the benefit of applying the changepoint detection technique over all variables globally, which increases the effective sample size used in PELT. We provide an implementation for this hybrid approach in our Python code to encourage future benchmarking.

Software. To allow the reader to use the algorithm, a Python implementation `github.com/srhmm/causalchange` is available. We include Jupyter notebooks `demo/ch5_time.ipynb` as well as `ch6_realworld_flux.ipynb` and `ch6_realworld_river.ipynb` to explore the real-world data.

7 | Causal Discovery from Mixtures of Populations

Case Study Subpopulation Paradox

The Cancer Genome Atlas (TCGA)-COAD dataset (Cancer Genome Atlas Research Network et al., 2013) contains multi-omics data, such as RNA expression and DNA methylation, from patients suffering from colon adenocarcinoma (COAD), a common type of colorectal cancer. For example, Figure 7.1 shows methylation (X) and gene expression (Y) for COAD patients (Chang et al., 2020). The relationship between methylation and gene expression is not uniform across all patients, but rather shows two different methylation-expression mechanisms (colors).

While all previous chapters assume multi-context data, sometimes a single given dataset can be a combination of different population subgroups or disease subtypes.

In the case of colon adenocarcinoma, the disease is known to be heterogeneous and to show distinct molecular characteristics in different patients (e.g., Fennell et al., 2019). Importantly, such disease subtypes are not always known a priori to the domain expert, therefore we do not assume multi-context data as before, but rather a single dataset composed of multiple subgroups with differences in causal mechanisms. As patients with different subtypes of a disease may vary in prognosis, disease progres-

This chapter is based on work published as Mameche et al. (2025b).

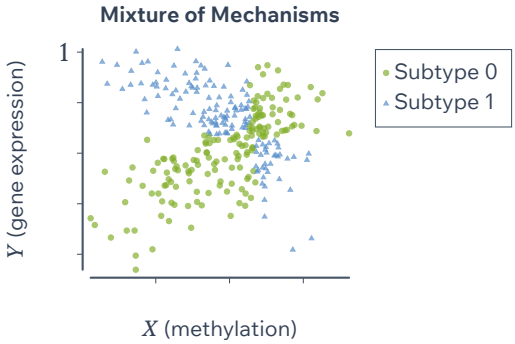


Figure 7.1 As an example of a mixture of causal mechanisms, we show methylation X (at site cg16012690) and gene expression Y (of gene CREB3L1) for patients from the Cancer Genome Atlas (TCGA) colon adenocarcinoma cohort. Patients show a heterogeneous methylation-expression profile (color). In contrast to the previous chapters, the assignment to these different subgroups is here unknown. *Data from Chang et al. (2020), based on data generated by the TCGA Research Network:* `https://www.cancer.gov/tcga`.

sion, and treatment response, we want to avoid treating the subgroups as a single population.

To address this problem, this chapter studies a mixtures-of-populations setting, where a given dataset is a combination of unknown subpopulations with differences in their causal mechanisms. After introducing this problem setting under a causal model (Section 7.1), we show consistency of score-based learning in this setting (Section 7.2). We also introduce a causal discovery algorithm capable of identifying the subpopulations (Section 7.3) and evaluate it empirically (Section 7.5).

7.1 Non-i.i.d. Data from Mixtures of Populations

Given continuous random variables $\mathcal{X} = \{X_1, \ldots, X_d\}$, we are now interested in unknown changes of their causal mechanisms. For this, we allow the coefficients of each X_j to depend on one of k_i unknown subgroups, represented through the values of a discrete latent variable Z_i. We assume the following structural equations,

$$X_j = f_j(Pa_j, La_j) + N_j \ . \tag{7.1}$$

where Pa_j are the observed causes, $La_j = Z_i$ is a latent variable, and N_j is independent noise.

In the motivating example (Figure 7.1), a latent variable Z influences $Y \mid X$ corresponding to the group assignment shown in color. To go beyond this bivariate case, we will consider a set of *multiple* latent variables, each affecting different observed variable sets.

We assume a set of unobserved random variables $\mathcal{Z} = \{Z_1, \dots, Z_{d_z}\}$, with $d_z \leq d$, each following a categorical distribution $Z_i \sim \text{Categorical}(\boldsymbol{\gamma}^i)$ with $Z_i \in \{1, \dots, k_i\}$. Each observed variable X_j is affected by at most one latent variable Z_i. When such a latent variable exists, we denote it by $La_j = Z_i$. Conversely, each latent variable may affect multiple observed variables.

The latent variables $\mathcal{Z}$ play a role similar to the variables Π introduced in previous chapters, but here contexts are not observed, and $\mathcal{Z}$ must be inferred jointly with the causal graph.

To address this setting, we simplify the setting here to parametric functional models. That is, we assume that f are parameterized by the collection of linear coefficient vectors $\boldsymbol{B}_j = \{\boldsymbol{\beta}_{j1}, \dots, \boldsymbol{\beta}_{jk_i}\}$, each of which has dimension equal to the number of parents of X_j, i.e., $\boldsymbol{\beta}_{jk} \in \mathbb{R}^{|Pa_j|}$ for all $k \in \{1, \dots, k_i\}$, and that noise $N_j \perp\!\!\!\perp Pa_j$ is additive Gaussian $N_j \sim \mathcal{N}(0, \sigma^2)$.

Then for each X_j with $La_j = Z_i$ we get

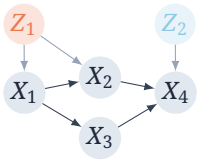
$$\begin{aligned} (X_j \mid Pa_j = y, La_j = k) &\sim \mathcal{N}\left(\boldsymbol{\beta}_{jk}{}^\top y, \sigma^2\right) && \text{for } k \in \{1, \dots, k_i\} \text{ and} \\ (X_j \mid Pa_j = y) &\sim \text{MLR}\left(\boldsymbol{B}_j, \boldsymbol{\gamma}^j, \sigma^2\right) \ , \end{aligned}$$

where $\text{MLR}\left(\boldsymbol{B}, \boldsymbol{\gamma}, \sigma^2\right)$ is the conditional distribution of a mixture of linear regressions with density

$$p^{\text{MLR}}(x, y; \boldsymbol{B}, \boldsymbol{\gamma}, \sigma^2) = \sum_{k=1}^{K} \gamma_k p^{\mathcal{N}}(x; \boldsymbol{\beta}_k{}^\top y, \sigma^2) = \sum_{k=1}^{K} \frac{\gamma_k}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\|\boldsymbol{\beta}_k{}^\top y - x\|^2}{2\sigma^2}\right) \ ,$$

and $p^{\mathcal{N}}(x; \mu, \sigma^2)$ is the density of the normal distribution with mean μ and variance σ^2.

We represent the above causal model as a DAG $G_{\mathcal{Z}} = (\mathcal{X} \cup \mathcal{Z}, \mathcal{E}_{\mathcal{Z}})$ over both observed $\mathcal{X}$ and latent $\mathcal{Z}$ random variables, between which we add edges



$X_j \to X_l$ whenever X_j is a cause of X_l, as well as $Z_i \to X_j$ whenever $La_j = Z_i$; since the $\mathcal{Z}$ were assumed exogenous and independent, we allow no incoming edges toward the $\mathcal{Z}$ themselves, plus we assume each latent node has at most one child. Of special interest is the subgraph $G = (\mathcal{X}, \mathcal{E})$ that is induced on $G_\mathcal{Z}$ by the node subset $\mathcal{X}$, with $\mathcal{E} = \{X_j \to X_l \mid X_j \to X_l \in \mathcal{E}_\mathcal{Z}\}$. Considering only the latter subgraph, we denote the set of all observed direct predecessors of X_j in G by $Pa_j \subset \mathcal{X} \setminus \{X_j\}$. Figure 7.2 depicts an example of such a graph $G_\mathcal{Z}$ (colored and black) and the resp. induced subgraph G (black).

We assume the causal Markov Property to hold in the full causal model $G_\mathcal{Z}$, which results in the following factorization over the observed variables.

Assumption 7.1 (Causal Markov Property)**.** *We assume the causal Markov property over $\mathcal{X}$, where*

$$p(\boldsymbol{x}) = \prod_{X_j \in \mathcal{X}} p^{\text{MLR}}(x, \boldsymbol{pa}_j; \boldsymbol{B}, \boldsymbol{\gamma}, \sigma^2) \,, \tag{7.2}$$

We also make the following simplifying assumptions that no causal dependencies among $\mathcal{Z}$ exist, such that the causal Markov property over these variables reads $p_\mathcal{Z}(\boldsymbol{z}) = \prod_{Z_i \in \mathcal{Z}} p(z_i)$.

Assumption 7.2 (Exogeneity of $\mathcal{Z}$)**.** *We assume the latent variables $\mathcal{Z}$ are exogenous, i.e., there are no edges $X_j \to Z_i$ nor edges $Z_i \to Z_l$ for any Z_i, Z_l, X_j.*

To learn this causal model, we show that score-based approaches can be extended to the mixed setting when we combine them with Expectation-Maximisation-based estimation of mixture assignments.

7.2 Score-Based CMM Discovery

We now study learning of the model using score-based causal discovery. As we consider the linear case, rather than the GP score we will now focus on the BIC score. Given a model hypothesis $\mathcal{H}$ assuming distribution $P_\mathcal{X}(\boldsymbol{\theta})$ with parameters $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d$ and observations $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ of $\mathcal{X}$, the score of $\mathcal{H}$ is

$$\mathrm{BIC}(\mathcal{H}) \coloneqq -2 \log p(\boldsymbol{x} \mid \hat{\boldsymbol{\theta}}) + \dim(\boldsymbol{\theta}) \log n\,, \tag{7.3}$$

where the first term is the scaled (log) likelihood of P_X evaluated at its maximiser $\hat{\boldsymbol{\theta}}$.

This score has the important property of decomposability (Chickering, 2002), stating that when P_X factorises as a product the Bayes Information Criterion (BIC) decomposes as a sum

$$p(\boldsymbol{x}) = \prod_{X_j \in \mathcal{X}} p(x_j | \boldsymbol{pa}_j) \qquad \Longrightarrow \qquad \mathrm{BIC}(\mathcal{X}) = \sum_{X_j \in \mathcal{X}} \mathrm{BIC}(X_j \mid Pa_j)\,. \tag{7.4}$$

To transfer these properties to our own model, we need to further account for the latent mixing variables. Here, at each scoring step and according to our model we need to compute the likelihood $p^{\mathrm{MLR}}(x, \boldsymbol{pa}_j; \boldsymbol{B}, \boldsymbol{\gamma}, \sigma^2)$, which requires an estimation of the Mixture of Linear Regressions (MLR) parameters. When an oracle that can compute the maximum likelihood estimates for the MLR model for a given number of mixtures K is available, we write

$$\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{B}}, \hat{\boldsymbol{\gamma}}, \hat{\sigma}) = \underset{(\boldsymbol{B}, \boldsymbol{\gamma}, \sigma) \in \Theta_K}{\arg\max}\ p^{\mathrm{MLR}}(x, y; \boldsymbol{B}, \boldsymbol{\gamma}, \sigma^2)\,, \tag{7.5}$$

where $\Theta_K = (\mathbb{R}^{|Pa_j|})^K \times \mathcal{S}^K \times \mathbb{R}_+$ is the space of the model parameters for K mixtures and $\mathcal{S}^K$ the K-dimensional probability simplex. While the computation of the Maximum Likelihood Estimate (MLE) estimates for both Gaussian Mixture Model (GMM) and MLR models is NP-hard in the general case, there exist efficient algorithms to compute well-performing local maximisers of the corresponding likelihoods is the Expectation Maximisation (EM) framework (Wu, 1983) and its variants (Bishop, 2006).

With these, we can define a latent-aware BIC,

$$\mathrm{BIC}^{\mathrm{ML}}_{\hat{Z}}(\mathcal{H}) = \max_{1 \le K \le k_{\max}} \mathrm{BIC}^{\mathrm{ML}}_{\hat{Z}}(\mathcal{H}, K) \tag{7.6}$$

and similarly $\mathrm{BIC}^{\mathrm{EM}}_{\hat{Z}}$, where $\mathrm{BIC}^{\mathrm{ML}}_{\hat{Z}}$ uses the oracle estimate and $\mathrm{BIC}^{\mathrm{EM}}_{\hat{Z}}$ the EM one. Then, the latent-aware versions inherit the decomposability property in Eq. (7.4), suggesting their integration into a score-based causal discovery algorithm.

Throughout, we assume the availability of the oracle version for our theoretical results.

Assumption 7.3 (Oracle BIC). *We assume the availability of consistent MLE estimates to obtain* $\mathrm{BIC}^{ML}_{\hat{Z}}(\mathcal{H})$.

A central observation is that mixture-of-regression conditionals deviate from the linear Gaussian setting that is known to be non-identifiable. The following lemma formalises this property.

Lemma 7.1 (Non-Gaussianity of Reverse Conditionals)**.** *Let Y and X be random variables such that $Y \mid X \sim \mathrm{MLR}(\boldsymbol{B}, \boldsymbol{\gamma}, \sigma^2)$, with parameters $\boldsymbol{\theta} = (\boldsymbol{B}, \boldsymbol{\gamma}, \sigma^2) \in \cup_{K \geq 2} \Theta_K$. Assume that $\gamma_k \neq 0$ for all k, and that the prior over $\boldsymbol{B}$ is absolutely continuous with full support. Then the conditional distribution $Y \mid X$ is almost surely non-Gaussian.*

Proof Sketch. Fix $K \geq 2$. The mixture distribution reduces to a single Gaussian only if either (i) a single component is active, or (ii) all regression coefficients coincide.

The first case corresponds to the vertices of the simplex S^{K-1}, which form a finite set and thus have Lebesgue measure zero. The second case corresponds to a proper linear subspace of $\mathbb{R}^{d \times K}$, which also has Lebesgue measure zero.

Hence, under an absolutely continuous parameter distribution, both cases occur with probability zero, and the mixture model is almost surely non-Gaussian. □

We now combine this property with the decomposability of the latent-aware score.

Theorem 7.1 (Consistency of Latent-Aware BIC)**.** *Under Assumptions 7.1, 7.2, and 7.3, together with standard regularity conditions, the latent-aware score $\mathrm{BIC}^{ML}_{\hat{Z}}$ is a consistent scoring criterion. Consequently, standard score-based search procedures recover the true Markov equivalence class almost surely as $n \to \infty$.*

Proof Sketch. We argue that the score satisfies the standard conditions for consistent score-based structure learning (formalised in Appendix G).

By Lemma 7.1, the conditional distributions are almost surely non-Gaussian, which removes the classical non-identifiability of linear Gaussian models and ensures that incorrect edge orientations induce a strictly worse likelihood.

Further, $\mathrm{BIC}^{\mathrm{ML}}_{\hat{Z}}$ inherits decomposability, so the score decomposes into local contributions. Any incorrect hypothesis either violates conditional independencies (leading to worse likelihood) or introduces unnecessary parameters (leading to a higher penalty).

Algorithm 7.1: CMM

input : Dataset $\mathcal{D}$ over $\mathcal{X}$
maximum number of mixture components $K_{\max}$
output: causal model $(G, \mathcal{Z})$

Initialize $\mathcal{Z} \leftarrow \emptyset, G_{\mathcal{Z}} \leftarrow \emptyset, G \leftarrow \emptyset, T \leftarrow [\,]$
while not all nodes are ordered **do**
  Infer source node $X_j \leftarrow \text{InferSource}(T, G)$ *// next node*
  Update graph $G \leftarrow \text{EdgeAdditions}(X_j, G)$ *// edge additions*
  Update graph $G \leftarrow \text{EdgePruning}(X_j, G)$ *// edge pruning*
  foreach $k = 1, \ldots, K_{\max}$ **do**
    Fit $(X_j \mid Pa_j = y) \sim \text{MLR}\left(\boldsymbol{B}_j, \boldsymbol{\gamma}^j, \sigma^2\right)$ with k components using EM *// mixture fitting*
  Set Z_j to best assignment via $\text{BIC}^{\text{EM}}_{\hat{Z}} = \max_{1 \le K \le K_{\max}} \text{BIC}^{\text{EM}}_{\hat{Z}}(K)$
  Append X_j to T
Infer $\mathcal{Z}, G_{\mathcal{Z}} \leftarrow \text{InferMixing}(G, \{Z_j\})$ *// latent structure*
return causal model $(G, \mathcal{Z})$

Thus, $\text{BIC}^{\text{ML}}_{\hat{Z}}$ satisfies the standard conditions for consistent score-based structure learning, and is therefore consistent. □

7.3 Algorithm

Here, we outline our algorithm for discovering both the latent discrete variables $\mathcal{Z}$ and the causal graph $G_{\mathcal{Z}}$. Following our theoretical insights, we propose a joint inference procedure that estimates both within a score-based framework. At each scoring step, we fit a MLR under a given candidate parent set. While our theory agrees with score-based causal discovery approaches (Chickering, 2002), our proof-of-concept instantiation mainly uses TOPIC.

We then discover the causal DAG G over the observed $\mathcal{X}$ by embedding the above inference step into TOPIC, which constructs the DAG in topological order as described in Chapter 4. Given a candidate edge $Pa_j \to X_j$, we infer an MLR using Expectation Maximisation (EM) (Bishop, 2006) for each number of components up to a given hyperparameter, $1 \le k \le$

$K_{\max}$. To score the edge, we compute the latent-aware BIC of Eq. (7.6). We also define the score difference g as the difference in BICs before and after adding X, $g(X, Y; G) = \text{BIC}_Z(Y \mid \text{pa}(Y) \cup \{X\}) - \text{BIC}_Z(Y \mid \text{pa}(Y))$.

Following this strategy, we discover a topological order T, DAG G as well as a mixture assignment Z_i for each variable X_j along the topological order. We finally consolidate these. With the reasoning that pairwise estimated Z_i, Z_l exhibit significant overlap when they correspond to the same mixing variable, we measure the similarity of the assignments, here using Adjusted Mutual Information (AMI), and merge variables when the AMI exceeds its expected value, similar to the ideas in Chapter 5.

CMM. We summarize our approach in Algorithm 7.1. The graph discovery steps are similar as in the TOPIC algorithm (l. 3–5). Within the per-variable score computation, we here perform mixture inference using EM (l.6–8). These steps are followed by a local discovery step that discovers $\mathcal{Z}$ from the per-variable results (l.10). The algorithm returns the causal structure over the observed variables along with $\mathcal{Z}$ (l. 11).

7.4 Related Work

Gaussian Mixture Models as well as their conditional counterparts, Mixtures of Linear Gaussian Regressions, are well studied (McLachlan & Peel, 2000; Hennig, 2000). Related to the idea of reinterpreting clustering and mixture modeling under a causal framework, there exists some works that assume a global mixing variable, also termed latent-class confounder (Mazaheri et al., 2023). Some works study for example the bivariate cases (Hu et al., 2018), discrete variables (Mazaheri et al., 2023) and causal inference settings (Kim et al., 2024; Mazaheri et al., 2025a). Most closely related to ours is the work of Kumar et al. (2024) which proposes Mixture-UT-IGSP (MixUTIGSP), combining a GMM with adequate component selection with UT-IGSP. Kumar et al. (2024) give insight into the sample complexity of identifying interventional mixtures, as well as establish identifiability of the iMEC in this setting using the two-stage approach.

7.5 Experiments

We evaluate our approach on two main questions,

(*i*) *Discovering Mixing Structure*: Given a causal graph, can our approach accurately discover the underlying mixing structure, including the number of mixing variables, number of their components, assignments, and sets of targeted observed variables?

(*ii*) *Discovering Causal Structure*: When the causal graph is unknown, can our approach recover the mixing structure as well as the causal graph?

7.5.1 Experimental Setup

We generate data according to our assumptions by drawing Erdős Rènyi DAGs using $X_j|Pa_j \sim \text{MLR}(\boldsymbol{B}_j, \boldsymbol{\gamma}^j, \sigma^2)$ with Gaussian additive noise $N_j \sim \mathcal{N}(0,1)$ and ensuring that the linear coefficients $\boldsymbol{B}_j$ are bounded away from zero, $\boldsymbol{\beta}_{jk} \in [-1,-0.25] \cup [0.25,1]$ and similarly from one another. To avoid issues related to Var-sortability and R^2-sortability (Reisach et al., 2023) we generate an internally-standardized structural causal model (iSCM) (Ormaniec et al., 2025).

The experiments address the effect of several parameters: the number of observed $N_X \in \{5,\dots 20\}$ and latent variables $N_Z \in \{0,\dots,10\}$, number of latent classes $K \in \{2,\dots,5\}$, fraction of observed variables $p_Z \in [0,1]$ affected by mixing, dag density $p_G \in [0,1]$, sample size $S \in \{200,...1000\}$ and a parameter controlling the size of the samples in each group $S_Z \in \{2,\dots,10\}$; for example, for K=2, we uniformly at random draw $\frac{S}{S_Z}$ samples belong class 0, otherwise assign class 1. By default, we show results for $N_X = 10, N_Z = 4, K = 2, p_Z = 0.5, p_G = 0.4, S = 500, S_Z = 5$.

7.5.2 Discovering Mixing Structure

We assess the quality of mixing assignments Z_i using the Adjusted Mutual Information (AMI) averaged over the estimated vs. true assignments for each observed node X_j in a simulated graph G, and the mixing structure in G_Z using F1 scores (binary edge existence $Z_i \to X_j$) and Jaccard indices (set overlap of each Z_i's observed targets). We include simple baselines that apply mixture modeling without causal information for comparison.

In Figure 7.3, we observe reliable recovery of mixing assignments (AMI) and their targets (F1, Jacc) for our approach given the true graph (dashed green) and for the full framework (solid green). Applying a clus-

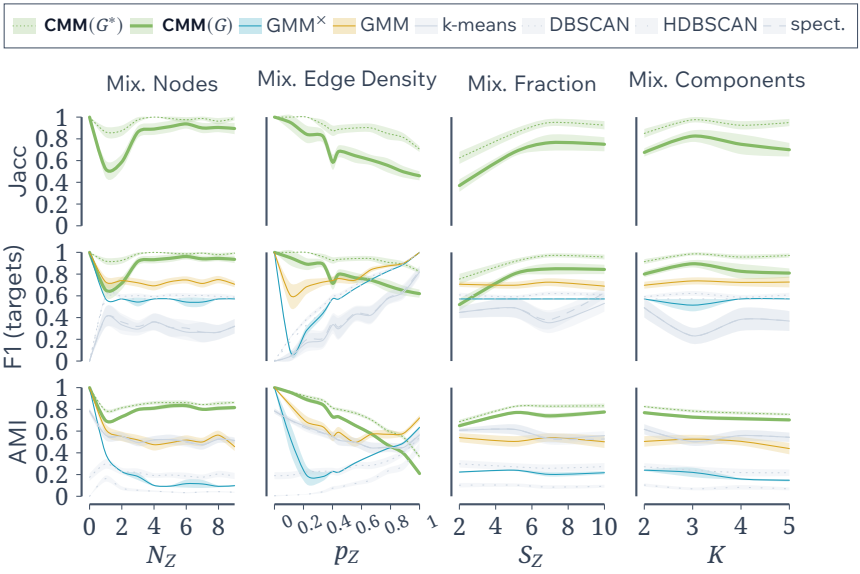


Figure 7.3 In synthetically generated CMMs, we evaluate the quality of the recovered mixing structure, evaluating the affected observed variables (F1 (target)), variable sets affected by the same latent variable (Jacc) as well as per-node mixing assignments (AMI).

tering baseline such as GMM over all nodes in the graph as in (Kumar et al., 2024) (blue, $\times$) is misspecified in case $N_Z > 1$ as per-node assignments are distinct, hence not surprisingly performance worsens as N_Z increases (left). Given this, we also apply the models to each node in turn (purple, gray), where gray variants show different instantiation choices. These perform not much better than random splitting in most cases (AMI, F1) and with accurate set recovery (Jacc).

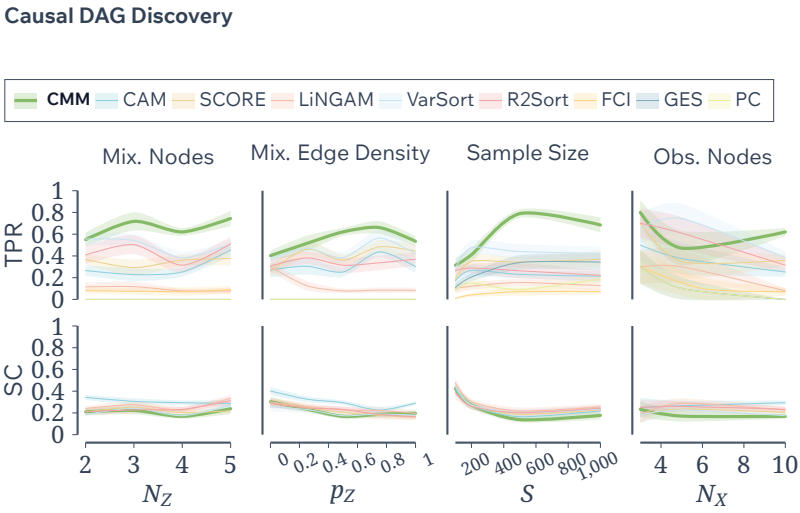


Figure 7.4 In synthetically generated CMMs, we evaluate the quality of the causal DAG over the observed variables in terms of the separation distance (SD) and true positive rate over directed edges (TPR).

7.5.3 Discovering Causal Structure

Next, we evaluate the learned causal DAG resp. CPDAG G against the ground truth G^*. We consider again the separation-based distance metrics (Wahl & Runge, 2025), including the Separation Distance (SD). To in particular demonstrate the strength of our approach in discovering edge orientations, we again compute orientation F1 scores and report true positive rates (TPR). Baselines include a range of causal discovery methods, CAM, SCORE, LiNGAM, FCI, GES, PC, VarSort, and R2Sort.

While some baselines perform reasonably well in mixed settings, our CMM maintains high accuracy across both structural and causal direction evaluations (Figure 7.4). We note that Mixture-UTIGSP is not meaningfully applicable here as there may be multiple mixing variables and no well-defined "observational" part of the dataset as required as input to UTIGSP, thus we will rather consider an experimental setting with a single mixing variable to compare to it.

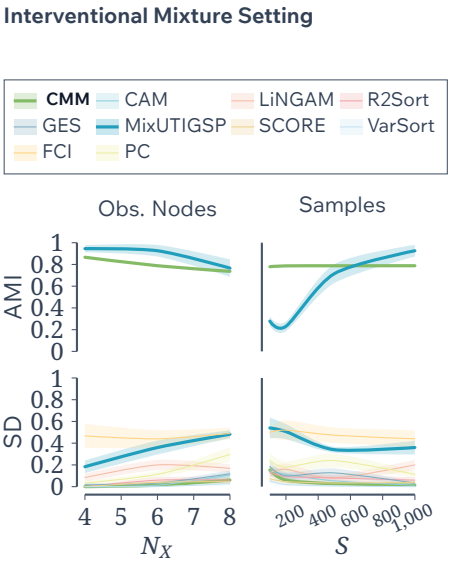


Figure 7.5 For data generated from the interventional mixture setting as in Kumar et al. (2024), we show the quality of mixture separation (AMI, cf. Figure 7.3) and causal DAG discovery (SD, cf. Figure 7.4).

Case Study: Mixtures of Interventions. Next we replicate the experimental setup of Kumar et al. (2024) using their data generators to evaluate performance in the mixtures-of-interventions setting. As Figure 7.5 shows, the GMM used in Mixture-UTIGSP performs well in recovering mixture assignments as expected given that its modeling assumptions are met in this setting. Our method performs comparably in mixture assignment recovery and more reliably recovers the causal graph (Figure 7.5 bottom). We also noticed a competitive performance of VarSort in this dataset suggesting potential Var-sortability which may inflate the performance of the discovery approaches.

	Mixed/Iv. Samples (AMI)				Causal Graph	
Target	CMM*	**CMM**	MixUTIGSP	Metric	**CMM**	MixUTIGSP
Akt	0.02	0.04	0.04	TP	5	0
PKC	0.26	0.43	**0.49**	FP	10	2
PIP2	0.10	**0.10**	0.03	FN	5	10
Mek	0.20	**0.20**	0.14	SD	**0.25**	0.75
PIP3	0.00	0.00	0.00	S/C	**0.27**	0.57

Table 7.1 We evaluate the mixing assignment, i.e., assignment to interventional conditions, discovered for the flow cytometry data (Sachs et al., 2005).

Table 7.2 We evaluate the causal DAGs discovered over the flow cytometry data, combined from multiple interventional conditions.

Case Study: Flow Cytometry Data. Last, we return to the flow cytometry data by Sachs et al. (2005), over 11 variables studied under different perturbations. In particular, the variables Akt, PKC, PIP2, Mek, and PIP3 were manipulated. As in previous analyses (Wang et al., 2017) we combine these subsamples into a larger dataset of size 5846 and test how well, for each given target node, its intervened subsample can be recovered (AMI). The dataset split found by Mixture-UTIGSP appears to match PKC best, similar for our per-variable splits which match PIP2 and MEK slightly better (Table 7.1), but neither method reaches convincing AMI, consistent with previous observations (Squires et al., 2020; Kumar et al., 2024) that the interventional targets are difficult to identify in this dataset. On causal discovery, our approach performs better in terms of separation distances (SD, S/C) albeit discovering a number of spurious (FP) edges (Table 7.2). All other baselines report similarly high false positive directions (10+ FP) with the exception of PC (2 FP). Also, some edge direction in the ground truth such as Raf → Mek are debated. We discover the Mek → Raf direction which has been previously discussed as agreeing better with given data (Mooij et al., 2020), where we also recall the confounding findings for this edge from Chapter 5.

7.6 Conclusion

We addressed the problem that real-world datasets rarely come from a fixed data-generating process, but could be a combination of subpopulations with distinct causal mechanisms. This gives rise to a causal mixture model (CMM) in the sense that each effect given its causes results from a finite mixture of linear regressions. After characterizing this model, we established theoretical results showing that consistent scoring criteria, such as the latent-aware BIC, allow causal identifiability. Complementary to these insights, we propose a practical implementation for causal discovery using this score which we demonstrate to have strong empirical performance in various settings.

In the real-world case study (Sachs et al., 2005), while the approaches recover the causal graph structure to an extent, we observed limited ability of both the CMM and Mixture-UTIGSP to discover the intervention targets, likely due to the strong assumption of linearity. As our proposed approach can in principle easily incorporate richer models, future work could study scoring criteria and theoretical guarantees for non-linear mixtures, encouraged by a proof-of-concept experiment that we provide in the original work (Mameche et al., 2025b). Future work could also study how to further interpret the meaning of a given Z, especially when it points to previously unknown states, for example, by exploring explainability approaches. Furthermore, invoking Expectation Maximization at every scoring step comes with drawbacks regarding the scalability of our method, especially when relying on random restarts, therefore subsequent analysis could devise algorithmic approaches to reduce this overhead.

Software. A Python implementation at `github.com/srhmm/causalchange` is available. To get started, the Jupyter notebook `demo/ch7_cmm.ipynb` shows an example usage.

8| Conclusion

Case Study Cocoa Controlled Trial

To settle an argument on which is the best type of chocolate, Charlotte wants to study the effects of different chocolate types on happiness. Charlotte designs a Randomized Controlled Trial (RCT) where participants receive either milk chocolate, dark chocolate, or no chocolate at all, and monitors their well-being. However, she faces complications during the trial.

"

Despite my efforts to keep the study groups blinded, some of the participants changed group mid-study. Indeed, those in the control group (who received no extra chocolate) decided to start raiding the chocolate of those in the interventional arms (dark chocolate or milk chocolate). [Also,] the fact that the study period overlapped with Halloween may have adversely affected the overall consumption of chocolate during the study. [Overall, I observed] no significant difference in happiness between any of the groups (Chan, 2008).

"

Central to many areas of science, and arguably central to human learning and intelligence, is answering questions of cause and effect. When human background knowledge is not available, the Randomized Controlled Trial (RCT) is the gold standard practice for answering causal questions; the study from Chan (2008), however, points to the practical challenges of RCTs, ranging from imperfect adherence, crossover effects between the randomized groups, to confounding and, in some

cases, infeasibility due to time constraints or ethical concerns.

Case Study (ctd) Cocoa Correlation

After conducting the chocolate-happiness trial, Charlotte researches ideas for a follow-up study on other effects of chocolate. She comes across a 2012 study (Messerli, 2012) remarking on a strong correlation between national chocolate consumption and the number of Nobel laureates (per-capita). One of the laureates even states,

> *I attribute essentially all my success to the very large amount of chocolate that I consume. Personally I feel that milk chocolate makes you stupid ... dark chocolate is the way to go (BBC News, 2012).*

Controlled studies aside, we can also attempt learning from purely observational studies. This, however, similarly requires caution given the well-known fact that *correlation is not causation*. The chocolate-Nobel prize association, for instance, is presumably either spurious, due to methodological issues such as the ecological inference fallacy, or due to confounding factors, such as access to education and resources (Maurage et al., 2013).

The two case studies give two strategies for learning descriptive models of the world, either through controlled experimentation and active interaction, or by passive observation of static processes. Given the challenges of both, this work explores a middle ground, studying cases where observations stem from systems observed under change, intervention, and over time.

8.1 Conclusion

To position our contributions, we revisit Pearl's ladder of causation (Pearl, 2009) in Figure 8.1. It shows levels of reasoning of increasing complexity, ranging from *association* by observation and measuring correlations, to *intervention* by analyzing the effect of changes and interventions, to answering *counterfactual* questions through retrospective thinking and imagination. In particular, the counterfactual level requires reasoning about an *alternative* world rather than the observed one, thereby not al-

lowing for the direct collection of data. At the interventional level, on the other hand, one can collect data by inducing *changes* to a system or process, such as via an RCT, or by measuring a system under change, in different contexts, or across time (non-i.i.d.). When data results from passive observation alone, we can position it at the associational level (i.i.d.).

Such i.i.d. data is the starting point of both statistical learning and of classical approaches in causal discovery (orange arrow). As we have seen, however, i.i.d.-ness is an idealized assumption for real-world systems. Therefore, our main research question is how to learn causal models from non-i.i.d. data (blue arrow).

To this end, we addressed the following research questions.

Question 1. (Multi-Context Causal Discovery) We first addressed observations from multiple contexts, where biases such as Simpson's paradox can arise (Population Paradox I, II). We formalized independent change of causal mechanisms within DAGs and SCMs and showed that under standard Markov and faithfulness assumptions, IC allows identification of causal directions (Theorem 3.2). In practice, we propose a score-based approach using Gaussian process models and MDL (Figure 3.3). We also proposed the causal discovery algorithm TOPIC. In some applications, data not only stems from different contexts but also from a process monitored over time, with unknown temporal regimes (Changepoint Challenge). Therefore, we generalized the framework to TCGs (Theorem 6.1) and proposed the algorithm SPACETIME to jointly learn the TCG and changepoints, supported with case studies from the environmental sciences (Figure 6.9).

Question 2. (Unknown Confounders) After noting causal sufficiency violations in the data by Sachs et al. (2005) (Pathway Problem II), we extended the scope to settings with latent confounders. We concluded that we can identify confounding from dependent distribution shifts of observed variables. We proposed a statistical test to detect such cases, even when the causal graph is unknown (Theorem 5.2). The COCO algorithm detects jointly confounded variable sets in practice (Figure 5.3).

Question 3. (Unknown Contexts) Last, we considered mixtures of populations, motivated by settings where an unknown subset of the population shows treatment resistance (Population Paradox III). We proposed a Causal Mixture Model (CMM) where the causal mechanism of each variable is a mixture of SCMs. The latent mixing variables introduce

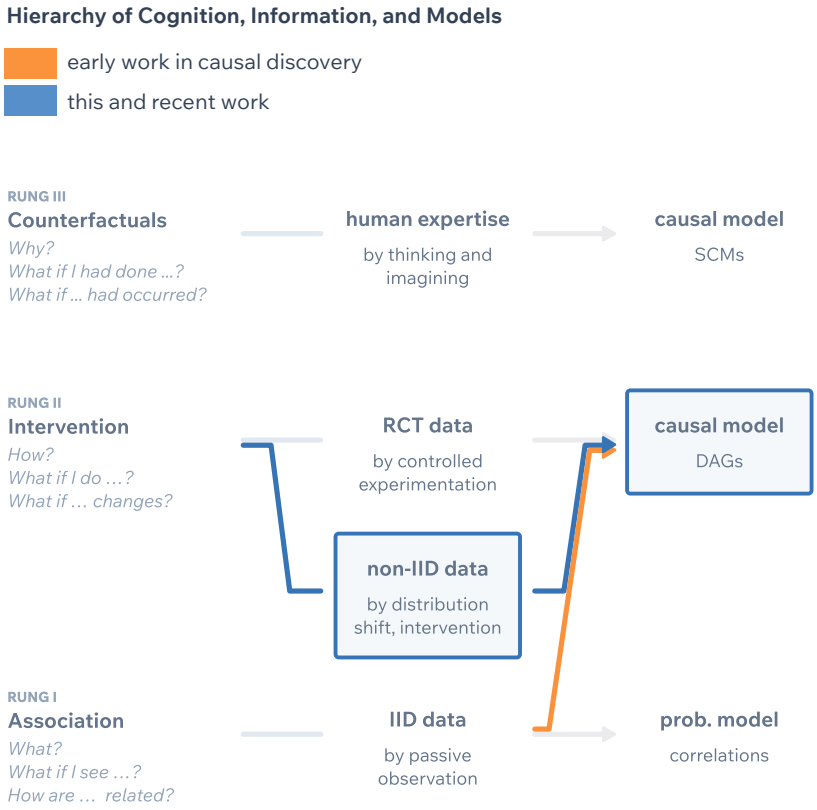


Figure 8.1 Pearl's ladder of causation (Pearl, 2009) showing different modes of reasoning (left), various forms of information and data (middle), and abstract descriptions of reality (right). The data with distribution changes that we considered forms a middle ground between the ideal of controlled experimentation through RCTs (middle) and i.i.d. observations from static distributions (bottom).

sufficient asymmetry to identify causal directions even when population memberships are unobserved (Theorem 7.1). Our algorithm integrates EM and the latent-aware BIC into TOPIC to discover both assignments to subpopulations (Figure 7.3) as well as the causal graph (Figure 7.4).

In summary, the different non-i.i.d. settings allow causal learning and identifiability under the principle of independence of causal mechanisms (ICM), particularly its implication of independent change. Different forms of the ICM (cf. Peters et al., 2017; Schölkopf et al., 2021) have been successfully applied to causal discovery, such as causal invariance (Peters et al., 2016; Heinze-Deml et al., 2018; Arjovsky et al., 2019), ICM-generative processes Guo et al. (2022); Huang et al. (2020) and the sparse shift principle (Perry et al., 2022). Overall, there is a general shift of attention in the field of causal discovery from assuming i.i.d. data towards addressing non-i.i.d. settings. The multi-context setting is the most well-studied (Mooij et al., 2020; Huang et al., 2020; Squires et al., 2020) or, more recently, population mixtures (Mazaheri et al., 2023; Kumar et al., 2024), similarly in the time series setting to multiple contexts (Günther et al., 2023b) or with unknown temporal regimes (Saggioro et al., 2020). In causal representation learning (CRL) settings, in particular, where the goal is the reconstruction of latent variables (Schölkopf et al., 2021), the use of multi-context datasets is commonly the starting point to allow for identifiability (Ahuja et al., 2023).

While this research progress is encouraging, transitioning from the orange to the blue learning paradigm (Figure 8.1) still requires future research efforts. We conclude with future work directions related to the contributions.

8.2 Future Work

In the following, we outline some future directions, starting from limitations of this work and moving towards broader downstream tasks and ML applications.

8.2.1 Direction 1: Relaxing Assumptions

While one of our goals is to allow causal discovery under more realistic assumptions, we addressed only a selection of assumptions here, such as i.i.d.-ness, temporal stationarity, and sufficiency.

First, we assume throughout this work that no *selection bias* exists. As we sketch in Pathway Problem III, violations of IC could point to latent selection variables similarly as to latent confounding variables. The inter-

play of selection bias and distribution change or intervention however requires further careful investigation; for example, Dai et al. (2025) note that the interaction between selection and intervention results in additional complexities as their time order is important, requiring a refined causal modeling structure such as twin graphs (Dai et al., 2025). An interesting open question is also how to *distinguish* between the presence of missingness and confounding, as well as how to *correct for* their presence, for example, through doubly robust regression (Boeken et al., 2023). Of note is also *measurement error* as a different source of bias that we do not address here (e.g., Zhang et al., 2017). Overall, the intersection of intervention, selection, and confounding remains an area that is arguably understudied.

Given that, with the exception of Chapter 5, we have restricted the learning approach to score-based methods, an interesting direction is to adopt constraint-based methods instead. That is, rather than MDL-based regularization as in a score-based paradigm, one could base IC-based discovery purely on *conditional discrepancy* (and invariance) testing, reminiscent to the approach by Perry et al. (2022). Related to this point, in most chapters, we assume functional models with additive Gaussian noise; hence, extensions for *heteroskedastic* noise (Immer et al., 2023) or, in the case of time series, history-dependent noise (Gong et al., 2022) are of interest. We also mostly focus on continuous-valued random variables and could hence shift focus to *categorical* variables. Here, the ideas in entropic causal discovery Kocaoglu et al. (2017) are similar in spirit to the MDL- or MI-based perspectives. Notably, the approach in Chapter 5 is in principle agnostic to the specific test used to identify causal mechanism shifts, and hence can be transferred directly to discrete settings using a suitable test (e.g., Marx & Vreeken, 2019a). As the most restrictive modelling assumption, Chapter 7 only addresses mixtures of linear regressions. EM-based approaches to estimate nonlinear mixtures exist, suggesting a plug-in approach that replaces the linear mixture estimation and scoring with nonlinear variants; theoretical guarantees for this case may however require additional study.

Besides the above, further study of *(change) faithfulness* assumptions is also of importance. For example, Mazaheri et al. (2025b) propose an intervention-adjacency (IA) faithfulness condition towards intervention-only causal discovery, an idea that is closely related to IC-only causal discovery under (change) faithfulness in the augmented causal model in Chapters 3-4.

Finally, on the topic of assumption choices, both *verifying* when spe-

cific assumptions hold, as well as *benchmarking* of algorithms under assumption violations and on real-world benchmarks where these might not hold (Brouillard et al., 2024) are important next steps towards application in real-world systems.

8.2.2 Direction 2: Model Choices

Besides the above assumptions, we can also question the specific causal model structures we assume, depending on the downstream application.

In Chapters 3-4, we assume that the same causal DAG applies in all contexts. When this does not hold, one could resort to estimating differences in the DAG structure across contexts (Wang et al., 2018), or using richer models such as *mixtures of DAGs* (Saeed et al., 2020), the latter especially interesting in the setting of Chapter 7 with unknown contexts. To this end, an additional Expectation-Maximisation would need to be placed around TOPIC to infer a mixture of DAGs. We also note that we assume the set of measured random variables in all contexts to remain fixed, while among others Triantafillou & Tsamardinos (2015) address *non-overlapping* variable sets.

In Chapter 5, we only address exogenous confounding, assuming confounders to be source nodes without causal interactions. In similar vein as Agrawal et al. (2023), one could consider *endogenous* unobserved variables, or allow a *hierarchical* confounding structure (Huang et al., 2022). Conversely, Chapter 6 does not model latent confounding in time series (cf. Malinsky & Spirtes, 2018); hence, the strategies in Chapter 5 could be explored here.

Meanwhile, while Chapter 7 relaxes the assumption of global latent-class confounders (Mazaheri et al., 2023; Kumar et al., 2024) to allow for multiple latent discrete variables, it makes simplifying assumptions on the latent structure, assuming them to be root nodes and independent of one another. Relaxations of this could address settings where populations have shared substructures.

It should also be noted that DAGs, with their property of acyclicity in particular, are not always a meaningful abstraction of real-world systems, especially in settings where feedback between variables is common. Hence, the further study of *cyclic SCMs* is another important direction. Lorch et al. (2024) instead learn stochastic differential equations (SDEs) whose stationary densities model how a system behaves under interventions. These stationary diffusion models do not require the assumption of

acyclicity nor a graphical abstraction, although one can be reconstructed. SDEs are similar to SCMs in providing a mechanistic model of causal interactions, but taking a different level of abstraction (Mooij et al., 2013).

Similarly, we can reconsider the level of abstraction of the causal DAG. In our case, we assume a small set of observed, semantic variables with meaningful causal structure, which is not suited to unstructured data or images, where semantic variables may be latent, such as representing objects or concepts. An increasingly large body of research addresses this setting through *causal representation learning* (CRL), where we are interested in both reconstructing such latent variables as well as the causal structure. Recent work mainly addresses identifiability theory. While not the focus of this work, many if not most works in CRL similarly consider multi-context and interventional settings (e.g., Ahuja et al., 2023; von Kügelgen et al., 2023).

8.2.3 Direction 3: Broadening of Scope

Lastly, while our focus is on the problem of causal discovery, the research direction of viewing distribution shifts and non-i.i.d.-ness through the lens of causality can also be formulated more broadly with other downstream tasks in mind.

Closest to causal discovery, which attempts to learn a causal model from data, is the *causal inference* setting, which a priori assumes a known causal model and focuses on treatment effect estimation problems. Here, a line of work also addresses distribution shifts with the study of *heterogeneous treatment effects* (e.g., Angus & Chang, 2021). Related to the mixture-of-population setting in Chapter 7, recent work proposes mixture modelling in the causal inference setting (Kim et al., 2024; Mazaheri et al., 2025a).

A different branch of related work addresses active learning settings such as adaptive intervention design (Hauser & Bühlmann, 2014), where the central question is how to choose interventions to identify causal structures most efficiently. This presumes the ability to induce active changes to a system, as is often possible in biology and genetics (Dominguez et al., 2016; Dibaeinia & Sinha, 2020), thereby moving closer towards the RCT setting. Finally, the availability of an SCM-based causal model can also allow addressing *counterfactual questions*, for example, by modeling causal mechanisms through diffusion models (Sanchez & Tsaftaris, 2022).

Moving towards broader areas of ML, a closely related task is *domain*

adaptation. For example, Magliacane et al. (2018) propose generalizing causal conclusions across contexts; Arjovsky et al. (2019) propose the use of invariant features in empirical risk minimization. Besides domain adaptation, the perhaps most closely related field of ML is *reinforcement learning* (Zeng et al., 2024), given that it involves sequential decision making of an agent while in *active* interaction with the environment, hence fitting well within the interventional rung of information collection. This point encourages the line of work on *causal* reinforcement learning, a field that is still in its early developments (Zeng et al., 2024).

Finally, moving further towards large ML models, we can ask the question of how far the observational, associational rung of reasoning (Figure 8.1) can take us, in the extreme of using vast amounts of data and equally large models. As such, Generative AI and *Large Language Models (LLMs)* have gained widespread use and popularity. In their current form, LLMs often achieve at best memorization of causal knowledge where genuine understanding can be questioned, such that memorization of biases in training data, faults in reasoning patterns, or even hallucinations are common (Wu et al., 2024). Given the previous points, the study of reinforcement-learning-based training of such systems, or of other ways to effectively allow these models access to interventional information and modes of counterfactual reasoning, is a worthwhile endeavor, albeit a much less straightforward research direction.

8.2.4 Outlook

Returning to the ladder of causation in Figure 8.1, we can use it to comment on the past, present, and future of the field of causal learning and discovery, in the context of the topic of this thesis.

Causal discovery has, in early works, started from the classical paradigm (orange) of learning from i.i.d. data (e.g., Pearl, 2009; Chickering, 2002). Given the missing access to information from the interventional and counterfactual rungs central to definitions of causality, this forms an underspecified learning task. Some even challenge whether structures recovered purely from conditional independence constraints should be claimed to have a causal interpretation (Dawid, 2010). The learning task becomes even more ill-posed when jointly performing causal discovery and representation learning, as is the focus of a large part of the recent literature.

Hence, much of the present work in causal discovery and CRL ad-

dresses non-i.i.d. settings (blue) which, due to coming from different contexts, populations, or time spans, provide information on how systems respond to *change*. Consequently, such data might be better suited to attempt causal discovery or, at least, allow doing so under a relaxed set of assumptions. This shows a bidirectional connection, with causal models being better suited to address non-i.i.d. data compared to probabilistic models (Chapter 1), but also non-i.i.d. data being better suited to learn causal models compared to correlational data (this Chapter).

Ultimately, causal learning from data is not possible without additional assumptions, but shifting from the orange to the blue learning paradigm promises learning under more realistic assumptions, and motivates to move further in this direction in the future.

A| Notation

A.1 Notation in Chapters 2-4

Notation	Meaning	Properties
$\mathcal{X}$	observed variables	$\mathcal{X} = \{X_1, \ldots, X_d\}$
X_j	jth observed variable	$X_j \in \mathcal{X}$
$\boldsymbol{x}$	realization of $\mathcal{X}$	$\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$
$\mathcal{D}$	multi-context dataset	$\mathcal{D} = \{\mathcal{D}^1, \ldots, \mathcal{D}^m\}$
$\mathcal{D}^c$	dataset in context c	$\mathcal{D}^c = \{\boldsymbol{x}^{c,1}, \ldots, \boldsymbol{x}^{c,n_c}\}$
$\boldsymbol{x}^{c,i}$	ith observation in context c	$\boldsymbol{x}^{c,i} \in \mathbb{R}^d$
$P_{\mathcal{X}}$	distribution over $\mathcal{X}$	
p	density function	
$P_{X \mid \mathcal{X}_S}$	conditional distribution of $X \mid \mathcal{X}_S$	sometimes $P(X \mid \mathcal{X}_S)$
P^c	distribution in context c	
p^c	density function in context c	
$x_j^{c,i}$	ith value of X_j in context c	

Table A.1 Notation related to random variables (r.v.s).

Notation	Meaning	Properties
G	causal DAG over $\mathcal{X}$	$G = (\mathcal{X}, \mathcal{E})$
$\mathcal{E}$	directed edges in G	$(X_k, X_j) \in \mathcal{E}$
Pa_j	parents of X_j in G	$Pa_j \subseteq \mathcal{X} \setminus \{X_j\}$
N_j	exogenous noise variable	
Π	latent variable set (mechanism assignment)	$\Pi = \{\Pi_1, \ldots, \Pi_d\}$
Π_j	latent variable for X_j	$\Pi_j \in \{1, \ldots, k_j\}$
π_j^c	realization of Π_j in context c	$\pi_j^c \in \{1, \ldots, k_j\}$
f_j^k	kth causal mechanism of X_j	active when $\pi_j^c = k$
$G_{\Pi,\mathcal{X}}$	augmented DAG over $\mathcal{X} \cup \Pi$	$G_{\Pi,\mathcal{X}} = (\mathcal{V}_{\Pi,\mathcal{X}}, \mathcal{E}_{\Pi,\mathcal{X}})$
$\mathcal{V}_{\Pi,\mathcal{X}}$	nodes of $G_{\Pi,\mathcal{X}}$	$\mathcal{V}_{\Pi,\mathcal{X}} = \mathcal{X} \cup \Pi$
$\mathcal{E}_{\Pi,\mathcal{X}}$	directed edges in $G_{\Pi,\mathcal{X}}$	
$\boldsymbol{x}_j^{(k)}$	pooled observations for X_j under mechanism k	
$\boldsymbol{pa}_j^{(k)}$	pooled parent observations for X_j under mechanism k	
$L(\boldsymbol{x}_j^{(k)} \mid \boldsymbol{pa}_j^{(k)})$	MDL score for kth mechanism	
$L(\mathcal{D}; G, \Pi)$	overall MDL score under (G, Π)	
$L(\mathcal{D}; \mathcal{M})$	abbreviates $L(\mathcal{D}; G, \Pi)$	

Table A.2 Notation related to the causal model over multiple contexts.

A.2 Notation in Chapter 5

Notation	Meaning	Properties
$\mathcal{V}$	full variable set	$\mathcal{V} = \mathcal{X} \cup \mathcal{L}$
$\mathcal{X}$	observed variables	$\mathcal{X} = \{X_1, \ldots, X_{d_x}\}$
$\mathcal{L}$	latent confounders	$\mathcal{L} = \{L_1, \ldots, L_{d_l}\}$
$G_{\mathcal{V}}$	causal DAG over $\mathcal{V}$	$G_{\mathcal{V}} = (\mathcal{V}, \mathcal{E}_{\mathcal{V}})$
$\mathcal{E}_{\mathcal{V}}$	directed edges in $G_{\mathcal{V}}$	
$G_{\mathcal{X}}$	induced subgraph over $\mathcal{X}$	$G_{\mathcal{X}} = (\mathcal{X}, \mathcal{E}_{\mathcal{X}})$
$\mathcal{E}_{\mathcal{X}}$	directed edges in $G_{\mathcal{X}}$	
Π^*	true mechanisms over $\mathcal{V}$	$\Pi^* = \{\Pi_1^*, \ldots, \Pi_{d_v}^*\}$
Π_j^*	true mechanisms for V_j	
Π	inferred mechanisms over $\mathcal{X}$	$\Pi = \{\Pi_1, \ldots, \Pi_{d_x}\}$
Π_j	inferred mechanisms for X_j	$\Pi_j \in \{1, \ldots, k_j\}$
$\mathrm{I}(\Pi_1, \Pi_2)$	mutual information (MI)	
$\mathrm{T}(\Pi_1, \ldots, \Pi_s)$	total correlation (TC)	

Table A.3 Notation related to latent confounding in multiple contexts.

A.3 Notation in Chapter 6

Notation	Meaning	Properties
$\mathcal{X}$	observed variables	$\mathcal{X} = \{X_1, \ldots, X_d\}$
$\mathcal{D}$	multi-context time-series data	$\mathcal{D} = \{\mathcal{D}^1, \ldots, \mathcal{D}^m\}$
$\mathcal{D}^c$	time series in context c	$\mathcal{D}^c = (\boldsymbol{x}^c(1), \ldots, \boldsymbol{x}^c(n))$
$\boldsymbol{x}^c(t)$	obs. vector at time t in context c	$\boldsymbol{x}^c(t) \in \mathbb{R}^d$
$G_{(t)}$	temporal causal graph (TCG)	$G_{(t)} = (\mathcal{V}_{(t)}, \mathcal{E}_{(t)})$
$\mathcal{V}_{(t)}$	nodes of the TCG	nodes $X^c_{j(t)}$ for all j, t, c
$\mathcal{E}_{(t)}$	directed edges in the TCG	
$G_{(\tau_{\max})}$	finite-lag graph	$G_{(\tau_{\max})} = (\mathcal{V}_{(\tau_{\max})}, \mathcal{E}_{(\tau_{\max})})$
$\tau_{\max}$	maximum time lag	$\tau \in \{0, \ldots, \tau_{\max}\}$
$\mathcal{T}$	changepoint set	$\mathcal{T} = \{1 = \tau_0 < \tau_1 < \cdots < \tau_{n_\mathcal{T}}$
I_r	rth temporal regime	$I_r = \{\tau_{r-1}, \ldots, \tau_r - 1\}$
$\pi_j^{(c,r)}$	mechanism assignment of X_j in context c, regime r	$\pi_j^{(c,r)} \in \{1, \ldots, k_j\}$
$\boldsymbol{x}_j^{(k)}$	pooled observations for X_j under mechanism k	
$\boldsymbol{pa}_j^{(k)}$	pooled parent observations for X_j under mechanism k	
$L(\boldsymbol{x}_j^{(k)} \mid \boldsymbol{pa}_j^{(k)})$	local temporal MDL score	
$L(\mathcal{D}; G, \mathcal{T}, \Pi)$	overall temporal MDL score	

Table A.4 Notation related to temporal causal graph (TCG) discovery and changepoint detection in time series.

A.4 Notation in Chapter 7

Notation	Meaning	Properties
$\mathcal{X}$	observed variables	$\mathcal{X} = \{X_1, \ldots, X_d\}$
$\mathcal{Z}$	latent mixing variables	$\mathcal{Z} = \{Z_1, \ldots, Z_{d_z}\}$, $Z_i \sim \text{Categorical}(\boldsymbol{\gamma}^i)$
La_j	latent variable assigned to X_j (mechanism selector)	$La_j \in \mathcal{Z} \cup \{\varnothing\}$
$G_{\mathcal{Z}}$	augmented causal DAG	$G_{\mathcal{Z}} = (\mathcal{X} \cup \mathcal{Z}, \mathcal{E}_{\mathcal{Z}})$
G	induced DAG over $\mathcal{X}$	$G = (\mathcal{X}, \mathcal{E})$
Pa_j	parents of X_j in G	$Pa_j \subseteq \mathcal{X} \setminus \{X_j\}$
$\boldsymbol{\beta}_{jk}$	regression coefficients of component k	$\boldsymbol{\beta}_{jk} \in \mathbb{R}^{\lvert Pa_j \rvert}$
γ_k^i	mixing weight of component k for Z_i	$\sum_{k=1}^{k_i} \gamma_k^i = 1$
MLR	mixture of linear regressions	cond. model for $X_j \mid Pa_j$
$p^{\text{MLR}}(x, \boldsymbol{pa}_j; \boldsymbol{B}, \boldsymbol{\gamma}, \sigma^2)$	local conditional density	$p(X_j \mid Pa_j)$ under MLR
$\boldsymbol{\theta} = (\boldsymbol{B}, \boldsymbol{\gamma}, \sigma^2)$	model parameters	
$\text{BIC}_{\hat{Z}}^{\text{ML}}(\mathcal{H})$	latent-aware BIC score	maximized over $K \leq k_{\max}$

Table A.5 Notation related to causal mixture models.

B| Background

In this section, we introduce additional background regarding structural causal models and causal graphs.

B.1 Structural Causal Models

Given a set of random variables $\mathcal{X} = \{X_1, \ldots, X_d\}$ that follow a distribution $P_{\mathcal{X}}$ that has a joint probability density p with respect to an appropriate measure, we encode their underlying data generating process (DGP) as a structural equation model (SEM) (Koller & Friedman, 2009). An SEM places a set of assumptions on the DGP in the form of one functional dependency $X_j = f_j(\mathcal{X}_j)$ for each random variable $X_j \in \mathcal{X}$ given a set $\mathcal{X}_j \subseteq \mathcal{X} \setminus \{X_j\}$. Of special interest are *structural causal models (SCMs)* (Bollen, 1989) in which f_j are interpreted as causal mechanisms,

$$X_j = f_j(\mathcal{X}_j, N_j) \qquad \text{with } \mathcal{X}_j \subseteq \mathcal{X} \setminus \{X_j\} \text{ and } N_j \in \mathcal{N}\,, \tag{B.1}$$

where the set of random variables $\mathcal{X}$ is extended to also include unobserved random variables $\mathcal{N} = \{N_1, \ldots, N_d\}$ corresponding to random noise. Above, $\mathcal{X}_j$ denotes the set of causal parents of X_j, therefore we write them throughout the chapters as $\mathcal{X}_j = Pa_j$.

It is also useful to establish a correspondence of SCMs with causal graphs (Pearl, 2009).

B.2 Bayesian Networks and Causal Graphs

A graphical representation of conditional independencies in $\mathcal{X}$ are *Bayesian networks* (Koller & Friedman, 2009) in the form of a (fully) directed acyclic graph (DAG) $G = (\mathcal{X}, \mathcal{E})$ where the nodes correspond to the random variables $\mathcal{X}$ and edges to a set $\mathcal{E} = \cup_{j=1}^{d} \{X_i \to X_j \mid X_i \in \mathcal{X}_j\}$. In particular, a DAG is called a Bayesian network over $\mathcal{X}$ if the joint probability density function can be written as

$$p(\boldsymbol{x}) = \prod_{X_j \in \mathcal{X}} p(x_j, \boldsymbol{x}_j)\,, \qquad \text{for } \mathcal{X}_j \subseteq \{X_1, \ldots, X_j\}\,. \tag{B.2}$$

Thus, a Bayesian network encodes a family of distributions consistent with the factorization of the joint density above.

Conditional independencies can be read from the graph in terms of the *d-separation* (Pearl, 2009).

Definition B.1 (d-Separation). *For any pairwise disjoint subsets $\mathcal{X}_S, \mathcal{Y}_S, \mathcal{Z}_S \subseteq \mathcal{X}$, and $\mathcal{X}_S, \mathcal{Y}_S \neq \emptyset$, we have that $\mathcal{X}_S$ and $\mathcal{Y}_S$ are d-separated in G given $\mathcal{Z}_S$, denoted $\mathcal{X}_S \perp_d \mathcal{Y}_S \mid \mathcal{Z}_S$ in G, under the following condition,*

$$\mathcal{X}_S \perp_d \mathcal{Y}_S \mid \mathcal{Z}_S \iff \begin{array}{l} \textit{all paths from any variable in } \mathcal{X}_S \\ \textit{to any variable in } \mathcal{Y}_S \textit{ are blocked by } \mathcal{Z}_S\,. \end{array}$$

We call a path blocked by $\mathcal{Z}_S$ if it either

- *traverses a section $\to W \to$, $\leftarrow W \leftarrow$ or $\leftarrow W \to$ for some node $W \in \mathcal{Z}_S$, or*
- *traverses a collider section $\to W \leftarrow$ where neither W nor any descendant of W is contained in $\mathcal{Z}_S$.*

When a distribution $P_{\mathcal{X}}$ exhibits all the conditional independencies that one can read from the graph G, G is commonly called an I-map of $P_{\mathcal{X}}$ denoted as $P_{\mathcal{X}} \in I_G$. The I-map defines an equivalence relation among all DAGs via the relation $G \equiv_M G' \iff I_G = I_{G'}$, called the *Markov equivalence class* (MEC).

A graphical representation of MECs can be given through the common notion of *completed partially directed acyclic graphs* (CPDAGs) (Chickering, 2002) also known under other names such as maximally oriented graphs (Meek, 1997). A partially directed graph (PDAG) $\mathcal{P}$ contains both

undirected and directed edges, and can be associated to an equivalence class $\mathscr{E}(\mathcal{P})$ with $G \in \mathscr{E}(\mathcal{P})$ if and only if $G, \mathcal{P}$ have the same skeleton and v-structures. The notion of completion of PDAGs allows for representing equivalence classes uniquely. To this end, for a given equivalence class $\mathscr{E}$ one distinguishes between *compelled* edges with the same directionality in every member of $\mathscr{E}$, and *reversible* edges otherwise. The completed PDAG $\mathcal{P}$ for $\mathscr{E}$ is then the one having a directed edge for every compelled edge in $\mathscr{E}$, and an undirected edge for every reversible edge in $\mathscr{E}$.

When each of the factors in the factorization of Eq. (B.2) correspond to a functional dependency of an SCM, we call the resulting graph a *causal DAG*.

We now return to the assumptions implicit in an SCM.

Assumption B.1 Causal Interpretation *Each functional dependency f_j corresponds to a true causal mechanism in the data, with $\mathcal{X}_j$ being the direct causes of the direct effect X_j.*

As a corollary, the corresponding causal graph contains no directed cycles.

Assumption B.2 Orientation and Acyclicity
The causal graph of an SCM is a DAG.

Hence, we can once again identify the direct causes of each effect X_j with its parents in the corresponding DAG, $\mathcal{X}_j = Pa_j$. The final assumptions regard exogeneity and independence of the noise variables.

Assumption B.3 Exogeneity of Noise *The random variables $\mathcal{N}$ are exogenous; that is, there are no edges $X_j \to N_i$, for any X_j and any N_i.*

Assumption B.4 Independence of Noise
The random variables $\mathcal{N}$ are mutually independent.

C| Appendix for Chapter 3

C.1 Assumptions

We first recall the assumptions relevant for the results.

Assumption 2.1 (Causal Markov Property)**.** *We assume the causal Markov property,*

$$p(\boldsymbol{x}, \boldsymbol{\pi}) = p(\boldsymbol{\pi}) \prod_{j=1}^{d} p(x_j \mid \boldsymbol{pa}_j, \pi_j) \ . \tag{2.2}$$

We state the faithfulness condition in Assumption 2.2 in detail in terms of the d-separation in Definition B.1.

Assumption C.1 (Causal Faithfulness)**.** *Every conditional independence relation in the joint distribution $P_{X,\Pi}$ is entailed by a d-separation in $G_{\Pi,X}$, i.e.,*

$$X_S \perp\!\!\!\perp Y_S \mid Z_S \text{ in } P_{X,\Pi} \qquad \Longleftrightarrow \qquad X_S \perp_d Y_S \mid Z_S \text{ in } G_{\Pi,X}$$

for any pairwise disjoint subsets $X_S, Y_S, Z_S \subseteq (X \cup \Pi)$.

Assumption 2.3 (Causal Sufficiency)**.** *There are no unobserved confounders, i.e., there are no unobserved common causes over (X, Π) in $G_{\Pi,X}$.*

Assumption 2.4 (Independent Change). *The mechanism assignment variables are mutually independent,*

$$\Pi_j \perp\!\!\!\perp \Pi_k \qquad \text{for all } j \neq k \,.$$

Assumption 2.5 (Change Faithfulness under Π). *The variables Π correspond faithfully to changes in the underlying causal mechanisms,*

$$\pi_j^c \neq \pi_j^{c'} \iff P^c(X_j \mid Pa_j) \neq P^{c'}(X_j \mid Pa_j)$$

for any contexts c, c' and variable X_j.

Assumption 2.6 (Sparse Change). *Let C, C' be contexts drawn i.i.d. from a distribution, and let*

$$p_j^{\text{chng}} = Pr\left(P^C(X_j \mid Pa_j) \neq P^{C'}(X_j \mid Pa_j)\right) \,,$$

then we assume $0 < p_j^{\text{chng}} < \frac{1}{2}$ for each X_j.

Assumption 3.1 (Additive Noise Model). *For each X_j we assume that the structural equation model in Eq. (2.1) is an additive noise model according to Eq. (3.1).*

Assumption 3.2 (Sufficient Capacity). *For each X_j and each structural causal mechanism f_j^k in Eq. (3.1), there exists a function in the model class $\mathcal{H}$ that can express f_j^k.*

The latter assumes that the functional model is rich enough to represent the true conditionals. For GPs, existing (information consistency) results show that GPs asymptotically have sufficient capacity to represent a rich set of functions.

Remark C.1 Information Consistency of GPs *Under standard conditions on the GP kernel, the GP posterior predictive distribution p_{GP} converges to p^* in Kullback–Leibler (KL) divergence,*

$$\mathbb{E}\left[\text{KL}(p^*(y \mid x) \parallel p_{\text{GP}}(y \mid x))\right] \longrightarrow 0 \quad \text{as } n \to \infty \,,$$

provided the true regression function lies in the closure of the RKHS induced by the kernel (see, e.g., Kakade et al., 2005; Seeger et al., 2008).

C.2 Theoretical Results

Recovery up to Markov Equivalence

Theorem 3.1 (Identifiability of MECs and Mechanism Assignments)**.** *Let the causal Markov property (Assumption 2.1), the faithfulness assumptions (Assumptions 2.2-2.5) and sufficiency (Assumption 2.3) hold and assume an ANM (Assumption 3.1) and Sufficient Capacity (Assumption 3.2). Then for any minimizer G, Π of Eq. (3.4),*

$$Pr(G \sim G^*,\ \Pi = \Pi^*) \to 1$$

as the number of contexts m and observations n_c for all c tend to infinity.

Proof. Let (G^*, Π^*) denote the true causal model and (G, Π) an arbitrary candidate. We show that any minimizer of the score must satisfy $G \sim G^*$ and $\Pi = \Pi^*$ in the asymptotic regime.

Under Sufficient Capacity (Assumption 3.2), the true conditionals are representable within the model class. By information consistency of GP models (Remark C.1), the empirical fit term converges to its population counterpart. Hence, in the limit $n \to \infty$, the score separates into

- a data-fit term, minimized when the fitted conditionals coincide with the true conditionals, and
- a complexity term penalizing additional structure such as extra parents or redundant mechanisms.

In particular, any mismatch between fitted and true conditionals induces a strictly positive asymptotic penalty.

Case 1. Recovery of the graph up to MEC. Suppose $G \not\sim G^*$. Then G is not Markov equivalent to G^*, and therefore implies a different set of conditional independence relations.

By the Causal Markov property (Assumption 2.1) and Faithfulness (Assumption C.1), the true distribution satisfies exactly the conditional

independencies encoded by G^*. Hence the true joint distribution cannot be represented exactly by a factorization according to G.

Consequently, at least one local conditional in the factorization induced by G must differ from the corresponding true conditional under G^*. This induces a strictly positive asymptotic data-fit penalty.

If instead $G \sim G^*$, then the joint distribution admits an equivalent factorization under G, and no asymptotic fit penalty arises from the graph alone.

Thus every score minimizer must satisfy $G \sim G^*$.

Case 2. Recovery of the mechanism assignments. Fix any graph G with $G \sim G^*$ and consider a candidate partition Π.

If Π merges two distinct true mechanism groups for some variable X_j, then by Change Faithfulness (Assumption 2.5) the corresponding true conditionals differ. Any single fitted conditional over the merged group must therefore mismatch at least one true conditional, yielding strictly positive asymptotic fit loss.

If instead Π splits a true mechanism group, then all resulting groups share the same true conditional. The fit term is unchanged, but the model introduces additional mechanism descriptions, increasing the complexity term.

Thus every $\Pi \neq \Pi^*$ is asymptotically penalized, and every minimizer must satisfy $\Pi = \Pi^*$.

Combining the two cases, every asymptotic minimizer satisfies

$$G \sim G^* \qquad \text{and} \qquad \Pi = \Pi^* .$$

□

Recovery of Edge Orientations

The proof of Theorem 3.2 relies on the existence of *informative pairs* of contexts, where the mechanism of one variable changes while that of another remains fixed. First we show that Sparse Change together with Independent Change guarantees the existence of such informative pairs.

Lemma C.1 (Existence of Informative Context Pairs under Sparse Change)**.** *Under Independent Change (Assumption 2.4) and Sparse Change (Assumption 2.6), for any pair of variables X_j, X_k with $j \neq k$, the probability that two*

randomly drawn contexts c, c' satisfy

$$\pi_j^c \neq \pi_j^{c'} \qquad \text{and} \qquad \pi_k^c = \pi_k^{c'}$$

is strictly positive. Consequently, as $m \to \infty$, such pairs of contexts exist with probability tending to one.

Proof. Let $A_j(c, c')$ denote the event that the mechanism of X_j differs between contexts c and c',

$$A_j(c, c') := \{\pi_j^c \neq \pi_j^{c'}\}.$$

By Sparse Change (Assumption 2.6), we have

$$0 < \Pr(A_j) = p_j^{\text{chng}} < \tfrac{1}{2}, \quad \text{and hence} \quad \Pr(A_j^c) = 1 - p_j^{\text{chng}} > \tfrac{1}{2}.$$

By Independent Change (Assumption 2.4), the events A_j and A_k are independent for $j \neq k$. Therefore,

$$\Pr\left(A_j \cap A_k^c\right) = \Pr(A_j)\Pr(A_k^c) = p_j^{\text{chng}}\left(1 - p_k^{\text{chng}}\right) > 0.$$

Thus, with strictly positive probability, two contexts c, c' satisfy

$$\pi_j^c \neq \pi_j^{c'} \quad \text{and} \quad \pi_k^c = \pi_k^{c'}.$$

Since contexts are sampled independently, the probability that such a pair is observed tends to one as $m \to \infty$. □

Theorem 3.2 (Identifiability of Causal Directions)**.** *Let the assumptions from Chapter 2 (Assumptions 2.1-2.6) hold and assume an ANM (Assumption 3.1) and Sufficient Capacity (Assumption 3.2). Then*

$$Pr(G = G^*,\ \Pi = \Pi^*) \to 1$$

as the number of contexts m and observations n_c for all c tend to infinity.

Proof. By Theorem 3.1, it remains to distinguish DAGs within the MEC of G^*. Let $G \sim G^*$ with $G \neq G^*$, and consider an edge $X_j \to X_k$ in G^* that is reversed in G.

By Lemma C.1, as $m \to \infty$ there exist contexts c, c' such that

$$\pi_j^c \neq \pi_j^{c'} \qquad \text{but} \qquad \pi_k^c = \pi_k^{c'} .$$

In the true graph, $X_j \in Pa_k^{G^*}$, so conditioning on $Pa_k^{G^*}$ blocks the path $\Pi_j \to X_j \to X_k$. Hence

$$P^c(X_k \mid Pa_k^{G^*}) = P^{c'}(X_k \mid Pa_k^{G^*}) .$$

In the reversed graph G, the true cause X_j is omitted from the parent set of X_k, so Π_j remains d-connected to X_k given Pa_k^G. By Change Faithfulness (Assumption 2.5),

$$P^c(X_k \mid Pa_k^G) \neq P^{c'}(X_k \mid Pa_k^G) .$$

Thus, under G, the true partition Π_k^* is no longer sufficient to model the conditional of X_k: the candidate must either merge distinct conditionals, causing asymptotic fit loss, or introduce additional mechanism groups, increasing complexity. In either case, G has strictly larger score than G^*.

Repeating the argument for every incorrectly oriented edge shows

$$\Pr(G = G^*, \Pi = \Pi^*) \to 1 \qquad \text{as } m \to \infty \text{ and } n_c \to \infty \text{ for all } c.$$

□

DI Appendix for Chapter 4

D.1 Assumptions

Throughout this section, we make the same set of assumptions as in Chapter 3, namely Assumptions 2.1-3.2.

D.2 Theoretical Results

Theorem 4.1 (Consistency of Topological Search)**.** *Let the assumptions of Theorem 3.2 hold. Assume the oracle O is correct (cf. Theorem 4.2). That is, at each iteration it returns an unresolved node whose true parents are already present in the current graph. Then TOPIC recovers the true DAG G^* as the number of contexts m and the sample sizes n_c tend to infinity.*

Proof. The proof proceeds by induction over the iterations of the algorithm.

Induction hypothesis. After iteration $k-1$, assume that (i) the selected nodes form a valid partial topological order of G^*, and (ii) for all resolved nodes X with $T^{k-1}(X) \neq -1$, all outgoing true edges $(X, Y) \in \mathcal{E}^*$ are contained in $\mathcal{E}$.

Base case. In a DAG, there exists at least one source node X with $Pa_X^{G^*} = \varnothing$. Such a node is admissible, and the oracle returns it as the first node. This establishes the base case.

Inductive step. Let $X_k = O(G, T^{k-1})$ be the next selected node. By correctness of the oracle, X_k is admissible, i.e., $Pa^{G^*}_{X_k} \subseteq Pa^{G}_{X_k}$. Hence assigning $T^k(X_k) = k$ preserves a valid topological order.

Now consider any unresolved node Y. If $(X_k, Y) \in \mathcal{E}^*$ and $(X_k, Y) \notin \mathcal{E}$, then $X_k \in Pa^{G^*}_Y \setminus Pa^G_Y$. By the same mechanism-shift argument as in Theorem 3.2, adding a missing true parent strictly reduces the local score of Y, so the algorithm inserts the edge (X_k, Y). Therefore, after processing X_k, all outgoing true edges from X_k are included in $\mathcal{E}$, preserving the induction hypothesis.

Regarding the pruning step, once all true parents of a node Y are included in Pa^G_Y, any additional parent $Z \in Pa^G_Y \setminus Pa^{G^*}_Y$ is redundant. By score consistency, removing such a parent does not worsen the score, whereas removing a true parent would incur a loss. Hence pruning removes all and only spurious edges.

By induction, the algorithm visits all nodes in a valid topological order and recovers all true edges while removing all spurious ones. Therefore, $G = G^*$. □

For Theorem 4.2, we show the following first.

Lemma D.1 (Directional Gain). *Let the assumptions of Theorem 3.2 hold. Let T^{k-1} be a correct partial topological order and let G contain all outgoing true edges from resolved nodes. Let X be unresolved and let $Z \in Pa^{G^*}_X \setminus Pa^G_X$ be an unresolved true parent of X missing from the current graph. Then, as the number of contexts m and the sample sizes n_c tend to infinity,*

$$\Delta_{Z\to X} > \Delta_{X\to Z} .$$

In particular,

$$\Delta_{X\to Z} - \Delta_{Z\to X} < 0 .$$

Proof. Consider the two local score gains

$$\Delta_{Z\to X} = L\Big(X \mid Pa^G_X\Big) - L\Big(X \mid Pa^G_X \cup \{Z\}\Big)$$

and

$$\Delta_{X\to Z} = L\Big(Z \mid Pa^G_Z\Big) - L\Big(Z \mid Pa^G_Z \cup \{X\}\Big) .$$

We compare these two quantities through their local score improvements under edge addition.

Since $Z \in Pa_X^{G^*} \setminus Pa_X^G$, the true parent Z of X is omitted from the current parent set of X. Hence in the augmented graph $G^*_{\Pi^*,X}$, the directed path

$$\Pi_Z \to Z \to X$$

is active given Pa_X^G. By Faithfulness (Assumption C.1) and Change Faithfulness (Assumption 2.5), the conditional distribution $P^c(X \mid Pa_X^G)$ therefore depends on the mechanism value of Π_Z. Under Independent Change (Assumption 2.4), there exist contexts c, c' for which this variation is realized, so the model omitting the true parent Z incurs a strictly positive asymptotic likelihood penalty in the local score of X. Adding the edge $Z \to X$ removes this mismatch by conditioning on the true cause, and therefore yields a strictly positive asymptotic gain,

$$\Delta_{Z\to X} > 0 \,.$$

Conversely, adding the reverse edge $X \to Z$ does not insert a missing true parent into the conditional of Z. The conditional of Z is determined by its true parents and its own mechanism, and conditioning on its descendant X cannot remove the mechanism variation entering through Π_Z. Equivalently, the reverse orientation does not block the causal asymmetry exploited in Theorem 3.2: under the incorrect direction, the fitted conditional of Z would still require additional context variation not present in the true mechanism assignment. Thus the asymptotic fit improvement achieved by $X \to Z$ is strictly smaller than that of $Z \to X$.

Therefore,

$$\Delta_{Z\to X} > \Delta_{X\to Z} \,,$$

which proves the claim. □

Theorem 4.2 (Oracle Correctness). *Let the assumptions of Theorem 3.2 hold. Let T^{k-1} be a correct partial topological order and let G contain all outgoing true edges from resolved nodes. Then the oracle*

$$O\left(G, T^{k-1}\right) = \underset{X \text{ unresolved}}{arg\,max} \left(\min_{Y \text{ unresolved}, Y \neq X} \Delta_{X\to Y} - \Delta_{Y\to X} \right)$$

returns an admissible node X such that $Pa_X^{G^*} \subseteq Pa_X^G$.

Proof. Consider for each unresolved node X the oracle score

$$s(X) := \min_{Y \text{ unresolved}, Y \neq X} \left(\Delta_{X \to Y} - \Delta_{Y \to X}\right) .$$

We show that $s(X) \geq 0$ for every admissible unresolved node X, while $s(X) < 0$ for every non-admissible unresolved node X. Hence every maximizer of $s(X)$ is admissible.

We first note that at least one unresolved admissible node exists. Indeed, since T^{k-1} is a correct partial topological order, the unresolved node with smallest true topological position has all true parents already resolved. Because G contains all outgoing true edges from resolved nodes, all these parent edges are already present in G, and thus this node is admissible.

Case 1: X is admissible. Assume X is unresolved and admissible, i.e., $Pa_X^{G^*} \subseteq Pa_X^G$. Let $Y \neq X$ be any other unresolved node.

If Y is not a descendant of X, then Y is not a missing true parent of X. Hence adding the edge $Y \to X$ does not add a missing cause to the conditional of X. By the same score-consistency argument as in Theorem 3.1, the asymptotic likelihood term does not improve, while the model complexity increases, so $\Delta_{Y \to X} \leq 0$.

If Y is a descendant of X, then mechanism shifts in X propagate along the true directed path from X to Y. Since X is unresolved, these effects are not yet fully blocked in the current local model for Y. Thus, by Change Faithfulness (Assumption 2.5) and Independent Change (Assumption 2.4), there exist contexts c, c' such that

$$P^c(Y \mid Pa_Y^G) \neq P^{c'}(Y \mid Pa_Y^G) \quad \text{but} \quad P^c(Y \mid Pa_Y^G \cup \{X\}) = P^{c'}(Y \mid Pa_Y^G \cup \{X\}) .$$

Hence adding the edge $X \to Y$ results in a non-negative asymptotic gain $\Delta_{X \to Y} \geq 0$.

Combining both cases, for every unresolved $Y \neq X$ we obtain$\Delta_{X \to Y} - \Delta_{Y \to X} \geq 0$ and therefore

$$s(X) = \min_{Y \neq X} \left(\Delta_{X \to Y} - \Delta_{Y \to X}\right) \geq 0 .$$

Case 2: X is not admissible. Assume X is unresolved but not admissible. Then there exists a missing true parent $Z \in Pa_X^{G^*} \setminus Pa_X^G$.

Because T^{k-1} is correct, such a parent Z must also be unresolved. By Lemma D.1, $\Delta_{Z \to X} > \Delta_{X \to Z}$, hence $\Delta_{X \to Z} - \Delta_{Z \to X} < 0$. Since Z is one of the unresolved competitors in the minimum defining $s(X)$, it follows that

$$s(X) = \min_{Y \neq X} \left(\Delta_{X \to Y} - \Delta_{Y \to X}\right) \leq \Delta_{X \to Z} - \Delta_{Z \to X} < 0 .$$

Combining the cases, every admissible unresolved node satisfies $s(X) \geq 0$, whereas every non-admissible unresolved node satisfies $s(X) < 0$. Since at least one admissible unresolved node exists, the oracle

$$O\left(G, T^{k-1}\right) = \underset{X \text{ unresolved}}{\arg\max}\ s(X)$$

must return an admissible node X. Therefore, we have $Pa_X^{G^*} \subseteq Pa_X^G$. □

E | Appendix for Chapter 5

E.1 Assumptions

We adapt the previous assumptions, Assumptions 2.1-2.6 to the causal model G over the full set $\mathcal{V} = \mathcal{X} \cup \mathcal{L}$ including observed $\mathcal{X}$ and latents $\mathcal{L}$, where $G_{\mathcal{V},\Pi}$ denotes the augmented graph over $\mathcal{V}, \Pi$.

Assumption 5.2 (Causal Markov Property over $\mathcal{V}$)**.** *We assume the causal Markov property over $\mathcal{V}$, namely*

$$p(\boldsymbol{v}, \boldsymbol{\pi}) = p(\boldsymbol{\pi}) \prod_{j=1}^{d_v} p\left(v_j \mid \boldsymbol{pa}_j^{G_{\mathcal{V}}}, \pi_j\right) ,$$

Assumption E.1 (Causal Faithfulness over $\mathcal{V}$)**.** *All conditional independences in the joint distribution over $(\mathcal{V}, \Pi)$ correspond to $d-$separations in $G_{\mathcal{V},\Pi}$.*

Assumption E.2 (Causal Sufficiency, relaxed)**.** *There are no unobserved confounders besides $\mathcal{L}$, i.e., there are no unobserved common causes over $(\mathcal{V}, \Pi)$ in $G_{\mathcal{V},\Pi}$ and all unobserved common causes over $\mathcal{X}$ are included in the full graph.*

Assumption E.3 (Independent Change)**.** *The mechanism-selection variables are mutually independent,*

$$\Pi_j^* \perp\!\!\!\perp \Pi_k^* \qquad \text{for all } j \neq k.$$

Assumption E.4 (Change Faithfulness under Π). *The mechanism-selection variables Π faithfully encode changes in the underlying causal mechanisms f,*

$$\pi_j^c \neq \pi_j^{c'} \Leftrightarrow P^c(V_j \mid Pa_j) \neq P^{c'}(V_j \mid Pa_j)$$

for any contexts c, c' and variable j.

Assumption E.5 (Sparse Change). *Let C, C' be contexts drawn i.i.d. from a distribution, and let*

$$p_j^{\text{chng}} = Pr\Big(P^C(V_j \mid Pa_j) \neq P^{C'}(V_j \mid Pa_j)\Big) \ ,$$

then we assume $0 < p_j^{\text{chng}} < \frac{1}{2}$ for each V_j.

The following assumptions are specific to the chapter.

Assumption 5.1 (Exogeneity of Confounders). *We assume the latent variables $\mathcal{L}$ are exogenous confounders, i.e., there are no edges $X_j \to L_i$ nor edges $L_i \to L_{i'}$ for any $L_i, L_{i'}$ and X_j.*

Assumption 5.3 (Confounder Minimality). *For every subset $\mathcal{X}_S$ of size at least $|\mathcal{X}_S| \geq 4$, at most $2\,|\mathcal{X}_S|$ edges enter $\mathcal{X}_S$ from latent confounders L_i with at least three children in $\mathcal{X}_S$.*

In difference to the other chapters, as we do not assume a specific functional model class, we do not require capacity assumptions here (cf. Assumptions 3.1-3.2, Assumption 6.3). This is important as in the presence of the confounders, regression models may not be well-specified.

E.2 Theoretical Results

Pairwise Confounding

Lemma 5.1 (Detecting Confounded Variable Pairs). *Let X_j, X_k be a pair of variables that is either confounded through $L \in \mathcal{L}$, or unconfounded with causal relationship $X_j \to X_k$. Let Π_j and Π_k be the variables showing changes of X_j and X_k under $G_{\mathcal{X}}$, and t the score of $\mathrm{I}(\Pi_j, \Pi_k)$ in (5.3). Then we have*

$$\lim_{m\to\infty} Pr(t > q_{1-\alpha}) = \begin{cases} 1, & X_j, X_k \text{ confounded} \\ \alpha, & \text{otherwise}, \end{cases}$$

for any $\alpha > 0$, where $q_{1-\alpha}$ is the $1-\alpha$-quantile of the standard normal distribution.

Proof. Since t is asymptotically normal (Vinh et al., 2009), the first assertion follows directly.

For the converse statement, note that for two confounded variables X, Y, their partitions satisfy

$$\mathbb{E}I(\Pi_1, \Pi_2) \geq \frac{m}{2} H(p) \gg \mathbb{E}I(\Pi_1', \Pi_2')$$

where $H(p) = -p\log(p) - (1-p)\log(1-p)$ is the binary entropy of the probability p of two different contexts belonging to different sets of the partition as defined in Assumption E.5. Note that the relation $\frac{n_c}{2} H(p) \gg \mathbb{E}I(\Pi_1', \Pi_2')$ follows from the fact that $\lim_{m\to\infty} \frac{1}{m} I(\Pi_1', \Pi_2') = 0$ Pr-almost surely so that $\mathbb{E}I(\Pi_1', \Pi_2')$ cannot be extensive in m. Since $I(\Pi_1, \Pi_2)$ also concentrates around its mean, the result follows. □

E.2.1 Joint Confounding and Minimality

Theorem 5.1 (Detecting Confounded Variable Groups)**.** *Let $\mathcal{X}_S$ be a set of variables such that every pair $X_j, X_k \in \mathcal{X}_S$ is confounded, and let Π denote the mechanism-assignment variables induced by the observed distribution shifts under $G_\mathcal{X}$. Then $\mathcal{X}_S$ is jointly confounded by the same latent variable L if and only if*

$$\lim_{m\to\infty} Pr\big(\mathrm{T}(\Pi_j, \Pi_k, \Pi_l) < \mathrm{I}(\Pi_j, \Pi_k) + \mathrm{I}(\Pi_k, \Pi_l)\big) = \begin{cases} 1, & X_j, X_k, X_l \text{ jointly confounded} \\ 0, & \text{otherwise}, \end{cases}$$

which holds for each triple X_j, X_k, X_l in $\mathcal{X}_S$.

Proof. The condition $T(\Pi_i, \Pi_j, \Pi_k) < I(\Pi_i, \Pi_j) + I(\Pi_j, \Pi_k)$ is equivalent to $I(\Pi_j, \Pi_k \mid \Pi_i) < I(\Pi_j, \Pi_k)$, which is true if and only if the dependence between the clusterings is due to a shared source, which can only happen

due to joint confounding of more than two variables at a time. Now, let us assume that some set S of $s \geq 4$ pairwise confounded nodes satisfying this inequality, does not share the same confounder between all nodes. W.l.o.g. let us call these variables $X_1, \ldots, X_s$. Then the way for every triplet to have a shared confounder while using the fewest incoming latent edges is for three distinct confounders to affect the sets $\{X_1, \ldots, X_{s-1}\}$, $\{X_2, \ldots, X_s\}$, and $\{X_1, X_2, X_s\}$. This requires $2(s-1)+3 = 2s+1$ edges into the set $X_1, \ldots, X_s$ in contradiction with Assumption 5.3. □

E.2.2 Confounding with Unknown Causal Graph

Lemma 5.2 (Pairwise Consistency). *Let X_j, X_k be a variable pair confounded by a latent L, with true mechanism shifts Π_j^*, Π_k^* and distribution shifts Π_j, Π_k under the misdirected edge. Then there exists some $\rho > 0$ such that*

$$Pr\left(\mathrm{I}\left(\Pi_j^*, \Pi_k^*\right) < \mathrm{I}\left(\Pi_j, \Pi_k\right)\right) = 1 - O\left(e^{-\rho n_c}\right) .$$

Proof. Similar to Perry et al. (2022) we show more precisely that

$$\left(\mathrm{I}(\Pi_X^*, \Pi_Y^*) < \mathrm{I}(\Pi_X^*, \Pi_Y^*)\right) = 1 - O\left(\left(p + (1-p)(1-p+p^2)\right)^{\lfloor n_c/2 \rfloor}\right) ,$$

by splitting the contexts into pairs c_{2i}, c_{2i+1} and note we will get a wrong result if and only if for *all* these pairs of contexts we have that a change in the mechanism of Y does not introduce an additional change in the mechanism of $X \mid Y$.

The probability of this not happening for any one pair is given by three parts: either the mechanism of Z already changes between the environments (probability p), or it does not (probability $1-p$) *and* either Y does not change (probability $1-p$) or both X and Y change (probability p^2).

Since the changes between any two contexts c_{2i}, c_{2i+1} are independent of each other, the probability of this happening in *all* environments is therefore $(p+(1-p)(1-p+p^2))^{\lfloor n_c/2 \rfloor}$ and since $p+(1-p)(1-p+p^2) \leq 1$ as convex combination of 1 and $(1-p+p^2)$, the result follows. □

For Theorem 5.2 we first show the per-variable case in the following lemma.

Lemma E.1 (Recovering Parents). *Let X_j be a target variable and let G and G' be two graphs in the MEC of the marginal distribution $p(X)$. Assume that only one of the two graphs correctly recovers the parents of X_j, $Pa_j = Pa_j^*$ and $Pa'_j \neq Pa_j^*$, and further assume that the number of latent confounders affecting X_j plus spurious siblings is bounded by $\frac{log(0.5)}{log(1-p)}$. Then we have*

$$\begin{aligned} Pr(\mathrm{I}(\Pi_j, \{\Pi_k : X_k \in Pa_j\}) < MI(\Pi'_j, \{\Pi'_k : X_k \in Pa'_k\})) \\ = 1 - O(e^{-\rho m}) \end{aligned}$$

Proof. More precisely, we will show that

$$\begin{aligned} &\Pr(\mathrm{I}(\Pi_j, \{\Pi_k : X_k \in Pa_j\}) < \mathrm{I}(\Pi'_j, \{\Pi'_k : k \in Pa'_j\})) \\ &= 1 - O\left(\left((1-(1-p)^r) + (1-p)^r(1-p+p^2)\right)^{m/2}\right), \end{aligned}$$

where we recall that m is the number of contexts, and furthermore r is the number of latent parents of X_j plus the number of other variables with which it is pairwise confounded. These variables are precisely those which could make us not detect changes between two context, just as in the previous proof changes in the mechanism of L between contexts could prevent us from detecting changes in the mechanisms of X or Y.

To this end, note that if $Pa'_j \neq Pa_j^*$ then there exists either a variable in Pa_j^* that is missing from Pa'_j or that is a child of X_j in Pa'_j. In either case, additional joint distribution shifts are introduced between X_j and these variables and therefore the mutual information increases. This increase in mutual information is guaranteed by the fact that $r \leq \frac{\log(0.5)}{\log(1-p)}$, so that the probability of mechanism between shifts in X_j is less than 0.5. □

Summing up over all variables gives us the following consistency of the entire causal ordering over X.

Theorem 5.2 (Consistency). *Let $G_{\mathcal{V}}$ be the true graph over $\mathcal{V}$ and $G_{\mathcal{X}}$ be the corresponding subgraph over $\mathcal{X}$. Assume further that for each X_j, the number of latent confounders plus spurious siblings in the MEC of the observed distribution $P_{\mathcal{X}}$ is bounded by $\frac{log(0.5)}{log(1-p)}$.*

Then with high probability, $G_{\mathcal{X}}$ is the unique minimum of total correlation,

$$Pr\Big(\underset{G,\Pi \text{ under } G}{arg\,min}\ T(\Pi_1, \ldots, \Pi_d) = \{G_{\mathcal{X}}\}\Big) = 1 - O(e^{-\rho m}),$$

where the minimization ranges over graphs G and the corresponding variables Π under this graph, i.e., $\Pi = \{\Pi_1, \ldots, \Pi_d\}$ show distribution shifts of each conditional $P(X_j \mid Pa_j)$ in G.

Proof. Let d be the number of observed variables, m the number of contexts, p the change probability in Assumption E.5 and r be an upper bound $r = \max\{r_j \mid j = 1, \ldots, d\}$ from the above Lemma E.1, where we recall that r_j is the number of latent parents of X_j plus the number of other variables with which it is pairwise confounded. Then we specifically show that

$$\begin{aligned}&\Pr\left(\arg\min_G T(\Pi_1, \ldots, \Pi_d) = \{G_X{}^*\}\right)\\ &= 1 - O\left(\frac{d^2(d-1)}{2}\left((1-(1-p)^r) + (1-p)^r(1-p+p^2)\right)^{m/2}\right).\end{aligned}$$

To this end, let us assume that the true causal ordering over X is given by $X_1 \leq \cdots \leq X_d$. Then note that by construction

$$T(\Pi_1, \ldots, \Pi_d) = \sum_j \mathrm{I}(\Pi_j, \{\Pi_k : k \in Pa_j\}),$$

so that the inside of our statement is simply the sum of all terms in Lemma E.1. Therefore, the total correlation being the unique minimum if the above proposition holds for all j and when compared against *any* other graph G, resulting in the union bound above. □

F | Appendix for Chapter 6

F.1 Assumptions

We adapt the assumptions from Chapter 3 to the temporal setting, assuming the temporal analogues of the causal Markov property, faithfulness, sufficient capacity, independent change, sparse change, and change faithfulness.

The assumptions regarding mechanism shifts extend the multi-context setting by treating context-regime pairs (c, r) as the indexing units of mechanism changes, so we restate them below.

Assumption F.1 (Independent Change). *The mechanism-selection variables are mutually independent across variables, i.e.,*

$$\Pi_j \perp\!\!\!\perp \Pi_k \qquad \text{for all } j \neq k\,,$$

where Π_j denotes the random variable governing the mechanism of X_j across context-regime pairs.

Assumption F.2 (Sparse Change). *Let (C, R) and (C', R') be context-regime pairs drawn i.i.d. from a distribution, and define*

$$p_j^{\text{chng}} = Pr\Big(P^{(C,R)}(X_j \mid Pa_j) \neq P^{(C',R')}(X_j \mid Pa_j)\Big)\,.$$

Then we assume

$$0 < p_j^{\text{chng}} < \tfrac{1}{2} \qquad \text{for each } X_j\,.$$

Assumption F.3 (Change Faithfulness under Π)**.** *The mechanism-selection variables Π faithfully encode changes in the underlying causal mechanisms f, in the sense that for any variable X_j and any context-regime pairs (c, r) and (c', r'),*

$$\pi_j^{(c,r)} \neq \pi_j^{(c',r')} \iff P^{(c,r)}(X_j \mid Pa_j) \neq P^{(c',r')}(X_j \mid Pa_j)\,.$$

The following assumptions are specific to the chapter.

Assumption 6.2 (Additive Noise Model)**.** *For each X_j we assume that the structural equation model is a temporal additive noise model according to Eq. (6.1).*

Assumption 6.3 (Sufficient Regime Coverage)**.** *Each true mechanism receives asymptotically many observations, i.e., for all j and k,*

$$\left|\left\{(c, r, t) \mid t \in I_r,\ \pi_j^{(c,r)} = k\right\}\right| \to \infty \quad \text{as } m, n \to \infty,$$

and each true regime I_r satisfies $|I_r| \to \infty$.

F.2 Theoretical Results

Theorem 6.1 (Consistency)**.** *Let the assumptions from Section 6.2 hold together with the temporal analogues of the assumptions from Chapter 3. Then for any minimizers $(G, \mathcal{T}, \Pi)$ of Eq. (6.2),*

$$Pr(G = G^*,\ \mathcal{T} = \mathcal{T}^*,\ \Pi = \Pi^*) \to 1 \qquad \text{as } m \to \infty \text{ and } n \to \infty.$$

Proof. Let $(G^*, \mathcal{T}^*, \Pi^*)$ denote the true temporal causal model and let $(G, \mathcal{T}, \Pi)$ be an arbitrary candidate. We show that any minimizer of the score must satisfy $G = G^*, \mathcal{T} = \mathcal{T}^*, \Pi = \Pi^*$ asymptotically.

Under Sufficient Capacity (Assumption 3.2) and the temporal additive noise model (Assumption 6.2), the true conditionals are representable within the model class. By information consistency of GP models (Remark C.1), the empirical fit term converges to its population counterpart. By Sufficient Regime Coverage (Assumption 6.3), each true mechanism

receives asymptotically many observations, so every pooled local score is estimated consistently. Hence, in the limit $m, n \to \infty$, the score separates into

- a data-fit term, minimized when the fitted conditionals coincide with the true conditionals, and
- a complexity term penalizing unnecessary graph structure, redundant changepoints, and redundant mechanism distinctions.

In particular, any mismatch between fitted and true conditionals induces a strictly positive asymptotic penalty.

Case 1. Recovery of the graph and mechanism assignments for fixed changepoints. Fix the true changepoints $\mathcal{T}^*$. Then the temporal score in Eq. (6.2) has the same structure as the multi-context score in Eq. (3.3), with context-regime pairs (c, r) taking the role of contexts. Therefore, the same arguments as in the proofs of Theorem 3.1 and Theorem 3.2 apply, with context-regime pairs (c, r) taking the role of contexts and relying again on Independent Change (Assumption F.1).

If $G \neq G^*$, then G either encodes incorrect conditional independence relations or assigns an incorrect causal direction, which leads to a strictly positive asymptotic fit penalty or increased description length. If $\Pi \neq \Pi^*$, then either distinct true mechanisms are merged, which worsens fit, or a true mechanism is split unnecessarily, which increases complexity without improving fit.

Thus, for fixed $\mathcal{T}^*$, every asymptotic score minimizer satisfies

$$G = G^* \qquad \text{and} \qquad \Pi = \Pi^* \, .$$

Case 2. Recovery of the changepoints. Now consider the segmentation. Suppose first that $\mathcal{T} \neq \mathcal{T}^*$ omits at least one true changepoint. Then there exists some estimated regime that contains observations from two distinct true regimes. By Change Faithfulness (Assumption F.3), for at least one variable X_j the corresponding conditional distributions differ across these two true regimes. Hence the candidate must fit a single pooled mechanism to observations generated by different true conditionals, which induces a strictly positive asymptotic fit penalty.

Conversely, suppose $\mathcal{T}$ introduces a spurious changepoint. Then some homogeneous true regime is split into two candidate regimes whose observations are generated by the same true conditional. In this case, the

asymptotic fit term does not improve, since the same true mechanism governs both parts, but the description length increases because the candidate introduces unnecessary mechanism distinctions.

Therefore every candidate $\mathcal{T} \neq \mathcal{T}^*$ is asymptotically penalized, and every score minimizer must satisfy $\mathcal{T} = \mathcal{T}^*$.

Combining the two cases, every asymptotic minimizer satisfies

$$G = G^*, \qquad \mathcal{T} = \mathcal{T}^*, \qquad \Pi = \Pi^* .$$

□

GI Appendix for Chapter 7

G.1 Assumptions

Assumption 7.1 (Causal Markov Property)**.** *We assume the causal Markov property over $\mathcal{X}$, where*

$$p(\boldsymbol{x}) = \prod_{X_j \in \mathcal{X}} p^{\mathrm{MLR}}(x, \boldsymbol{pa}_j; \boldsymbol{B}, \boldsymbol{\gamma}, \sigma^2) \,, \tag{7.2}$$

Assumption 7.2 (Exogeneity of $\mathcal{Z}$)**.** *We assume the latent variables $\mathcal{Z}$ are exogenous, i.e., there are no edges $X_j \to Z_i$ nor edges $Z_i \to Z_l$ for any Z_i, Z_l, X_j.*

Assumption 7.3 (Oracle BIC)**.** *We assume the availability of consistent MLE estimates to obtain* $\mathrm{BIC}^{ML}_{\hat{\mathcal{Z}}}(\mathcal{H})$.

In addition, we assume the standard regularity conditions required for consistency of BIC-based score comparison, as well as faithfulness and causal sufficiency with respect to the graphical model under consideration (cf. Assumption 2.2, Assumption 2.3).

G.2 Score Consistency

We recall the notion of a consistent scoring criterion, following Chickering (2002).

We write $P_{\mathcal{X}} \in I_G$ if the conditional independencies implied by G hold in the distribution $P_{\mathcal{X}}$.

Definition G.1 (Consistent Scoring Criterion)**.** *Given data $\boldsymbol{X}$ of size n sampled from distribution $P_{\mathcal{X}}$, a scoring criterion S is consistent if for any two graph hypotheses $\widehat{G}_1, \widehat{G}_2$*

1. *$P_{\mathcal{X}} \in I_{\widehat{G}_1}$ and $P_{\mathcal{X}} \notin I_{\widehat{G}_2}$ implies*

$$S(\widehat{G}_1; \boldsymbol{X}) > S(\widehat{G}_2; \boldsymbol{X}),$$

2. *$P_{\mathcal{X}} \in I_{\widehat{G}_1} \cap I_{\widehat{G}_2}$ and $\widehat{G}_1$ has fewer parameters than $\widehat{G}_2$ implies*

$$S(\widehat{G}_1; \boldsymbol{X}) > S(\widehat{G}_2; \boldsymbol{X}).$$

G.3 Theoretical Results

Lemma 7.1 (Non-Gaussianity of Reverse Conditionals)**.** *Let Y and $\mathcal{X}$ be random variables such that $Y \mid \mathcal{X} \sim \mathrm{MLR}(\boldsymbol{B}, \boldsymbol{\gamma}, \sigma^2)$, with parameters $\boldsymbol{\theta} = (\boldsymbol{B}, \boldsymbol{\gamma}, \sigma^2) \in \cup_{K \geq 2} \Theta_K$. Assume that $\gamma_k \neq 0$ for all k, and that the prior over $\boldsymbol{B}$ is absolutely continuous with full support. Then the conditional distribution $Y \mid \mathcal{X}$ is almost surely non-Gaussian.*

Proof. Fix $K \geq 2$. Consider the conditional density

$$p^{\mathrm{MLR}}(y, \boldsymbol{x}; \boldsymbol{B}, \boldsymbol{\gamma}, \sigma^2) = \sum_{k=1}^{K} \gamma_k \, p^{N}(x; \boldsymbol{\beta}_k{}^{\top} y, \sigma^2).$$

We show that this mixture can coincide with a single Gaussian density only on a parameter set of measure zero.

For such a collapse to occur, one of the following degenerate cases must hold:

(i) only one mixture component is active, i.e. $\gamma_k = 1$ for some $k \in \{1, \ldots, K\}$;

(ii) all components have identical regression coefficients, i.e. $\boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_K$.

For (i), the admissible mixing weights lie at the vertices of the simplex $\mathcal{S}^{K-1}$. Since this is a finite set, it has Lebesgue measure zero inside $\mathcal{S}^{K-1}$.

For (ii), the admissible coefficient matrices form the subset of $\mathbb{R}^{d\times K}$ whose columns are all equal. This is a proper linear subspace of $\mathbb{R}^{d\times K}$, and therefore has Lebesgue measure zero.

Hence the parameter values for which the mixture collapses to a single Gaussian form a union of two measure-zero sets, and thus themselves have measure zero. Since the parameter distribution is assumed to be absolutely continuous with respect to Lebesgue measure, this set is assigned probability zero.

Therefore, almost surely, the mixture model does not reduce to a Gaussian, and the reverse conditional $Y \mid X$ is non-Gaussian almost surely. □

Theorem 7.1 (Consistency of Latent-Aware BIC). *Under Assumptions 7.1, 7.2, and 7.3, together with standard regularity conditions, the latent-aware score* $\mathrm{BIC}^{ML}_{\hat{Z}}$ *is a consistent scoring criterion. Consequently, standard score-based search procedures recover the true Markov equivalence class almost surely as* $n \to \infty$.

Proof. We verify the conditions of Definition G.1.

Non-Gaussianity. By Lemma 7.1, the conditional distributions $X_j \mid Pa_j$ are almost surely non-Gaussian. Hence, the model avoids the classical non-identifiability of linear Gaussian structural equation models, where different graph structures can induce the same observational distribution.

Likelihood separation. Consider any incorrect graph hypothesis $\widehat{G} \neq G^*$. Then either: (i) $\widehat{G}$ encodes incorrect conditional independence relations, or (ii) $\widehat{G}$ assigns an incorrect edge orientation.

In both cases, the corresponding conditional distributions differ from the true ones. By standard likelihood theory, this induces a strictly worse asymptotic log-likelihood compared to the true model.

Score consistency. The score $\mathrm{BIC}^{\mathrm{ML}}_{\hat{Z}}$ decomposes into a sum of local terms and consists of a likelihood component and a complexity penalty. For any incorrect model, either the likelihood term is strictly worse (by Step 2), or the model introduces unnecessary parameters, which increases the penalty term.

Therefore, $\mathrm{BIC}^{\mathrm{ML}}_{\hat{Z}}$ satisfies both conditions of Definition G.1.

From the above it follows that $\mathrm{BIC}^{\mathrm{ML}}_{\hat{Z}}$ is a consistent scoring criterion, and score-based search procedures recover the true Markov equivalence class almost surely as $n \to \infty$. □

HI List of Case Studies

Chapter	Case Study	References
1	Population Paradox	Public Health England (2021)
3	Population Paradox (revisited)	*toy example*
4	Pathway Problem	Dibaeinia & Sinha (2020)
5	Confounding Case Study	Mooij et al. (2020)
6	Time Trial	Cornes et al. (2018)
7	Subpopulation Paradox	Chang et al. (2020)
8	Cocoa Controlled Trial	Chan (2008)
8	Cocoa Correlation	Messerli (2012)

Table H.1 Case studies throughout the chapters.

Chapter	Case Study	Usage Examples
1	Population Paradox	`demo/ch1_example_simpson.ipynb`
6	Time Trial	`demo/ch6_realworld_river.ipynb`
7	Population Paradox	`demo/ch7_example_colon.ipynb`

Table H.2 Jupyter notebooks showing the case studies, maintained online at `github.com/srhmm/thesis`.

Index

Bibliography

Agrawal, R., Squires, C., Prasad, N., and Uhler, C. The decamfounder: Nonlinear causal discovery in the presence of hidden variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2023.

Ahuja, K., Mahajan, D., Wang, Y., and Bengio, Y. Interventional causal representation learning. In *International Conference on Machine Learning (ICML)*, 2023.

Angus, D. C. and Chang, C.-C. H. Heterogeneity of treatment effect: Estimating how the effects of interventions vary across individuals. *JAMA*, 2021.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Assaad, C., Devijver, E., and Gaussier, E. Survey and evaluation of causal discovery methods for time series. *Journal of Artificial Intelligence Research*, 2022.

Baldocchi, D. Measuring fluxes of trace gases and energy between ecosystems and the atmosphere – the state and future of the eddy covariance method. *Global Change Biology*, 2014.

BBC News. Does chocolate make you clever?, 2012. URL `https://www.bbc.com/news/magazine-20356613`.

Bhattacharya, R., Nagarajan, T., Malinsky, D., and Shpitser, I. Differentiable causal discovery under unmeasured confounding. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

Bishop, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, 2006.

Boeken, P., de Kroon, N., de Jong, M., Mooij, J. M., and Zoeter, O. Correcting for selection bias and missing response in regression using privileged information. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.

Bollen, K. A. Structural equation models with observed variables. *Structural equations with latent variables*, 1989.

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

Brouillard, P., Squires, C., Wahl, J., Kording, K. P., Sachs, K., Drouin, A., and Sridhar, D. The landscape of causal discovery data: Grounding causal discovery in real-world applications. *Conference on Causal Learning and Reasoning (CLeaR)*, 2024.

Bühlmann, P., Peters, J., and Ernest, J. Cam: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 2014.

Cancer Genome Atlas Research Network, Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J. M. The cancer genome atlas pancancer analysis project. *Nat Genet*, 2013.

Chan, K. A clinical trial gone awry: The chocolate happiness undergoing more pleasantness (chump) study. *Canadian Medical Association Journal (CMAJ)*, 2008.

Chang, W., Wan, C., Yu, C., Yao, W., Zhang, C., and Cao, S. Robmixreg: an r package for robust, flexible and high dimensional mixture regression. *bioRxiv*, 2020.

Chickering, D. M. Optimal structure identification with greedy search. *J. Mach. Learn. Res.*, 2002.

Cornes, R., Schrier, G., Van den Besselaar, E., and Jones, P. An ensemble version of the e-obs temperature and precipitation data sets. *Journal of Geophysical Research Atmospheres*, 2018.

Dai, H., Ng, I., Sun, J., Tang, Z., Luo, G., Dong, X., Spirtes, P., and Zhang, K. When selection meets intervention: Additional complexities in causal discovery. In *International Conference on Learning Representations (ICLR)*, 2025.

Dawid, A. P. Beware of the dag! In *Proceedings of Workshop on Causality: Objectives and Assessment at NIPS 2008*, 2010.

Dibaeinia, P. and Sinha, S. Sergio: A single-cell expression simulator guided by gene regulatory networks. *Cell systems*, 2020.

Dominguez, A. A., Lim, W. A., and Qi, L. S. Beyond editing: Repurposing crispr–cas9 for precision genome regulation and interrogation. *Nature Reviews Molecular Cell Biology*, 2016.

Donath, W. E. and Hoffman, A. J. Algorithms for partitioning of graphs and computer logic based on eigenvectors of connection matrices. *IBM Technical Disclosure Bulletin*, 1972.

Eberhardt, F. and Scheines, R. Interventions and causal inference. *Philos. Sci.*, 2007.

Elidan, G., Lotner, N., Friedman, N., and Koller, D. Discovering hidden variables: A structure-based approach. *Neural Information Processing Systems (NeurIPS)*, 2000.

Fennell, L., Dumenil, T., Wockner, L., Hartel, G., Nones, K., Bond, C., Borowsky, J., Liu, C., McKeone, D., Bowdler, L., Montgomery, G., Klein, K., Hoffmann, I., Patch, A.-M., Kazakoff, S., Pearson, J., Waddell, N., Wirapati, P., Lochhead, P., Imamura, Y., Ogino, S., Shao, R., Tejpar, S., Leggett, B., and Whitehall, V. Integrative genome-scale dna methylation analysis of a large and unselected cohort reveals 5 distinct subtypes of colorectal adenocarcinomas. *Cellular and Molecular Gastroenterology and Hepatology*, 2019.

Gao, S., Addanki, R., Yu, T., Rossi, R. A., and Kocaoglu, M. Causal discovery in semi-stationary time series. In *Neural Information Processing Systems (NeurIPS)*, 2024.

Gao, S., Addanki, R., Yu, T., Rossi, R. A., and Kocaoglu, M. Causal discovery-driven change point detection in time series. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2025.

Ghassami, A., Kiyavash, N., Huang, B., and Zhang, K. Multi-domain causal structure learning in linear systems. In *Neural Information Processing Systems (NeurIPS)*, 2018.

Gong, W., Jennings, J., Zhang, C., and Pawlowski, N. Rhino: Deep causal temporal relationship learning with history-dependent noise. *arXiv preprint arxiv:2210.14706*, 2022.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A Kernel Two-Sample Test. *J. Mach. Learn. Res.*, 2012.

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

Günther, W., Miersch, P., Ninad, U., and Runge, J. Clustering of causal graphs to explore drivers of river discharge. *Environmental Data Science*, 2023a.

Günther, W., Ninad, U., and Runge, J. Causal discovery for time series from multiple datasets with latent contexts. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2023b.

Guo, S., Tóth, V., Schölkopf, B., and Huszár, F. Causal de finetti: On the identification of invariant causal structure in exchangeable data. *arXiv preprint arXiv:2203.15756*, 2022.

Hägele, A., Rothfuss, J., Lorch, L., Somnath, V. R., Schölkopf, B., and Krause, A. Bacadi: Bayesian causal discovery with unknown interventions. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.

Hauser, A. and Bühlmann, P. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *J. Mach. Learn. Res.*, 2012.

Hauser, A. and Bühlmann, P. Two optimal strategies for active learning of causal models from interventional data. *International Journal of Approximate Reasoning*, 2014.

Heinze-Deml, C., Peters, J., and Meinshausen, N. Invariant causal prediction for nonlinear models. *J. Causal Inf.*, 2018.

Hennig, C. Identifiability of models for clusterwise linear regression. *J. Classification*, 2000.

Hoyer, P., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. Nonlinear causal discovery with additive noise models. In *Neural Information Processing Systems (NeurIPS)*, 2009.

Hu, S., Chen, Z., Nia, V. P., Chan, L., and Geng, Y. Causal inference and mechanism clustering of a mixture of additive noise models. In *Neural Information Processing Systems (NeurIPS)*, 2018.

Huang, B., Zhang, K., Lin, Y., Schölkopf, B., and Glymour, C. Generalized score functions for causal discovery. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2018.

Huang, B., Zhang, K., Zhang, J., Ramsey, J., Sanchez-Romero, R., Glymour, C., and Schölkopf, B. Causal discovery from heterogeneous/nonstationary data. *J. Mach. Learn. Res.*, 2020.

Huang, B., Low, C. J. H., Xie, F., Glymour, C., and Zhang, K. Latent hierarchical causal structure discovery with rank constraints. *Neural Information Processing Systems (NeurIPS)*, 2022.

Hyvärinen, A., Zhang, K., Shimizu, S., and Hoyer, P. O. Estimation of a structural vector autoregression model using non-gaussianity. *J. Mach. Learn. Res.*, 2010.

Immer, A., Palumbo, E., Marx, A., and Vogt, J. Effective bayesian heteroscedastic regression with deep neural networks. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Neural Information Processing Systems (NeurIPS)*, 2023.

Janzing, D. and Schölkopf, B. Causal inference using the algorithmic markov condition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 2010.

Janzing, D., Balduzzi, D., Grosse-Wentrup, M., and Schölkopf, B. Quantifying causal influences. *The Annals of Statistics*, 2012.

Kakade, S. M., Seeger, M. W., and Foster, D. P. Worst-case bounds for gaussian process models. In *Neural Information Processing Systems (NeurIPS)*, 2005.

Kaltenpoth, D. *Don't confound yourself: Causality from biased data*. PhD thesis, Saarländische Universitäts-und Landesbibliothek, 2024.

Kaltenpoth, D. and Vreeken, J. Causal Discovery with Hidden Confounders using the Algorithmic Markov Condition. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2023a.

Kaltenpoth, D. and Vreeken, J. Nonlinear Causal Discovery with Latent Confounders. In *International Conference on Machine Learning (ICML)*, 2023b.

Kim, K., Kim, J., and Kennedy, E. H. Causal k-means clustering. *arXiv preprint arxiv:2405.03083*, 2024.

Knopfmacher, A. and Mays, M. A survey of factorization counting functions. *International Journal of Number Theory*, 2005.

Kocaoglu, M., Dimakis, A. G., Vishwanath, S., and Hassibi, B. Entropic causal inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017.

Kolch, W. Kolch, w. coordinating erk/mapk signalling through scaffolds and inhibitors. *Nature reviews. Molecular cell biology*, 2005.

Koller, D. and Friedman, N. *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, 2009.

Kolmogorov, A. N. Three approaches to the quantitative definition of information. *International Journal of Computer Mathematics*, 1968.

Kraemer, G., Camps-Valls, G., Reichstein, M., and Mahecha, M. D. Summarizing the state of the terrestrial biosphere in few dimensions. *Biogeosciences*, 2020.

Krich, C., Migliavacca, M., Miralles, D. G., Kraemer, G., El-Madany, T. S., Reichstein, M., Runge, J., and Mahecha, M. D. Functional convergence of biosphere–atmosphere interactions in response to meteorological conditions. *Biogeosciences*, 2021.

Kumar, A., Shiragur, K., and Uhler, C. Learning mixtures of unknown causal interventions. In *Neural Information Processing Systems (NeurIPS)*, 2024.

Lauritzen, S. *Graphical Models*. Oxford Statistical Science Series. Clarendon Press, 1996.

Lemeire, J. and Janzing, D. Replacing causal faithfulness with algorithmic independence of conditionals. *Minds and Machines*, 2013.

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

Lorch, L., Krause, A., and Schölkopf, B. Causal modeling with stationary diffusions. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2024.

Magliacane, S., van Ommen, T., Claassen, T., Bongers, S., Versteeg, P., and Mooij, J. M. Domain adaptation by using causal inference to predict invariant conditional distributions. In *Neural Information Processing Systems (NeurIPS)*, 2018.

Malinsky, D. and Spirtes, P. Causal structure learning from multivariate time series in settings with unmeasured confounding. In *Proceedings of 2018 ACM SIGKDD Workshop on Causal Disocvery*, 2018.

Mameche, S., Kaltenpoth, D., and Vreeken, J. Discovering invariant and changing mechanisms from data. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2022.

Mameche, S., Kaltenpoth, D., and Vreeken, J. Learning causal models under independent changes. *Neural Information Processing Systems (NeurIPS)*, 2023.

Mameche, S., Vreeken, J., and Kaltenpoth, D. Identifying confounding from causal mechanism shifts. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2024.

Mameche, S., Cornanguer, L., Ninad, U., and Vreeken, J. Spacetime: Causal discovery from non-stationary time series. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025a.

Mameche, S., Kalofolias, J., and Vreeken, J. Causal mixture models: Characterization and discovery. *Neural Information Processing Systems (NeurIPS)*, 2025b.

Marx, A. and Vreeken, J. Testing conditional independence on discrete data using stochastic complexity. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019a.

Marx, A. and Vreeken, J. Telling cause from effect by local and global regression. *Knowledge and Information Systems*, 2019b.

Marx, A. and Vreeken, J. Identifiability of cause and effect using regularized regression. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2019c.

Maurage, P., Heeren, A., and Pesenti, M. Does chocolate consumption really boost nobel award chances? the peril of over-interpreting correlations in health studies. *The Journal of nutrition*, 2013.

Mazaheri, B., Gordon, S., Rabani, Y., and Schulman, L. Causal discovery under latent class confounding. *arXiv preprint arxiv:2311.07454*, 2023.

Mazaheri, B., Squires, C., and Uhler, C. Synthetic potential outcomes and causal mixture identifiability. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2025a.

Mazaheri, B., Zhang, J., and Uhler, C. Faithfulness and intervention-only causal discovery. In *ICML 2025 Workshop on Scaling Up Intervention Models*, 2025b.

McLachlan, G. J. and Peel, D. *Finite Mixture Models*. Probability and Statistics – Applied Probability and Statistics Section. Wiley, 2000.

Meek, C. *Graphical Models: Selecting causal and statistical models*. PhD thesis, Carnegie Mellon University, 1997.

Messerli, F. H. Chocolate consumption, cognitive function, and nobel laureates. *New England Journal of Medicine*, 2012.

Mian, O., Marx, A., and Vreeken, J. Discovering fully oriented causal networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

Mian, O., Mameche, S., and Vreeken, J. Learning causal networks from episodic data. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2024.

Montagna, F., Noceti, N., Rosasco, L., Zhang, K., and Locatello, F. Causal discovery with score matching on additive models with arbitrary noise. In *Conference on Causal Learning and Reasoning (CLeaR)*, 2023.

Mooij, J., Magliacane, S., and Claassen, T. Joint causal inference from multiple contexts. *J. Mach. Learn. Res.*, 2020.

Mooij, J. M., Janzing, D., and Schölkopf, B. From ordinary differential equations to structural causal models: the deterministic case. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 440–448, 2013.

Ormaniec, W., Sussex, S., Lorch, L., Schölkopf, B., and Krause, A. Standardizing structural causal models. In *International Conference on Learning Representations (ICLR)*, 2025.

Pamfil, R., Sriwattanaworachai, N., Desai, S., Pilgerstorfer, P., Georgatzis, K., Beaumont, P., and Aragam, B. Dynotears: Structure learning from time-series data. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

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

Perry, R., Kügelgen, J. V., and Schölkopf, B. Causal discovery in heterogeneous environments under the sparse mechanism shift hypothesis. In *Neural Information Processing Systems (NeurIPS)*, 2022.

Peters, J. and Bühlmann, P. Structural intervention distance for evaluating causal graphs. *Neural Comput.*, 2015.

Peters, J., Janzing, D., and Schölkopf, B. Causal inference on time series using restricted structural equation models. In *Neural Information Processing Systems (NeurIPS)*, 2013.

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

Peters, J., Bühlmann, P., and Meinshausen, N. Causal inference using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016.

Peters, J., Janzing, D., and Schölkopf, B. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017.

Pezzullo, A. Sources of bias in covid-19 infection fatality rate estimation. *European Journal of Public Health*, 2022.

Public Health England. Sars-cov-2 variants of concern and variants under investigation in england: Technical briefing 20. Technical report, Public Health England, 2021. URL https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1009243/Technical_Briefing_20.pdf.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In *Neural Information Processing Systems (NeurIPS)*, 2007.

Rahmani, A. and Frossard, P. Causal temporal regime structure learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2025.

Rasmussen, C. E. and Williams, C. K. I. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

Reisach, A., Tami, M., Seiler, C., Chambaz, A., and Weichwald, S. A scale-invariant sorting criterion to find a causal order in additive noise models. *Neural Information Processing Systems (NeurIPS)*, 2023.

Richardson, T. S., Evans, R. J., Robins, J. M., and Shpitser, I. Nested markov properties for acyclic directed mixed graphs. *The Annals of Statistics*, 2023.

Rolland, P., Cevher, V., Kleindessner, M., Russell, C., Janzing, D., Schölkopf, B., and Locatello, F. Score matching enables causal discovery of nonlinear additive noise models. In *International Conference on Machine Learning (ICML)*, 2022.

Rothenhäusler, D., Meinshausen, N., Bühlmann, P., and Peters, J. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2021.

Runge, J. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2020.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D., and Nolan, G. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 2005.

Saeed, B., Panigrahi, S., and Uhler, C. Causal structure discovery from distributions arising from mixtures of dags. In *International Conference on Machine Learning (ICML)*, 2020.

Saggioro, E., de Wiljes, J., Kretschmer, M., and Runge, J. Reconstructing regime-dependent causal relationships from observational time series. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 2020.

Sanchez, P. and Tsaftaris, S. A. Diffusion causal models for counterfactual estimation. In *Conference on Causal Learning and Reasoning (CLeaR)*, 2022.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. On causal and anticausal learning. In *International Conference on Machine Learning (ICML)*, 2012.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. Toward causal representation learning. *Proceedings of the IEEE*, 2021.

Seeger, M. W., Kakade, S. M., and Foster, D. P. Information consistency of nonparametric gaussian process methods. *IEEE Transactions on Information Theory (IEEE TIT)*, 2008.

Shah, R. D. and Peters, J. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 2020.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., Kerminen, A., and Jordan, M. A linear non-gaussian acyclic model for causal discovery. *J. Mach. Learn. Res.*, 2006.

Shpitser, I., Evans, R. J., and Richardson, T. S. Acyclic linear sems obey the nested markov property. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2018.

Simpson, E. H. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 1951.

Solus, L., Wang, Y., and Uhler, C. Consistency guarantees for greedy permutation-based causal inference algorithms. *Biometrika*, 2021.

Spirtes, P., Meek, C., and Richardson, T. An algorithm for causal inference in the presence of latent variables and selection bias. *Computation, causation, and discovery*, 1999.

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

Squires, C. and Uhler, C. Causal structure learning: A combinatorial perspective. *Foundations of Computational Mathematics*, 2022.

Squires, C., Wang, Y., and Uhler, C. Permutation-based causal structure learning with unknown intervention targets. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2020.

Tjøstheim, D. Granger-causality in multiple time series. *Journal of Econometrics*, 1981.

Triantafillou, S. and Tsamardinos, I. Constraint-based causal discovery from multiple interventions over overlapping variable sets. *Journal of Machine Learning Research*, 2015.

Vinh, N. X., Epps, J., and Bailey, J. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *International Conference on Machine Learning (ICML)*, 2009.

Vinh, N. X., Epps, J., and Bailey, J. Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. *J. Mach. Learn. Res.*, 2010.

von Kügelgen, J., Gresele, L., and Schölkopf, B. Simpson's paradox in covid-19 case fatality rates: a mediation analysis of age-related causal effects. *IEEE Transactions on Artificial Intelligence (IEEE TAI)*, 2021.

von Kügelgen, J., Besserve, M., Wendong, L., Gresele, L., Kekić, A., Bareinboim, E., Blei, D. M., and Schölkopf, B. Nonparametric identifiability of causal representations from unknown interventions. In *Neural Information Processing Systems (NeurIPS)*, 2023.

Wahl, J. and Runge, J. Separation-Based Distance Measures for Causal Graphs. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2025.

Wang, Y., Solus, L., Yang, K. D., and Uhler, C. Permutation-based causal inference algorithms with interventions. In *Neural Information Processing Systems (NeurIPS)*, 2017.

Wang, Y., Squires, C., Belyaeva, A., and Uhler, C. Direct estimation of differences in causal graphs. In *Neural Information Processing Systems (NeurIPS)*, 2018.

Watanabe, S. Information theoretical analysis of multivariate correlation. *IBM Journal of Research and Development*, 1960.

Wu, A., Kuang, K., Zhu, M., Wang, Y., Zheng, Y., Han, K., Li, B., Chen, G., Wu, F., and Zhang, K. Causality for large language models. *arXiv preprint arXiv:2410.15319*, 2024.

Wu, C. F. J. On the Convergence Properties of the EM Algorithm. *The Annals of Statistics*, 1983.

Xu, S., Mameche, S., and Vreeken, J. Information-theoretic causal discovery in topological order. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2025.

Yang, K. D., Katoff, A., and Uhler, C. Characterizing and learning equivalence classes of causal dags under interventions. In *International Conference on Machine Learning (ICML)*, 2018.

Yu, Y., Chen, J., Gao, T., and Yu, M. Dag-gnn: Dag structure learning with graph neural networks. In *International Conference on Machine Learning (ICML)*, 2019.

Zanga, A., Ozkirimli, E., and Stella, F. A survey on causal discovery: Theory and practice. *International Journal of Approximate Reasoning*, 2022.

Zeng, Y., Cai, R., Sun, F., Huang, L., and Hao, Z. A survey on causal reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. Kernel-based conditional independence test and application in causal discovery. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2011.

Zhang, K., Gong, M., Ramsey, J., Batmanghelich, K., Spirtes, P., and Glymour, C. Causal discovery in the presence of measurement error: Identifiability conditions. *arXiv preprint arxiv:1706.03768*, 2017.

Zhang, X.-H., Tee, L. Y., Wang, X.-G., Huang, Q.-S., and Yang, S.-H. Off-target effects in crispr/cas9-mediated genome engineering. *Molecular Therapy - Nucleic Acids*, 2015.

Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. Dags with no tears: Continuous optimization for structure learning. *Neural Information Processing Systems (NeurIPS)*, 2018.