Don't Confound Yourself: Causality from Biased Data

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 David Kaltenpoth

Saarbrücken, 2024

Date of colloquium	25 November 2024
Dean of faculty	Roland Speicher
Chair	Isabel Valera
Reviewer	Prof. Dr. Jilles Vreeken
Reviewer	Prof. Dr. Gerhard Weikum
Reviewer	Prof. Dr. Khun Zhang
Academic Assistant	Lénaïg Cornanguer

Abstract

The core assumption of not only the sciences but of our broader (western) worldview is that nature adheres to certain laws which can be described in terms of causal relationships. This assumption has repeatedly proved successful in improving our understanding of how the world works. However, as we explore ever more complex domains, it has become clear that identifying these causal links is extraordinarily challenging. While collecting observational data is straightforward, it is unfortunately the case that "correlation is not causation". Instead, the gold standard in causal inference is to run randomized controlled trials (RCT). But these often face ethical, economic, or physical limitations. Even when such an RCT is feasible, the inclusion criteria are frequently too stringent for derived causal estimates to generalize to other settings.
In this thesis we therefore investigate under which assumptions we can distinguish causal relationships between observed variables from biases due to unobserved confounding or selection in purely observational data. Furthermore, we investigate to what extent we can discover the causal graph over both observed and unobserved variables. Our main tools will be the algorithmic model of causality and the independence of causal mechanisms. That is, if different causal mechanisms convey no information about each other, then by exploiting the violations of this assumption, we can discover the effects of unmeasured variables. We will show that such deviations exist at multiple levels of the causal description, both at the level of the parameters of a single causal model over the observed variables, as well as the level of mechanism changes between different causal models over the same set of variables.

Zusammenfassung

Die grundlegende Annahme nicht nur der Wissenschaften, sondern auch unserer breiteren (westlichen) Weltanschauung ist, dass die Natur bestimmten Gesetzen folgt, welche in Form von kausalen Beziehungen beschrieben werden können. Diese Annahme hat sich wiederholt als erfolgreich erwiesen, um unser Verständnis der Welt zu verbessern. Doch je komplexer die Bereiche die wir erforschen werden, desto deutlicher wird, dass es außerordentlich herausfordernd ist, diese kausalen Verbindungen zu identifizieren. Das Sammeln von Beobachtungsdaten ist zwar unkompliziert, doch leider gilt dass „Korrelation ist nicht Kausation“. Stattdessen ist der Goldstandard für kausale Effekte das Durchführen von randomisierten kontrollierten Studien (RCT). Diese stehen jedoch oft vor ethischen, ökonomischen oder physischen Einschränkungen. Selbst wenn ein RCT durchgeführt werden kann, sind die Einschlusskriterien oft zu streng um die abgeleiteten kausalen Effekte auf andere Szenarion zu verallgemeinern. In dieser Arbeit untersuchen wir, unter welchen Annahmen wir kausale Beziehungen zwischen beobachteten Variablen von Verzerrungen durch unbeobachtete Störfaktoren oder Selektionsprozessen unterscheiden können, ohne Verwendung von experimentelle Daten. Weiterhin untersuchen wir, inwiefern wir kausale Graphen über sowohl beobachtete als auch unbeobachtete Variablen entdecken können. Unsere Hauptwerkzeuge werden das algorithmische Modell der Kausalität und die Unabhängigkeit kausaler Mechanismen sein. Das heißt, wenn kausale Mechanismen keine Informationen übereinander übermitteln, dann können wir durch das Ausnutzen der Verletzungen dieser Annahme die Auswirkungen von ungemessenen Variablen entdecken. Wir werden zeigen, dass solche Abweichungen auf mehreren Ebenen der kausalen Beschreibung existieren, sowohl auf der Ebene der Parameter eines einzelnen Modells über die beobachteten Variablen, als auch auf der Ebene von Mechanismusänderungen zwischen verschiedenen kausalen Modellen über den selben Variablen.

This is for those who gave everything and asked for nothing.

Acknowledgments

First, I am grateful for my advisor's patience and optimism even at times where nothing seemed to be moving forward. Many times, it would have been all too easy to quit, and perhaps if he had been any less patient I would have. Second, I am grateful for the complete freedom my parents gave me at almost all times of my life, even when I went severely off the tracks. Without all the experiences I gained on the misadventures made possible by their permissiveness, I likely would not be where I am now. And third, I would like to thank all the kind souls I have met over the years. The world can seem like a dark place, but it would be much darker still if they weren't around.

Table of Contents

Chapter 1

From Plato to Probabilistic Models of Causality

> *Everything that becomes or changes must do so owing to some cause; for nothing can come to be without a cause.*
>
> Plato, Timaeus 28a

Plato was, of course, not the first person ever to consider the nature of reality regarding questions of causal relations. In fact, asking the question "why?" is a fundamental part of human nature. As we observe the world around us, we do our utmost to weave our observations into a coherent whole. While it has been popular in recent years to refer to brains as "prediction machines," this is a mischaracterization of what brains do; more adequately, we should refer to them as *causal prediction machines*. Humans do not only exploit statistical regularities present in their observed data, but create causal models as to how these regularities arise (Cisek and Kalaska, 2010; Ramstead et al., 2023) In fact, children begin developing notions of causality already at the earliest stages of development, well before they can even formulate or conceptualize such notions, or even understand what exactly they are doing (Buchsbaum et al., 2012; Weisberg and Gopnik, 2013).

Questions of "why are things the way they are?" or "how did things come to be this way?" are inherently causal questions: what are the underlying *causal mechanisms* generating the observed structure of the universe, and how can we come to know them? The most basic step towards developing such explanations is to *observe* the world around us, noticing correlations between

distinct factors of reality and forming various hypotheses about the interplay of and relationship between these factors.

Further, as individual agents within the world taking actions, we inherently generate *interventions*: by interacting with our surroundings, we consciously or unconsciously *intervene on* some variables of the system, so that we no longer passively observe the world's processes as they would run their natural course, but instead, as actors, produce and obtain interventional data resulting from this interference with the causal system we are part of. This kind of interventional data is generally considered to be fundamental to our ability to refine our causal models of the world.

At the same time, to the extent that our actions are intentional, this already assumes that we can, at least to some extent, predict which actions will benefit our goals and which will not. Our interactions with the world are founded on the very assumption that we can use the causal models we have crafted from past observations to predict the results of our actions.

Throughout their development, humans form an intuitive grasp of many different domains: certain aspects of physics, psychology, economics, and biology, as well as countless others. These intuitive models are called "folk" theories—often derogatorily—and their accuracy can vary dramatically, not only across different areas but also within them (Gelman and Legare, 2011). While our naïve model of physics is highly effective in unconsciously predicting the trajectory of a ball thrown directly toward us, it fails spectacularly when it ricochets off a wall. Similarly, our folk model of psychology is good at some aspects of psychology but fails horribly at others (Kahneman and Tversky, 2013).[1]

The goal of science is, therefore, to make up for the limitations and failures of human intuition. The scientists' task is twofold in its endeavor: to develop better answers to the questions of "why?" and "how?" and to develop appropriate methodologies to answer these questions in the first place. To derive such answers and develop such methods, scientists develop theoretical, both qualitative and quantitative, models to make predictions, gather observational data, and orchestrate experiments to verify the models' predictions. In order to analyze the data they gather and the experiments they undertake and to ensure the correctness of the derived causal conclusions, they must also develop the statistical and causal frameworks within which it becomes possible to answer these questions and verify the answers.

Clearly, the success of this methodology hinges critically both on our ability to develop such causal frameworks as well as on our ability to obtain data that can be used to validate our models within them. While data collection

[1] Although much of the last 20 or more years of research in heuristics and biases, and especially in behavioral economics, has turned out not to replicate and should be taken with a grain of salt, many of the underlying psychological principles have stood up to repeated attempts at replication.

is often conducted experimentally, such experiments are not always feasible. In fields such as economics or epidemiology, the relevant experiments are often unethical, financially prohibitive, or outright impossible with our current technology (Grossman and Mackenzie, 2005). However, even when such experiments are feasible, all is not necessarily well.

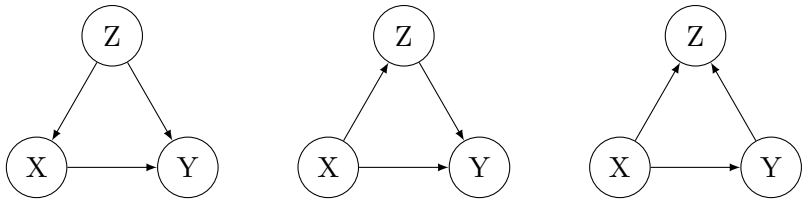
RCTs & Reality: Not All That Glitters is a Gold Standard

Observation is the cornerstone of developing causal models not only for individuals but also for science as a whole: a good scientific theory has to be consistent with all our empirical observations, so that our observed correlations impose *hard constraints* on the set of viable causal models that could be governing the phenomena we are interested in. Any model that cannot explain these observed correlations is incomplete at best and straight-up wrong at worst. For this reason, simply by passively observing the world around us and noting whatever patterns and correlations we find, we can already obtain glimpses into the underlying causal structure of reality.

Yet, while these patterns and associations impose constraints, they rarely suffice to discern real causation from *spurious correlation*. External factors not accounted for in the data gathered may surreptitiously influence the observed outcomes, leaving us with plausible but incorrect models and misleading narratives spun from these models. This interference by *latent confounders*—unmeasured variables which affect multiple observed variables—plagues all sciences (Hemkens et al., 2016; Secrest et al., 2020) and leads to incorrect estimates of causal effects when care is not taken. That is, to be confident in our causal estimates, we need to develop methods that can determine which correlations are due to causal influences and which are spurious.

Traditional attempts to control such biases are commonly used in the econometrics, epidemiology, and psychology literature; however, they *already presume knowledge of the causal graph* (Wysocki et al., 2022). When the causal graph is not known, controlling for other covariates by "regressing them out", or in any of a number of other ways, can be either helpful or harmful to the estimation of the true causal parameters, and it is only by knowing this graph that we can distinguish between the two cases. We show the three basic cases in Figure 1.1. If we do not know the causal relationship between X, Y, Z, we *cannot* tell whether the total causal estimate for $X \to Y$ when controlling for Z is better or worse than the estimate when we do not control for Z. Only when Z is a common cause for X and Y will the estimate resulting from controlling for Z be improved. Otherwise, we introduce an unknown amount of bias.

To resolve these ambiguities, Randomized Controlled Trials (RCTs) are nowadays considered the gold standard for obtaining data that is not subject to such confounding (Deaton and Cartwright, 2018). By randomly assigning partici-



(a) Common Cause **(b)** Mediator **(c)** Collider

Figure 1.1: *Different causal relationships lead to different results when controlling for* Z. (a) Controlling for the common cause Z will produce the correct total causal effect for $X \to Y$. (b, c) Controlling for either mediator or collider Z can introduce *unknown amounts of bias* into the estimate of the total causal effect of $X \to Y$, which was, in fact, correctly estimated when we did *not* control for them.

pants to different groups, RCTs endeavor to ensure that all extraneous variables are equidistributed among groups. This equidistribution ensures that unmeasured factors have equal effects in both experimental and control groups on average, thus guaranteeing *internal validity* of the findings. In this controlled setting, the cause-effect relationships between variables become clear, allowing us to obtain insight into the true effect of one variable on another.

The most significant problem with this approach is, of course, paucity of such experimental data. Even when RCTs are possible and ethical, the higher cost of running tightly controlled experiments compared to obtaining merely observational data leads to theoretical identifiability of causal effects not always translating into practically distinguishable causal estimates. Since sample size is relatively small, there can be high variance in the distribution of all other factors, leading to these other factors *not* being equal in the two experimental and control groups (Deaton and Cartwright, 2018). While these concerns can be ameliorated by stratification or other randomization schemes designed to create equal distributions in the treated and control populations (Kang et al., 2008), these can only adjust explicitly for differences in the *observed* covariates, but not the unobserved ones (Arah, 2017). To remove biases in the unobserved covariates, we would again need large sample sizes, the very problem that these advanced randomization schemes were supposed to solve.

However, even barring issues of ethics, cost, and the associated low sample size, RCTs are not without their problems. In their pursuit of internal validity, study designs necessarily deviate from the messiness of real-world scenarios. Effects are often measured in a highly simplified setting, and the participants of a study are often a convenience sample chosen for their ease of availability or whose inclusion is otherwise subject to highly stringent selection criteria, and thus not truly representative of the larger population (Naci and Ioannidis, 2013). Many other factors, such as non-response bias, survivorship bias, admission rate bias,

and healthy user bias, among countless others, can exacerbate these issues even further. As such, even given perfect knowledge of the inclusion criteria, the causal estimates may not even translate between subsets of the population matched for their characteristics (Averitt et al., 2020). Even where RCTs solve the issue of internal validity, the question of *external validity* therefore remains: can results derived from a controlled environment (*ecological validity*) on an unrepresentative sample of the population (*population validity*), be generalized to the sprawling complexities of the real world? In other words, is it possible to *transport* the causal estimate derived in one (sub-)population to another?

Of course, the Taoists of old already knew that "the causal effect that can be measured is not the true causal effect."[2] In recent years, the replication crisis gives us further reason to believe that the answer is an emphatic *no* (Open Science Collaboration, 2015; Camerer et al., 2018; Gordon et al., 2020). While many other aspects, such as flawed statistical methodology, file drawer effects, p-hacking, and other questionable research methodology, up to outright scientific fraud, are undoubtedly large contributors to the non-replicability of many of the results, we must ask to what extent the fundamental underlying issue might be the lack of external validity in our studies (Yarkoni, 2022). While this disregard for external validity is understandable, given the academic pressure to produce surprising but minimally viable results for publication, combined with the difficulty of evaluating the external validity of causal estimates in general, it leads to *a priori* flawed study designs with little to no chance of replication in the first place. Thus, while RCTs endeavor to refine our understanding of the causal landscape by filtering out noise, they are often at risk of missing the very facets of reality which we are interested in, providing us instead with false or misleading narratives that can lead to long-lasting real-world harm.[3]

This problem of external validity in RCTs also explains the failure of the commonly proposed approach of combining large quantities of biased observational data with small quantities of unbiased experimental data (Statnikov et al., 2015; Cheng and Cai, 2021; Kladny et al., 2023; Colnet et al., 2024). The issue is that these data combination approaches are a fundamentally flawed attempt at overcoming the real issues posed by the two types of data. The underlying assumption, that the experimental data is unbiased (i.e., internally valid), virtually never obtains in the first place. Other attempts, such as using meta-analytic methods, have been made to ensure that the obtained causal estimates may be transported between study populations (Dahabreh et al., 2020;

[2]Originally, "The Tao that can be spoken is not the true Tao."

[3]It also does turn out that, at least in the field of medicine, the results produced by RCTs are almost invariably indistinguishable from those obtained from purely observational studies (Anglemyer et al., 2014; Bun et al., 2020; Bröckelmann et al., 2022). That is, either RCTs never gave us the internally valid results we wanted in the first place, or we never needed RCTs to obtain them.

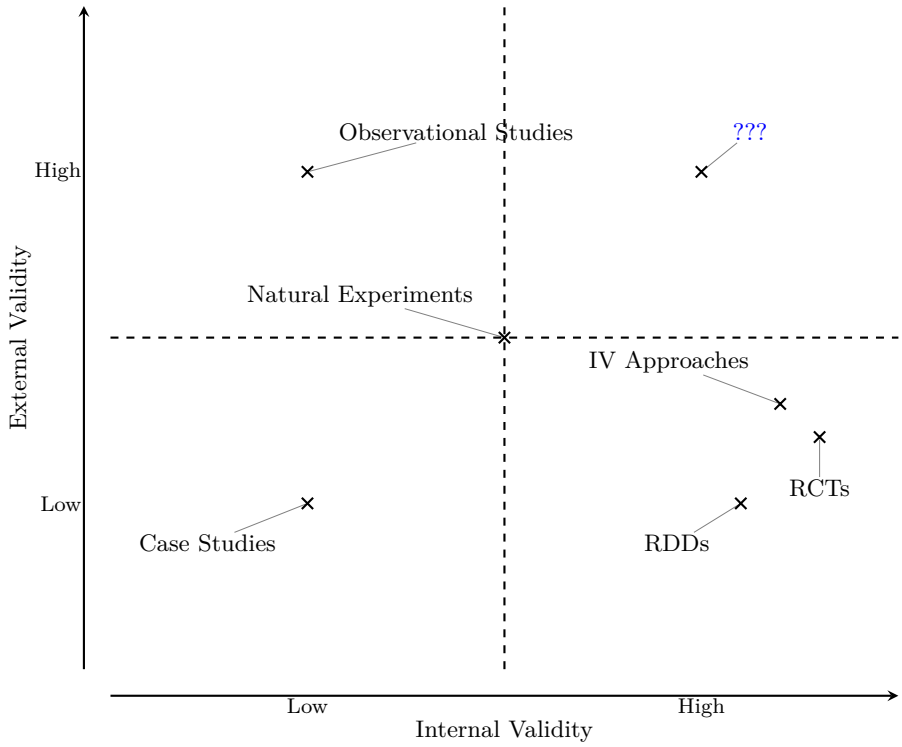


Figure 1.2: *Different types of studies have different benefits and drawbacks.* In this dissertation we are interested in solutions satisfying both high internal and external validity.

Markozannes et al., 2021). Unfortunately, these methods can currently deal only with the case of data obtained from multiple RCTs rather than combining multiple types of data collected from different (non-)experimental contexts.

Other approaches, such as *instrumental variable*-based methods (also known as *Mendelian randomization* when the instrument is genetic variation; Angrist et al., 1996; Sanderson et al., 2022), and *regression discontinuity designs* (RDDs; Imbens and Lemieux, 2008) both assume precise knowledge about subsets of the causal graph, but are also open to further questions. Is the instrument indeed valid? Do estimates from the regression discontinuity apply to other regions of the covariate range? Meanwhile, *natural experiments* (Rosenzweig and Wolpin, 2000) may at times be good alternatives to designed experiments as they do not artificially restrict the included subjects, but it is often impossible to tell to what extent causal estimates derived from such data are representative of the causal mechanism at large as compared to the idiosyncratic circumstances particular to that time and place. In contrast to these

approaches designed to leverage (quasi-)experimental data, our two goals in this thesis are different (see Figure 1.2). We are concerned with the questions of whether and how we can learn *unbiased*, internally and externally valid, causal structures *from purely observational data*. First, under which conditions can we *provably* ensure the internal validity of our results and learn causal structures and effects directly from observational data? Second, how can we provably determine if the (potentially experimental) data we have obtained is affected by selection bias, thereby violating external validity? Moreover, under which conditions can we recover the original distribution from it?

Causalchemy: From Leaden Observation to Golden Insights

> *"One box contains a key," said the king, "to unlock your chains; and if you find the key you are free. But the other box contains a dagger for your heart, if you fail."*
> *And the first box was inscribed: "Either both inscriptions are true, or both inscriptions are false."*
> *And the second box was inscribed: "This box contains the key."*
> *The jester reasoned thusly: "Suppose the first inscription is true. Then the second inscription must also be true. Now suppose the first inscription is false. Then again the second inscription must be true. So the second box must contain the key, if the first inscription is true, and also if the first inscription is false. Therefore, the second box must logically contain the key."*
> *The jester opened the second box, and found a dagger.*
> *"How?!" cried the jester in horror, as he was dragged away. "It's logically impossible!"*
> *"It is entirely possible," replied the king. "You merely wrote those inscriptions on two boxes, and then I put the dagger in the second one."*
>
> Adapted from The Parable of the Dagger

As we have noted, while observational data constrains the facts that need to be correctly predicted by a proposed scientific explanation, these constraints will generally only partially determine the true causal model. One must, therefore, invoke additional assumptions to bridge the gap between observation and causation (Spirtes et al., 2000; Pearl, 2009). These assumptions serve to distinguish between different *observationally equivalent* mechanisms generating the data based on external information that is not included in the data itself. These assumptions vary from setting to setting and often require expert domain knowledge, but common assumptions include sparsity, (non-)linearity,

(in-)variance, (non-)Gaussianity, (un-)confoundedness, and lack of selection bias (or pure selection). That is, it is often assumed that the correct causal model is the sparsest among all equivalent models, that the causal mechanisms are linear (or strictly nonlinear), that the causal mechanisms are stable over time or across environments (or not), that the exogenous sources of variation are Gaussian (or strictly non-Gaussian), that there is no latent confounding as well as no selection bias (or that all correlations are due to latent variables).

While none of these assumptions are likely to hold exactly, and we will revisit all of them in the following chapters, for now, we will focus on the last and most egregious: that all correlations are due to endogenous causal relations between the measured variables rather than due to latent variables inducing confounding or selection bias among the observed variables.

First, while the other assumptions are often satisfied at least approximately, it is *virtually never* the case that all relevant variables have been measured, and in fact, in most complex domains, we neither know what the relevant variables are nor would we be able to measure all of them even if we did. To assume that there are no latent variables is to assume away the biggest challenge in most fields of science. In fact, the contribution of the latent factors to the observed correlations can easily be larger than those of endogenous causal mechanisms (Boyle et al., 2021; Barth et al., 2022). Or, if by some chance this assumption does hold, the reason is more likely than not that the data was obtained in a lab, with all variables subjected to strict control and the causal relations already known or suspected from the start.

Second, the assumption is commonly unverifiable, both ex-ante and ex-post. When such methods return causal gobbledygook with more spurious than correct edges, this must be taken at face value because no tools exist to determine which edges are real and which are due to the effect of latent confounders. In contrast, assumptions of (non-)linearity, (in-)variance, and (non-)Gaussianity of exogenous noise variables are all testable after a method has been applied to data, by checking whether the error residuals have the expected properties.

Thus, while the assumption that all relevant variables have been observed, known as *causal sufficiency*, is persistent in the literature on causal discovery and inference, it is often both untestable and untenable in practice.

As such, the overarching goal of this thesis is to determine whether we can do away with this assumption. More precisely, we develop methods for testing the validity of this assumption and determining the extent of its violation (Chapters 2, 5), as well as exploiting the structure in our data to recover the latent variables and performing causal discovery over both observed *and* unobserved variables when the assumption does not hold (Chapters 3, 4, 6).

More precisely, our first problem will be the simple question: can we detect at all whether our data *might* be confounded based on observational data alone?

Problem Statement 1 (Confounded or Causal)**.** Given covariates X and a target variable Y, does X cause Y, or are the correlations best explained due to latent confounding caused by not controlling for an unobserved variable Z?

Asked differently, which patterns in the data can we exploit to distinguish between these two different potential models? It turns out that one of the most important factors is the number of observed covariates X. When only few jointly confounded covariates X are observed, it is almost impossible to distinguish between latent confounding and causal connections between variables. In contrast, when a larger number of variables is observed, patterns induced by latent confounding become more predominant, and we can use these patterns to distinguish between the cases. As such, while the common intuition of "just measure more variables (and control for them)" shared by many social scientists is misguided in the sense that controlling for everything *does not* solve the problem, and can even *exacerbate* it to an unknown degree (Wysocki et al., 2022), measuring more covariates is, in fact, useful when adequate procedures are employed to make use of them to detect latent confounding.
What if the covariates *are* jointly confounded? Commonly used causal discovery methods may not be applicable, but is there anything we *can* do to learn the underlying causal structure? Ideally, can the same patterns used to distinguish between causality and confounding also be used to learn which correlations are due to causal mechanisms and which are due to latent confounding?

Problem Statement 2 (Causal Discovery with Hidden Confounders)**.** Given only data over the observed variables X but not the data for the unobserved variables Z, can we discover a joint causal network over X and the unobserved confounding factors Z affecting them?

This problem is, of course, far too general for us to solve directly, so we start by tackling the following simpler version, assuming linear causal relationships.

Problem Statement 2a) (Causal Discovery with Hidden Confounders—Linear Case)**.** Given data only for the observed variables X, and assuming that all causal relationships between both the unobserved Z and the observed X are purely linear, can we discover a joint causal network over both the observed variables X as well as the unobserved Z?

It turns out that under some additional assumptions on the causal structure, which mainly amount to sparsity of the underlying causal graph, the answer to this question is *yes*. In essence, by exploiting the linearity of the causal relations, we can decompose the correlations between the observed variables into those attributable to latent factors—inducing specific low-dimensional patterns in these correlations—and those due to direct causal effects between observed variables—deviations from the patterns expected from pure confounding.

Next, since strict linearity of causal mechanisms is a strong requirement, we consider whether we can use the framework developed in answering the last question to also deal with some nonlinearities in our causal mechanisms.

Problem Statement 2b) (Causal Discovery with Hidden Confounders—Nonlinear Case)**.** Given data only for the observed variables X but not the unobserved variables Z, do nonlinear causal mechanisms exist for which we can discover a joint causal network over X and Z?

We begin by showing that recent identifiability results for nonlinear ICA do not directly translate into any results for this question. We then proceed to show that the post-nonlinear causal model precisely fits the bill (Zhang and Hyvärinen, 2009). It permits us to describe the correlations between observed variables solely in terms of linear relations, enabling us to use the tools and theory we developed to tackle the previous question.
Having obtained positive results for the case of confounding, we turn to the related but distinct problem of selection bias. While one might intuitively expect that solutions for one kind of problem translate readily into solutions for the other, this is unfortunately not the case. We therefore ask whether there are any other patterns we can exploit in the case of selection bias. That is, under which conditions can we tell apart data subject to (artificial) selection bias from data that arose "naturally"? Can we determine how strong these selection effects are? Can we recover the true underlying distribution?

Problem Statement 3 (Dealing with Selection Bias)**.** Given data for the observed variables X, are they causally connected, or are the correlations best explained by selection effects due to conditioning on an unobserved collider Z?

To study this problem, we consider the case of linear selection effects, i.e., threshold effects for including data points based on a linear combination of the observed values. We show that this kind of selection bias can be discovered for parametric (specifically, exponential) families of distributions and certain nonparametric families of distributions for which we have prior information about their invariance structure (e.g., rotational invariance).
Last, we study if data from multiple contexts is useful for detecting latent confounders. A central observation in the last years on causal discovery has been that using data from multiple sources, such as data from different hospital sites or observational and experimental data, allows us to obtain stronger identifiability on the underlying causal graph and its mechanisms (Mooij et al., 2020; Huang et al., 2020). However, the vast majority of these approaches still assume causal sufficiency and are, therefore, not designed to deal with latent confounding or selection. Therefore, we take some first steps towards determining to what extent the presence of latent confounding can be detected when we have access to such multiple data sources.

Problem Statement 4 (Confounding across Contexts)**.** Given data only for the observed variables X across multiple contexts $c \in C$, but not the unobserved variables Z, how readily can we determine which variables are jointly confounded by the same latent factor $Z_i \in Z$?

From the answers to our previous questions we already know that we can figure out a lot about the structure under parametric assumptions. Given data from multiple environments, however, it turns out that the answer is yes under quite generic assumptions, requiring neither parametricity nor strong structural assumptions on the underlying causal model. Instead, we only require some weak assumptions about the mechanism changes across environments.

We summarize the contributions of this thesis and the relevant research articles on which it is based in the following table. All chapters are based on these research articles, albeit many structural changes were made. These include changes of notation, restructuring of results for narrative coherence, and inclusion of additional results and discussions relating them to each other and to other related work that did not fit within the pages of the original publications. For the publications forming the basis of Chapters 2–5, the author took the lead in ideation, developing the theory, experiments, and writing. For the article on which Chapter 6 is based, the author provided the initial idea, contributed to the theory and writing, and took on a more supervisory role.

Reference	**Chapter**
D. Kaltenpoth and J. Vreeken. We Are Not Your Real Parents: Telling Causal from Confounded using MDL. In *SDM*, pages 199–207. SIAM, 2019	Chapter 2
D. Kaltenpoth and J. Vreeken. Causal Discovery with Hidden Confounders Using the Algorithmic Markov Condition. In *UAI*, pages 1016–1026. PMLR, 2023a	Chapter 3
D. Kaltenpoth and J. Vreeken. Nonlinear Causal Discovery with Latent Confounders. In *ICML*, pages 15639–15654. PMLR, 2023c	Chapter 4
D. Kaltenpoth and J. Vreeken. Identifying Selection Bias from Observational Data. *AAAI*, 37(7):8177–8185, 2023b	Chapter 5
S. Mameche, J. Vreeken, and D. Kaltenpoth. Identifying Confounding from Causal Mechanism Shifts. In *AISTATS*. PMLR, 2024	Chapter 6

Table 1.1: Publications on which this thesis is built and their corresponding chapters.

Notation

Symbol	Description
$X = (X_1, \dots, X_m), Y$	Observed variables
$Z = (Z_1, \dots, Z_l)$	Unobserved variables
V	General set of variables, usually $X \cup Y \cup Z$
$\mathcal{I}, \mathcal{J} \subseteq [m] = \{1, \dots, m\}$	Index sets
$V_{\mathcal{I}} = \{V_i : i \in \mathcal{I}\}$	A subset of variables corresponding to index set $\mathcal{I}$
$P(V)$	Joint probability distribution over V
x, y, z, v	Samples for the corresponding set of variables
$G = (V, E)$	Graph with vertices V and edges E
$X_i \to X_j$	Edge $(X_i, Y_j) \in E$
$X_{\mathcal{I}} \to Y$	Set of edges $\{X_i \to Y : i \in \mathcal{I}\} \in E$
$\mathrm{Pa}_G(Y), \mathrm{Pa}_i$	The set of parents of Y in G, and $\mathrm{Pa}(X_i)$
G^*, Pa^*	True graph, and parent sets in this graph
A, B, C	Matrices
U	Orthogonal matrix
ε	Exogenous noise variables
$X \perp\!\!\!\perp Y \mid Z$	Conditional independence
f, g, τ, ν	(Nonlinear) Functions
K	Kolmogorov Complexity
$\overset{+}{=}, \overset{+}{\leq}$	(In-)equality up to additive constants
$I(x : y)$	Algorithmic Mutual Information
$I(X; Y)$	Statistical Mutual Information
KL	Kullback-Leibler Divergence
$L(x, M), L(x; \mathcal{M})$	MDL Scores
$\mathcal{C}$	Confidence score of a method
$\Pi = \{\pi_1, \dots, \pi_r\}$	Partition
$s \in \mathcal{S}$	Settings
P^s	Probability distribution in setting s

Chapter 2

Telling Causal from Confounded

"If you can't solve a problem, then there is an easier problem you can solve: find it."

George Pólya, Mathematical Discovery on Understanding, Learning, and Teaching Problem Solving

Causal inference is one of the most challenging and important problems in statistics (Pearl, 2009). As we have seen, the commonly proposed "gold standard" of using RCTs or other controlled data sources comes with its own set of problems. It by no means guarantees that causal estimates will be accurate, much less so that causal graphs will be (Deaton and Cartwright, 2018).
Therefore, we focus on a different way of tackling the problem: finding conditions under which the causal effects (or networks) are identifiable from purely observational data (Peters et al., 2017; Glymour et al., 2019). That is, we want to find assumptions under which the observed data allows us to uniquely determine the causal factorization of the joint distribution of all variables.
One of the most common assumptions in causal inference is that of *causal sufficiency*. That is, to make sensible statements on the causal relationship between two statistically dependent random variables X and Y, it is assumed that no hidden confounder Z exists that causes both X and Y. When this assumption holds, all correlations between two variables X and Y that cannot be *explained away* by conditioning on other observed variables must be due to direct causal effects between X and Y. In practice, this assumption is virtually

always violated—because we do not know all the relevant factors, or because we cannot measure everything—and existing methods relying on this assumption will return supposedly "causal" graphs containing many spurious edges (Tu et al., 2019). Worse yet, the assumption is not only often violated, but most methods do not, even *can not*, check whether the assumption holds.

In this chapter, we begin by asking the following simple question: given a set of variables $X = (X_1, \ldots, X_m)$ and a target Y, does X cause Y or does there exist some set of unmeasured confounders $Z = (Z_1, \ldots, Z_l)$ co-causing both of them? To answer this question, we build upon the algorithmic model of causality (AMC) introduced by Janzing and Schölkopf (2010). Within this framework of causality, the simplest—measured in terms of Kolmogorov complexity—factorization of the joint distribution coincides with the true causal model. In the simplest case of only two variables without latent confounding, if X causes Y, then the complexity of the factorization in the causal direction, $X \to Y$, should be lower than in the anti-causal direction, $Y \to X$,

$$K(P(X)) + K(P(Y \mid X)) < K(P(Y) + K(P(X \mid Y))\,.$$

We propose, similarly, that if a third variable Z confounds X and Y, then the complexity of the factorization according to the model, including this confounder *should* satisfy the corresponding inequality

$$K(P(Z)) + K(P(X \mid Z)) + K(P(Y \mid Z)) < K(P(X)) + K(P(Y \mid X))\,.$$

Note that we do not include $K(P(Z))$ on the right-hand side here. Our claim is, therefore, that including the (correct) latent confounder is strictly better than *any* proposed causal model over solely the observed variables.

Of course, since we have not measured the confounding factor Z, we cannot evaluate the Kolmogorov complexity terms involving it. We will, therefore, employ latent factor models to estimate the joint distribution $P(X, Y, Z)$ (Loehlin, 1998). Within the AMC, the true confounder helps us compress the data optimally so that no other model, including latent confounders, can perform any better. Hence, if we can nevertheless find such a latent factor model outperforming purely causal models, we know that there must exist *some* latent confounder, even if the specific model we fit to the data is itself misspecified. To address the fact that K itself is not computable (Li and Vitányi, 2009), we use the Minimum Description Length (MDL) principle (Grünwald, 2007), which provides us with a statistically well-founded variational upper bound for the Kolmogorov complexity K by restricting the set of permitted programs.

To introduce our approach and set the tone for the remainder of this thesis, we begin by introducing the required background for causal discovery in Section 2.1, including a precise description of our problem in Section 2.1.2 and some intuitively appealing attempts at solutions, which are unfortunately not

all that, in Section 2.1.3. We then introduce the algorithmic model of causality in Section 2.2, which in particular lets us formalize the notion of independence of causal mechanisms in Section 2.2.3. The algorithmic model of causality will underpin the remainder of this thesis as a *conceptual tool to think with.*
In Section 2.3, we extend the algorithmic model of causation to include latent confounding directly instead of considering them only as violations of the standard formulation. Since Kolmogorov complexity itself is not computable, in Section 2.4 we restrict the problem from *all* programs to subsets of programs for which the code lengths can be computed. This is similar to the commonly used approach of variational inference, where we optimize over some tractable subset of the distributions we would, in fact, like to optimize over. We then show that it performs well empirically on both synthetic and real-world data, and also in comparison with other approaches, in Section 2.6. We wrap up with Section 2.7, where we discuss some limitations of our approach and the literature on the topic in general, as well as outline some additional problems that we will further discuss in subsequent chapters. We include proofs for all theoretical statements in Appendix A.2, and include here only proof sketches.

2.1 Causal Inference and Confounding

We consider here the setting where we are given n samples from the joint distribution $P(X, Y)$ over two statistically dependent continuous-valued random variables X and Y. We require Y to be a scalar, i.e., univariate, but allow $X = (X_1, \ldots, X_m)$ to be of arbitrary dimensionality so that it may be univariate or multivariate. We also allow a set of unmeasured variables, $Z = (Z_1, \ldots, Z_l)$ to influence the observed variables X, Y so that $P(X, Y)$ is the marginal distribution of the joint distribution $P(X, Y, Z)$ over both observed and unobserved variables. We will write V to refer to a generic set of variables, which can generally be taken to be $V = X \cup Y \cup Z$. We begin by introducing the basic framework for causal discovery and inference that will be used throughout the rest of this manuscript.

2.1.1 Causal Basics

It is impossible to infer causal effects from observational data without making any assumptions (Spirtes et al., 2000; Pearl, 2009). That is, to reason about the effects of an intervention (or counterfactual) on the distribution $P(V)$ generating our data, we require assumptions on the (properties of) a causal model in the first place. Without any such assumptions, many processes with different causal relationships are consistent with the observed distribution $P(V)$.
To start with, we introduce Bayesian Networks (BNs), which provide a graphical representation of the distribution $P(V)$ (Koller and Friedman, 2009).

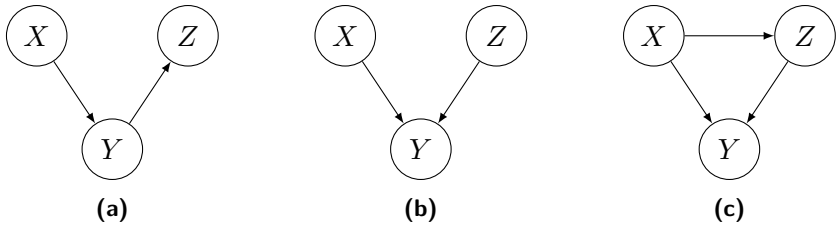


Figure 2.1: *Three proposed Bayesian networks for the distribution* $P(X, Y, Z) = P(X)P(Z)P(Y \mid X, Z)$. All three graphs entail different conditional independences. As such, only graph (b) captures the independences $X \perp\!\!\!\perp Z$ and $X \not\perp\!\!\!\perp Z \mid Y$ correctly.

Definition 2.1 (Bayesian Network)**.** A Bayesian Network for a distribution $P(V)$ is a directed acyclic graph (DAG) $G = (V, E)$ with nodes V and edges $E \subseteq V \times V$, denoted by $X \to Y$ when $(X, Y) \in E$, such that

$$P(V) = \prod_{Y \in V} P(Y \mid \mathrm{Pa}_G(Y)),$$

where $\mathrm{Pa}_G(Y) = \{X \in V : X \to Y \in E\}$ is the set of parents of Y in G. For ease of notation, we also write $S \to Y$ for subsets $S \subseteq \mathrm{Pa}_G(Y)$ when all variables $X_i \in S$ have an edge $X_i \to Y$.

Of course, for any given distribution $P(V)$, there are many different ways of factorizing it and, thus, many different corresponding BNs G. To constrain the set of admissible networks we therefore require further assumptions.
Consider the example where X and Y are independent, $X \perp\!\!\!\perp Y$. While we could consider the graph $G = (\{X, Y\}, \{X \to Y\})$, the corresponding factorization $P(X, Y) = P(X)P(Y \mid X)$ would not capture the independence of X and Y. In a sense, this graph is *not minimal* in that it allows us to model correlations that are not, in fact, present in the observed distribution that is to be modeled. We, therefore, want to find BNs G that capture precisely the (in-)dependences of the variables that hold in the observed distribution $P(V)$. In particular, if we want to interpret G causally, we need to be able to look at G and know which variables would change if we were to intervene on any given target variable $V_i \in V$. In the previous example of independent X and Y, changes in either variable will not influence the other. This leads us to our first assumption, the Markov condition (MC; Pearl, 2009).

Definition 2.2 (Markov Condition)**.** Each variable $Y \in V$ is independent, in $P(V)$, of all its non-descendants, given its parents $\mathrm{Pa}_G(Y)$ in G.

What does this mean? Ignoring the variable Y's effects, all the relevant information about it is contained in its parents $\mathrm{Pa}_G(Y)$. That is, in a (causally)

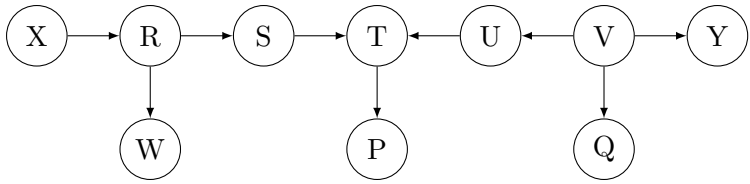


Figure 2.2: *Illustration of d-separation in a causal graph based on Pearl (2009).* X and Y are d-separated by the empty set $\emptyset$ because of the collider T on the path between them, but would become d-connected if we condition on either the collider T or its child P. U and Y are not d-separated, but become d-separated by conditioning on V.

Markovian system, the current state is relevant to the future, but how we arrived at the current state is irrelevant (Scheines, 1997). This is in contrast with the state-dependence of a system, also known as hysteresis (Ewing, 1882).
In particular, if G satisfies the MC for $P(V)$, a graphical criterion exists for deriving statistical independence in $P(V)$ from G. That is, under the MC, properties of G can be translated into properties of $P(V)$.
To derive useful properties from a BN, we require two more properties for our network, G and distribution P.
First, an (undirected) path p in G of length r is be a sequence of nodes $V_{i_1}, \ldots, V_{i_r} \in V$ such that there exists an edge between $V_{i_j}, V_{i_{j+1}}$ for all $j = 1, \ldots, r-1$. We call a node $C \in V$ a *collider* on the path p in G if $C = V_{i_j}$ for some j and such that $V_{i_{j-1}} \to C \leftarrow V_{i_{j+1}}$. An *unblocked path* between R, S is a path between two nodes $X \in R$ and $Y \in S$ containing no colliders C.
We call two sets $R, S \subseteq V$ d-connected by a set T if there exists a path p between R and S such that a) no non-collider on p lies in T and b) T contains C or a descendant of C for every collider in p. The sets R and S are called *d-separated* by T if they are not d-connected by T.

Example 2.1. Let us consider the graph in Figure 2.2 based on Pearl (2009).

a) The variables X, Y are d-separated by the empty set $\emptyset$ because the only path between them goes through the collider T. However, this path would become unblocked if we condition on either T (a collider) or P (a descendant of a collider).
b) The variables U, Y are not d-separated by $\emptyset$ because V is a common parent of both, but they are d-separated by V.

According to the MC, any two subsets R and S that are d-separated by T are, in fact, *independent* in P, i.e., $R \perp\!\!\!\perp S \mid T$. Therefore, the MC lets us draw inferences about conditional independences in $P(V)$ based on the graphical structure of the BN G. In order to learn such a graph G from data, we also need to be able to do the reverse; that is, draw inferences about the graphical

structure of G from conditional independence properties in $P(V)$. This is called the *Faithfulness* assumption.

Definition 2.3 (Faithfulness)**.** A Bayesian Network G is faithful to a probability distribution P if any three sets satisfying the independence $R \perp\!\!\!\perp S \mid T$ in P satisfy the same d-separation in G.

Note that by construction, the faithfulness assumption reduces to tests of pairwise conditional independence. By considering all paths between single nodes $X \in R$ and $Y \in S$, we are essentially assuming that conditional independence between R, S is violated, $R \not\perp\!\!\!\perp S \mid T$, because there exists at least one pair $X \not\perp\!\!\!\perp Y \mid T$ that are not conditionally independent.

Example 2.2. We consider three examples where faithfulness does (not) hold.

a) The faithfulness assumption generically holds for generalized additive models in which each variable is described by a sum of its parents

$$V_j = \sum_{V_i \in \mathrm{Pa}(V_j)} f_{ji}(V_i) + \varepsilon_j \,,$$

where $\varepsilon_j \perp\!\!\!\perp \mathrm{Pa}(V_j)$ and all f_{ji} are assumed to be non-zero functions.

b) To see what we mean by "generically" in the previous example, let us consider the simple linear model given by

$$\begin{aligned} V_1 &= \varepsilon_1 \\ V_2 &= \alpha_{21} V_1 + \varepsilon_2 \\ V_3 &= \alpha_{32} V_2 + \alpha_{31} V_1 + \varepsilon_3 \end{aligned}$$

In particular, the marginal effect of V_1 on V_3 is given by $\alpha_{31} + \alpha_{32}\alpha_{21}$. Therefore, if the parameters satisfy $\alpha_{32} = -\alpha_{31}/\alpha_{21}$, the parameters cancel out exactly, leading to $V_1 \perp\!\!\!\perp V_3$, and thus faithfulness being violated. This kind of fine-tuning is, of course, very specific, and if we assume the parameters to be sampled at random from a continuous probability distribution, this happens with probability zero.

c) Another example where the faithfulness assumption fails is

$$\begin{aligned} X_1, X_2 &\sim \mathrm{Bern}(0.5) \\ Y &= X_1 \oplus X_2 \,, \end{aligned}$$

in which case both X_1 and X_2 are marginally independent of Y, but the sets $\{X_1, X_2\}$ and $\{Y\}$ are not independent. Note that once more, this is a highly specific case: if either of the two X_i were $\sim \mathrm{Bern}(p)$ with probability $p \neq 0.5$, faithfulness would, in fact, hold.

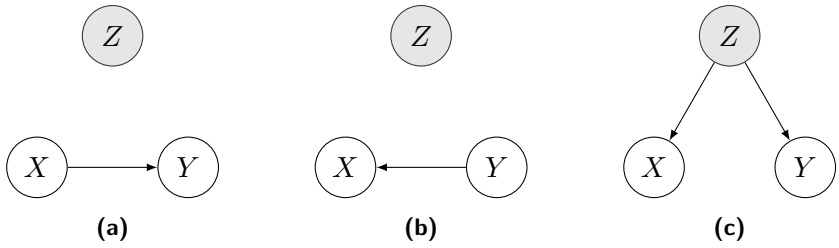


Figure 2.3: *The need for causal sufficiency.* When variable Z is unobserved, then the conditional independence relations holding between X and Y (no independences at all) are indistinguishable between all three causal graphs.

When the faithfulness assumption holds, all independences in $P(V)$ are reflected by d-separation in G. When both the MC and faithfulness hold, conditional independences in $P(V)$ correspond *exactly* to d-separation in G. That is, G is *minimal* for representing the conditional independence relationships of $P(V)$ in the sense that no graph with fewer edges can produce the same independence relationships. In particular, when we try to discover a graph G capturing the conditional independences in $P(V)$, sparsity constraints will be computationally convenient and provide an adequate inductive bias by enforcing a real property of the true graph.
However, These two conditions are insufficient to guarantee that a discovered graph G corresponds to the true causal network generating $P(V)$. For example, in Figure 2.3, in all generating models, the same conditional independences—none whatsoever—hold among the observed variables X, Y, regardless of whether X causes Y, Y causes X, or the unobserved Z causes both X and Y. Therefore, one more assumption, called *causal sufficiency*, is required for independences and d-separations to tell us anything.

Definition 2.4 (Causal Sufficiency)**.** All common parents of all observed variables V are themselves observed and thus included in V.

That is, we cannot distinguish between the generating models without assuming sufficiency based on only graphical criteria. When all three assumptions hold, the true causal graph can be recovered up to a certain equivalence relation called *Markov equivalence*. Consider, for example, the graphs in Figure 2.4 over variables X_1, X_2, X_3. There are no conditional independences between any of the variables, and all of the graphs capture precisely the same set of distributions. This set of graphs entailing exactly all the same independence constraints is called the *Markov Equivalence Class* (MEC) of the graph G (Pearl, 2009).

Definition 2.5 (Markov Equivalence)**.** Two graphs G and H are called Markov equivalent, denoted $G \sim H$ if they entail precisely the same set of d-separation

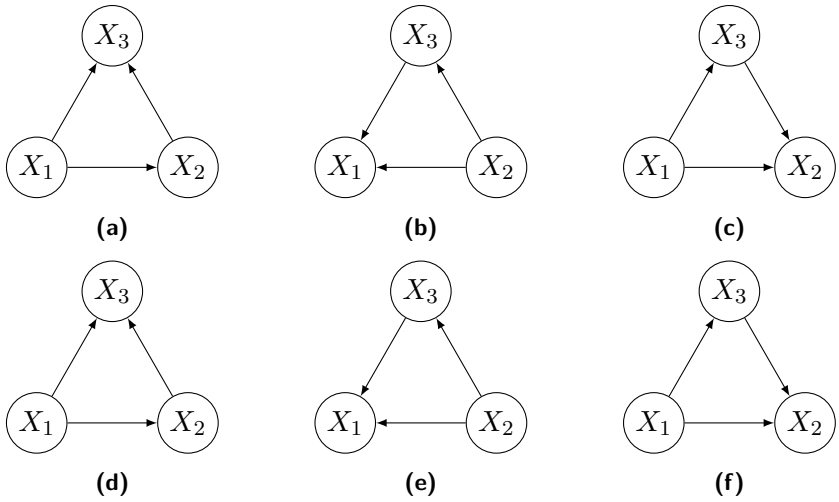


Figure 2.4: *Markov Equivalence.* All six networks contain precisely the same d-separation conditions, despite depicting different causal mechanisms generating the data.

relationships. The Markov equivalence class of a graph G is the set $\{H : H \sim G\}$ of all its Markov equivalent graphs.

It can be shown that all graphs H in the Markov equivalence class of G can be characterized by two things. First, they all share the same undirected edges, called the *skeleton* of G. Second, they share the same set of *unshielded colliders*, where a collider $X \to C \leftarrow Y$ is called unshielded if there exists no edge between X and Y (Meek, 1995).
Another way to formulate this is in terms of *Markov blankets* (sometimes called *Markov boundary*). The Markov blanket $\mathrm{MB}(Y)$ of the variable Y is the smallest set of variables $B \subset V$ not containing Y such that Y is independent of all other variables, $V \setminus (B \cup \{Y\})$, conditional on B. When the MC and faithfulness hold, the Markov blanket is unique and consists precisely of the parents, children, and other parents of its children. Graphs G and H are then Markov equivalent if and only if all Markov blankets $\mathrm{MB}_G(Y) = \mathrm{MB}_H(Y)$ coincide.
Unfortunately, as we will repeatedly see in the experiments in future chapters, the Markov equivalence classes of discovered graphs are generally exponentially large in the number of discovered edges, and thus not very insightful (Chickering, 2002a; He et al., 2015b). That is, a discovered MEC commonly leaves many of the causal directions unspecified, allowing for a great many different causal graphs and permitting only little causal insight into the result.
In many cases, this limited causal interpretability of the resulting MEC is insufficient for the task at hand, so we would like to go beyond Markov equiv-

alence. To do so, we employ *structural causal models* (SCMs; Wright, 1921; Pearl, 2009), also known as structural equation models or structural functional models. These models correspond to specific parametrizations of the causal mechanisms underlying the joint distribution $P(V)$.

Definition 2.6 (Structural Causal Model)**.** A structural causal model for the distribution $P(V)$ is a triple $(G, F, P(\varepsilon))$ such that G is a BN for $P(V)$ and $F = (f_1, \ldots, f_m)$ is a set of functions such that each V_i can be written as

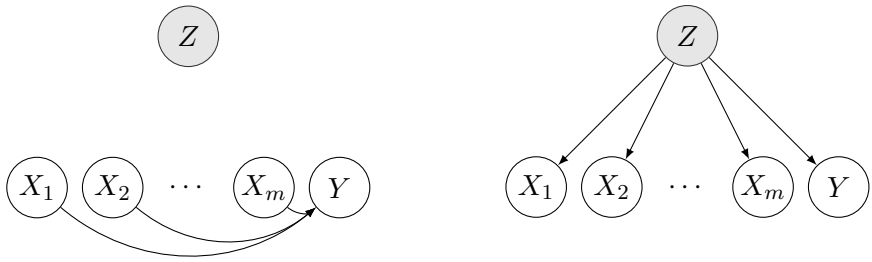
$$V_i = f_i(\mathrm{Pa}_i, \varepsilon_i)\,,$$

where ε are independent random variables and Pa_i are the parents of V_i in G.

We have already seen three examples of SCMs in Example 2.2 and will see many more throughout this thesis.
Conversely, given a set F of functions f_i and noise variables ε_i such that the joint distribution of V can be described by the relationships

$$V_i = f_i(V_{\mathcal{I}_i}, \varepsilon_i), \quad \mathcal{I}_i \subseteq \{1, \ldots, n\}\,,$$

then, if we construct the graph G by setting $\mathrm{Pa}_i = \mathcal{I}_i$, it will satisfy all conditions required of a BN for $P(V)$. Note that for any given $P(V)$ and BN G, there are infinitely many sets of functions F and noise variables ε which could generate $P(V)$. For example, if F, ε parametrize $P(V)$, then so does any pair $\tilde{F}, \tilde{\varepsilon}$ such that $\tilde{\varepsilon}_i = g_i(\varepsilon_i)$ and $\tilde{f}_i(\cdot, \cdot) = f_i(\cdot, g_i^{-1}(\cdot))$ for invertible functions g_i.
Once a specific choice of functions has been made, however, the causal network often becomes uniquely *identifiable*. That is, *all* of its edges become directed, thereby reducing the size of its MEC exponentially. In this case, it makes sense to speak of the *ground truth* network G^* uniquely describing the structure of $P(V)$. This is the case for linear non-Gaussian models (Shimizu et al., 2006; Hoyer et al., 2008b), linear Gaussian models with equal variances (Peters and Bühlmann, 2014), additive noise models (Peters et al., 2014), and post-nonlinear models (Zhang and Hyvärinen, 2009), among others. While the choice of model class in these works is generally inspired by theoretical tractability of the analysis, they both cover pragmatic choices for fields like chemistry (Ludden, 1991) and biology (Álvarez Buylla et al., 2016; Runge, 2023; Sakurada and Ishikawa, 2024), and also incorporate a wide range of cases (Cramér, 1936).
For the rest of this thesis, when we use $P(V)$, we assume that the MC, causal faithfulness condition, and causal sufficiency all hold over the entire set V. In particular, we assume that the set $V = X \cup Y \cup Z$ will contain all variables, including all confounders Z. Of course, when Z is not observed, and only X and Y are observed, causal sufficiency no longer holds over only these variables.



(a) Causal **(b)** Confounded

Figure 2.5: *Causal and Confounded graphs with nodes $X_1, \ldots, X_m$, Y, and Z.* Our goal will be to determine which of the two models fits our data better.

2.1.2 What's Our Problem?

With the basics out of the way, we can now describe the specific problem we are interested in solving. In the previous section, we have described conditions that allow for discovering causal graphs (or at least their MEC) when all relevant variables have been observed. Therefore, our question is: how would we know whether we have measured all relevant variables? In general, this is a challenging problem, so as a first attempt at discovering confounders, we consider two relatively simple cases of causal models in Figure 2.5. Our goal is to decide between the model in which X causes Y, depicted in Figure 2.5a, and the model where a set of confounders Z are responsible for all the observed correlations between X and Y, shown in Figure 2.5b.

Problem Statement. Given a sample $x, y = x^n, y^n$ from $P(X, Y)$, determine whether the distribution $P(X, Y)$ is the marginal distribution generated by a causal model where $X \to Y$ with independent Z, or where $X \leftarrow Z \to Y$.

Of course, to solve this problem, we will require further assumptions on the two competing models generating $P(X, Y, Z)$. Before we go into the details of this, however, let us consider some simple and intuitive ideas we could try to use to determine which of the two cases obtains.

2.1.3 Simple Solutions That Do Not Work

As we have noted above, when looking at the causal relationships between X and Y, the case where X causes Y and where a Z causes both X and Y are indistinguishable from looking at conditional independences between X and Y alone. In both cases, $X \not\perp\!\!\!\perp Y$ are not independent and cannot be rendered independent by conditioning on any other observed variables.

Moreover, given only a sample x, y from the distribution $P(X, Y)$ for which $X \not\perp\!\!\!\perp Y$ holds, but knowing nothing about Z or $P(Z)$, we cannot directly test if $X \perp\!\!\!\perp Y \mid Z$. A simple approach would be to see if we can *generate* a $\widehat{Z}$ such that $X \perp\!\!\!\perp Y \mid \widehat{Z}$; for example, through sampling or optimization. However, suppose the confounders $\widehat{Z}$ are not constrained in any way. In that case, it is easy to see that it is *always* possible to generate a $\widehat{Z}$ that achieves this independence, even when there are no confounders Z. A trivial example is to set $\widehat{Z} = X$.

Deciding based on model evidence rather than conditional independence fares no better. That is, we can try to find a $\widehat{Z}$ achieving better model evidence, $P(\widehat{Z})P(X|\widehat{Z})P(Y|\widehat{Z}) > P(\widehat{Z})P(X)P(Y|X)$. Besides having to choose a prior on Z, we already achieve this, once more, by initializing $\widehat{Z} = X$, regardless of whether there was a confounder in the underlying causal model or not.

Essentially, the problem with both of these approaches is that it is simply *too easy* to find a $\widehat{Z}$ where these conditions hold, which in large part is due to the fact that we do not take the complexity of $\widehat{Z}$ into account, and hence face the problem of overfitting. To avoid this, we take an algorithmic information theoretic approach, such that we can take both the complexity of $\widehat{Z}$ and its effect on X and Y into account in a principled manner.

2.2 The Algorithmic Model of Causality

What do we mean when we say that the above attempts at a solution do not take into account the complexity of $\widehat{Z}$, and how do we measure the complexity of a variable in the first place? Over the years, many different measures of complexity have been proposed in the field of complex systems and outside of it. These measures include measures of nonlinearity (Strogatz, 2018), interactions (Cliff et al., 2023), adaptation (Del Fabbro and Christen, 2022), self-organization (Kauffman, 1995), hierarchy (Corominas-Murtra et al., 2013), diversity (Ulanowicz, 2009), feedback loops (Turner and Baker, 2019), closeness to the "edge of chaos" (Langton, 1990), and circuit complexity (Vollmer, 1999). For the most part, these are difficult to compute and approximate, are highly specialized to the study of specific systems in which they give interesting insights, are often difficult to reason about, and have *no causal interpretation* whatsoever. Instead of these highly specific measures, we, therefore, make use of the *Kolmogorov complexity* from the field of *algorithmic information theory* (Li and Vitányi, 2009), which can be given a causal interpretation through the *algorithmic model of causality* (Janzing and Schölkopf, 2010). The underlying intuition is that the laws of physics are both *computable* and that the specific forms of these laws *simple*. In particular, a causal description of the universe should be simpler than a non-causal description of the universe.

As we have seen for SCMs, we can describe the distribution $P(V)$ by writing each variable $V_i \in V$ with its structural equation

$$V_i = f_i(\mathrm{Pa}_i, \varepsilon_i)\,,$$

where $\{\varepsilon_i\}_{V_i \in V}$ are mutually independent noise variables.
In the *algorithmic model of causality*, we assume that each function f_i is a *computable function* (Grzegorczyk, 1957). Unfortunately, introducing the entire formal framework of computability theory is beyond the scope of this section, so we shall give essentially correct but informal definitions of the required concepts (Cooper, 2017). A computable function will then be no more than a function that can be approximated arbitrarily well by a (computer) program. More specifically, we call a number $x \in \mathbb{R}$ computable if there exists an algorithm that can be used to approximate it arbitrarily well. For example, e is computable because it can be approximated by evaluating increasing partial sums of the form $\sum_{i=0}^{n} \frac{1}{n!}$. A computable sequence $(x_i)_{i\in\mathbb{N}}$ is a sequence of computable numbers that can all be computed by the same program, such as the sequence of all these approximations of e.

Definition 2.7 (Computable Function)**.** A function $f : D \subseteq \mathbb{R}^d \to \mathbb{R}$ is called computable if and only if the following conditions hold

a) For every tuple $(x_i)_{i\in\mathbb{N}}((x_{1i}, \ldots, x_{di}))_{i\in\mathbb{N}}$ of computable sequences of real numbers , the sequence $(f(x_i))_{i\in\mathbb{N}}$ is also computable, and
b) there exists a computable function $d : \mathbb{N} \to \mathbb{N}$ such that if $\|x - y\| < \frac{1}{d(n)}$ then $|f(x) - f(y)| < \frac{1}{n}$.

In essence, assumption a) tells us a computable function f is another program so that the composition of f and the program generating its inputs is again a program. Meanwhile, assumption b) is known as *effective continuity*, requiring us to be able not only to show that the function is continuous but also to give an explicit (and computable) description of how the ε and δ in the commonly used $\varepsilon - \delta$ definition of continuity are related to each other.
Note that for any compact (i.e., closed and bounded) domain, the set of all computable functions is *dense* in the space of continuous functions, and therefore capable of capturing any physically relevant aspect of the universe arbitrarily well (Grzegorczyk, 1957). If this were *not* the case, correct theories of the universe would be fundamentally *non-constructive*: while a mathematically accurate description of the laws of physics might exist, we would not be able to use them to construct models and make predictions about the world.
Given such computable functions $F = \{f_i\}_i$—which imply the structure in G—and $\varepsilon = \{\varepsilon_i\}_i$, all variables V_i take on deterministic values, so that the joint distribution $P(V)$ is described entirely by the functions f and the distribution of ε. In particular, $P(V)$ is computable so long as F and $P(\varepsilon)$ are. To measure

the complexity of $P(V)$, we can therefore measure the complexities of f and ε. To do this, we use algorithmic information theory (Li and Vitányi, 2009).

2.2.1 Kolmogorov Complexity and Algorithmic Mutual Information

Now that all our functions are considered computable, how would we use this fact to measure their complexity? The most natural and straightforward way of doing this is to measure the *length of the shortest program*, which computes, or approximates, the function (and then halts). This length depends on the choice of the specific programming language—more formally referred to as *universal Turing machine*—used. Fortunately, it turns out that the resulting quantity, called *Kolmogorov complexity*, is independent of this choice up to an additive constant (Li and Vitányi, 2009).

Definition 2.8 (Kolmogorov Complexity)**.** Let $\mathcal{U}$ be a universal Turing machine (Li and Vitányi, 2009). Then the the Kolmogorov complexity for finite strings and functions is defined as follows.

a) If $r \in \{0,1\}^*$ is a finite string

$$K(r) = \min_{p \in \{0,1\}^*} \{|p| : \mathcal{U}(p) = r\} \ .$$

b) If f is a (computable) function, then[1]

$$K(f) = \min_{p \in \{0,1\}^*} \left\{ |p| : \forall x \in \mathrm{dom}(f) \, \forall q \in \mathbb{N} : |\mathcal{U}(x,p,q) - f(x)| \leq \frac{1}{q} \right\} .$$

Here, the terms $\mathcal{U}(p)$ and $\mathcal{U}(x,p,q)$ can be understood simply as a computer (with a specific programming language) running the program p (with the inputs x, q). The Kolmogorov complexity is therefore defined as the *length of the best possible compression* of the string r, respectively function f, and makes use of *all* structure in the given object (Li and Vitányi, 2009). In particular, when this object is generated from an i.i.d. source, Kolmogorov complexity cannot be asymptotically improved upon, even when we are given access to the true model generating the object, as we will see in Equation (2.4).
Next, we define the *conditional* Kolmogorov complexity for string r given some side information s as the length of the shortest program which computes r starting from the input s (Li and Vitányi, 2009),

$$K(r \mid s) = \min_{p \in \{0,1\}^*} \{|p| : U(p,s) = r\} \ ,$$

[1] If f is not computable, then there is no such program and $K(f) = \infty$.

and similarly for two computable functions f and g, or combinations of computable functions f and strings r. For example, when $f(x) = f(x;r) = x^r$ where $r \gg 0$ is some natural number, then $K(f \mid r) \ll K(f)$ since the entire "difficulty" in computing f is knowing the parameter r.
Thus, when we see that $K(r \mid s) \ll K(r)$, this tells us that s contains a lot of algorithmic information about r that can be extracted by a short program p. In contrast, when $K(r \mid s) \approx K(r)$, then s contains no algorithmic information about r, e.g., when s is simply a random string sampled independently of r.

Example 2.3. Let us look at a few examples.

a) Let $s = r \oplus e$ where $e = (0, \ldots, 0, 1, 0, \ldots, 0)$ contains precisely one non-zero entry. Then, instead of computing r from scratch, we could run the following simple two-step program p: first, copy the input s. Second, specify the one index whose corresponding bit needs to be flipped. The first step is independent of the size of the input. The second stop requires a description of the index, which can be done in $O(\log n)$ bits. Together, they, therefore, need much fewer than n bits.
b) Let $r \in \{0,1\}^*$ be a string whose bits are i.i.d. uniformly distributed in $\{0,1\}$, and let s be another string such string, sampled independently from r. Then s gives us no information about r.
c) Let r, s be as in the previous case, but now let both have their entries distributed according to Bern(p) with $p \neq 0.5$. Then, despite s being statistically independent of r, it *does* provide us with relevant information: an (initial) estimate of the probability p, knowing which we can compress r much more cheaply than if we had not known it.

The above intuition leads us to the following definition of *algorithmic mutual information* (AMI) between two objects (Li and Vitányi, 2009).

Definition 2.9 (Algorithmic Mutual Information)**.** We define the algorithmic mutual information $I(a : b)$ between two objects a, b as

$$I(a : b) \stackrel{+}{=} K(a) - K(a \mid b^*) \stackrel{+}{=} K(b) - K(b \mid a^*)\,,$$

where t^* refers to the shortest program computing t and $\stackrel{+}{=}$ refers to equality up to additive constants independent of a and b.

Here, we need to condition on the shortest programs for a and b instead of a and b themselves to avoid asymmetries due to inefficient encodings. For example, if we use the function $f(x) = x^r$ from above, then we could write an arbitrarily complex program to compute it and thus make it difficult to gain any information from f about r with a shorter program than what was needed to compute r in the first place. By using the optimal compression f^*, we avoid such cases and ensure symmetry of the measure (Li and Vitányi, 2009).

We call two objects a and b *algorithmically independent* when $I(a : b) \stackrel{+}{=} 0$. In this case, their best joint compression is simply the concatenation of their individual compressions, i.e., $K(a, b) = K(a) + K(b)$. For more than two objects $a_1, \ldots, a_m$, we define algorithmic independence in terms of independence between pairs $a_{\mathcal{I}}, a_{-\mathcal{I}}$ for all subsets $\mathcal{I} \subseteq [m] \coloneqq \{1, \ldots, m\}$, where we write $a_{\mathcal{I}} = (a_i)_{i \in \mathcal{I}}$ and $a_{-\mathcal{I}} = a_{[m] \setminus \mathcal{I}}$. Equivalently, when all the a_i are independent, we have $K(a_1, \ldots, a_m) = \sum_{i=1}^{m} K(a_i)$.

2.2.2 Algorithmic and Statistical Independence

Since algorithmic independence is a rather abstract concept that is much less well-understood than statistical independence, in this section we explain some connections between the two notions of independence.
To start with, algorithmic independence is a stronger condition than statistical independence. Consider, for example, the case of $a \in \{0, 1\}^n$ being a long string such that $K(a) \stackrel{+}{=} n$. Assume that we obtain two identical copies of a from some process. Clearly, a is not algorithmically independent of a. However, we can easily find a distribution where the two strings could have been generated independently: Take the distribution P on $\{0, 1\}^n$ such that $P(X = a) = 1$. Then the two strings $x_1 = a, x_2 = a$ could be generated independently from the distribution $P \otimes P$. While this example may be highly specific, it nevertheless highlights the difference between the two approaches. Since two objects x, y are algorithmically dependent if and only if we can write $y = f(x^*)$ for some simpler function $K(f) < K(y)$, a reasonable question is to ask under which conditions a random variable X and the output $f(X)$ under a function f are statistically independent. The answer is, unfortunately, rather disappointing.

Lemma 2.1. *Let $X \sim P$ and f be some (measurable[2]) function. Then X and $f(X)$ are statistically independent if and only if f is constant.*

Proof sketch. If X and $f(X)$ are independent, then for all B we have $P(f(X) \in B) = P(f(X) \in B)^2 \in \{0, 1\}^2$, which can only happen when $f(X)$ is constant. □

This lemma tells us that the "gap" between the two types of independence is quite small. However, their semantics are nevertheless quite different. For example, in independent mechanism analysis (Gresele et al., 2021), the independence of nonlinear functions is phrased in terms of the orthogonality of the columns of the Jacobian matrix at every point. That is, variation in one dimension of the output is (locally) independent of variation in the other dimensions

[2]Since all computable functions are continuous and all continuous functions measurable, this is not a concern for us

of the output. While these assumptions *could* be formulated in terms of *some* statistical independence of *some* distributions, such a formulation would likely not be a fruitful way to think about the problem.

Another way to see that the difference between statistical and algorithmic independence is not very large is to compare algorithmic and statistical mutual information. Here, the statistical mutual information for two discrete random variables is given by (Cover and Thomas, 1999),

$$I(X;Y) = \sum P(x,y) \log \frac{P(x,y)}{P(x)P(y)} \geq 0\,.$$

Assuming that we have strings generated from an i.i.d. source on its symbols, we then have the following relationship between the mutual information and the average Kolmogorov complexity (Li and Vitányi, 2009),

$$I(X;Y) - K(P) \leq E_{X,Y\sim P(X,Y)}[I(x:y)] \overset{+}{\leq} I(X;Y) + 2K(P)\,.$$

That is, statistical dependence implies (almost) algorithmic dependence for "typical" samples from the distribution P. We can, therefore, use statistical dependencies between variables as a proxy for algorithmic dependencies. In essence, the *only* algorithmic information shared between two statistically independent samples from the same distribution P is that distribution.

2.2.3 Independence of Causal Mechanisms

How can we apply these notions of complexity and independence to a causal model described by the DAG G? To think about this, let us consider the Markov property of G for the distribution $P(V)$,

$$P(V) = \prod_{V_i \in V} P(V_i \mid \mathrm{Pa}_G(V_i))\,.$$

As we have seen, many graphs G can satisfy this factorization, even among those capturing precisely the independence constraints present in the distribution $P(V)$. To distinguish between different graphs, we require a stronger Markov condition. Not only should the variables V_i be generated solely from their parents $\mathrm{Pa}_G(V_i)$, but the mechanisms by which they are generated should also be independent of the mechanisms generating the other variables.

This idea is called *independence of causal mechanisms* (Janzing and Schölkopf, 2010). That is, all mechanisms $P(\cdot \mid \mathrm{Pa}_G(\cdot))$ in the factorization above should be independent of each other. When this independence is framed in terms of al-

gorithmic independence, this means precisely that the complexity "factorizes",

$$K(\{P(V_i \mid \mathrm{Pa}_G(V_i))\}_{V_i \in V}) \stackrel{+}{=} \sum_{V_i \in V} K(V_i \mid \mathrm{Pa}_G(V_i))\,.$$

Because of the similarity of this equation with factorization of $P(V)$ in the causal Markov condition, this notion is called the *algorithmic Markov condition* (Janzing and Schölkopf, 2010).

Postulate 2.1 (Algorithmic Markov Condition)**.** *A causal DAG G is admissible as causal graph for the distribution $P(V)$ if and only if*

$$K(P(V)) \stackrel{+}{=} \sum_{V_i \in V} K(P(V_i \mid \mathrm{Pa}_G(V_i)))\,. \tag{2.1}$$

Being an *algorithmic* version of the causal Markov condition, this means that not only is a variable statistically independent of its non-descendants given its parents, but so are the causal mechanisms of all variables. Furthermore, when we generate V via an SEM $(G, F, P(\varepsilon))$ in which all components are algorithmically independent, then G is indeed an admissible causal graph for $P(V)$. Therefore, the simplest description of an algorithmic Markov process is the causal model generating the distribution $P(V)$ (Janzing and Schölkopf, 2010).
In particular, in the bivariate case, it states that if X causes Y, then

$$K(P(X)) + K(P(Y \mid X)) \stackrel{+}{\leq} K(P(Y)) + K(P(X \mid Y))\,,$$

which can be used to determine the direction of causality where purely statistical criteria cannot be used to distinguish between the direction of the edge.
Since the algorithmic Markov condition is a stronger requirement than the causal Markov condition, the former implies the latter. How about faithfulness? While there is no general relationship between the CMC and faithfulness, it turns out that the algorithmic Markov condition often *does* entail faithfulness. To see why this is so, we first look at an example where it is *not* the case.

Example 2.4. Let X_1, X_2, X_3 satisfy the following deterministic equations,

$$\begin{aligned} X_1 &= \varepsilon_1 \\ X_2 &= f(X_1) \\ X_3 &= X_1 + X_2\,, \end{aligned}$$

where f is bijective and both f and f^{-1} are sufficiently complex. It is easy to see that in this model, all necessary information about X_3 is already available

in either of the variables X_1 or X_2, and once we include either of the edges $X_1 \to X_3$ or $X_2 \to X_3$, causal faithfulness would require the other to be absent. However, the complexity of $K(P(X_3 \mid X_1, X_2))$ is very small, while the complexity of both $K(P(X_3 \mid X_1))$ and $K(P(X_3 \mid X_2))$ are large, requiring an additional encoding of f or f^{-1}. Hence, the model with the shortest encoding is the complete graph G, which is not faithful to the distribution P.

The reason why this example does not satisfy faithfulness is that the equations are fully deterministic without any exogenous noise (besides X_1). In contrast, for non-deterministic models, it has been shown that when a graph G satisfies the algorithmic Markov condition for $P(V)$, then it will also satisfy the faithfulness condition (Lemeire and Janzing, 2013).
Note that when there is no admissible graph G for $P(V)$, other explanations, such as latent confounding or models with selection bias, must be considered. We study how we can determine that confounding is likely to be the issue with a proposed graph over the observed variables X.

2.3 Telling Causal from Confounded by Simplicity

The non-existence of admissible graphs for the distribution $P(X, Y)$ gives us a hint towards the existence of a latent confounder Z, but it does not provide us with a constructive criterion for determining whether such non-existence is due to confounding, selection bias, or other issues. To derive such a constructive criterion, we instead *explicitly* include latent confounders Z in the evaluation of the model complexity and show that by comparing between models with and without such Z, we can consistently and fairly determine if such a latent Z should exist. We then explain how our setup can be used in practice by combining the statistically well-founded approximations of Kolmogorov complexity provided by the minimum description length principle (Grünwald, 2007) with latent factor models to derive explicit model classes for comparing between models with and without confounders (Loehlin, 1998).

2.3.1 Confounding in the Algorithmic Model of Causality

Since we assume that X, Y, Z with $Z \to X, Y$ have a joint causal DAG satisfying causal sufficiency, there is an admissible causal DAG satisfying over this (unknown) larger set of variables. However, when we observe only the variables X and Y, then causal sufficiency over only these two variables is violated, and there will generally be no admissible network G for $P(X, Y)$ in the sense of the algorithmic Markov condition above. Specifically, this means that any presumed set of causal functions relating the variables X, Y to each other will not be algorithmically independent (Janzing and Schölkopf, 2010).

For simplicity of notation, let us for now include Y among the X_i and consider whether the variables X are confounded by some Z. We know that no admissible causal graph exists if the variables X are confounded. But, being a purely negative statement about the non-existence of a suitable DAG, it does not give us any information as to *why* no such admissible graph exists. Nor will we be able to derive from it an easily computable criterion for discovering latent confounding down the line. Therefore, we explicitly model the unobserved confounder Z to obtain such a readily operationalizable criterion.
That is, we apply the algorithmic Markov condition on the extended variable set X, Z including the (correct) latent variables $Z = (Z_1, \ldots, Z_l)$, where the Z_j's are assumed to be independent. Then, if we knew the joint distribution $P(X, Z)$, we could identify the corresponding minimal Kolmogorov complexity causal graph G as the minimizer,

$$K(P(X,Z)) = \min_G \sum_{i=1}^{m} K(P(X_i \mid \mathrm{Pa}_i)) + \sum_{j=1}^{k} K(P(Z_j))\,, \tag{2.2}$$

where by Pa_i we denote here the parents of X_i among $\{X_j, Z_k\}$ in the full network. Adding the terms $K(P(Z_j))$ mirrors our assumption that all Z_j are jointly independent and that no reverse causality from X to Z exists.
Two questions arise from this idea. First, even knowing $P(X, Z)$, could we actually obtain $K(P(X, Z)) < K(P(X))$? That is, can the distribution become simpler by including more variables? At first glance, it seems unlikely that this will work. However, an intuitive example where this is plausible is for mixture models. When we consider a distribution P that is a mixture of, e.g., Gaussian distributions, then the "natural" parametrization we give is to state the mixture probabilities and the parameters of the individual distributions in the mixture. In general, even when we have an analytically tractable distribution $P(X, Z)$, the simplest way we know to describe $P(X)$ is often by *explicitly* modeling $P(X, Z)$ and marginalizing out Z. Second, since we do not even know which set of latent confounders Z is correct, and thus which distribution $P(X, Z)$ would be appropriate for the proposed comparison, can we ask instead if there exists *any* distribution $P(X, Z)$ with correct marginal distribution $P(X)$ for which $K(P(X, Z)) < K(P(X))$ holds? Moreover, if such a distribution $P(X, Z)$ indeed exists, does that mean that the data comes from a model involving a latent variable? The following Theorem gives the, perhaps surprising, positive answer to these questions.

Theorem 2.2 (Kolmogorov Does Not Incorrectly Detect Confounders)**.** *For any distribution $P(X)$, the following inequality holds*

$$\inf_{P(X,Z)\in\mathcal{P}} K(P(X,Z)) \overset{+}{\leq} K(P(X))\,,$$

where the infimum is over the set $\mathcal{P}$ of all joint distributions $P(X, Z)$ with fixed marginal $P(X)$ and jointly independent Z. Conversely, if a joint distribution $P(X, Z) \in \mathcal{P}$ exists such that the inequality

$$K(P(X, Z)) \overset{+}{<} K(P(X)),$$

holds;[3] *then the true generating mechanism of X includes latent variables influencing some subset $X_S \subseteq X$ of the observed variables.*

Proof sketch. The first part follows directly from choosing a distribution $P(Z = a) = 1$, which has constant complexity independent of $P(X)$. For the second part, if there are no confounders for X, then all information required to compress the distribution $P(X)$ is already available in the graph G^*_X. □

This formulation gives us a principled manner to identify whether a given set of variables Z is a (likely) confounder of X. With the above, we can score the hypothesis $Z \to X$. However, it also allows us to fairly score the hypothesis that X is unaffected by any confounders. This is the case if none of the proposed $P(X, Z)$ obtain lower Kolmogorov complexities than $P(X)$. Equivalently, when $P(Z) = \delta(Z, z)$ is a deterministic point mass with $P(Z = z) = 1$, then $K(P(X, Z)) \overset{+}{=} K(P(X))$ so that the case of unconfounded distributions can be included within this larger framework. That is, the comparison between complexities $K(P(X))$ and $K(P(X, Z))$ is not biased towards either case of always preferring the confounded or always preferring the unconfounded model. What Theorem 2.2 tells us is, therefore, that if we can find *any* set of additional variables Z such that $K(P(X, Z)) < K(P(X))$, then there in fact does exist *some* true set of latent confounders Z affecting (some subset of) the variables X. Note, however, that it does not guarantee that if a confounder is involved in the true generating process, it will always be found. For a simple counterexample, consider $Z \sim N(0, \sigma^2)$ and noiseless relations $X = aZ, Y = bZ$. Then the model containing only X and Y can be described by two parameters $a^2\sigma^2$ and b/a, whereas the model involving Z requires three parameters σ^2, a, b. It is, therefore, not guaranteed that the model involving Z should be strictly less complex than the one containing only X and Y.
Given that a distribution $P(X, Z)$ is not guaranteed to exist, why *should* we ever expect such a distribution $P(X, Z)$ to exist? We will provide an intuitive explanation of this expectation in the following hypothesis.

[3] Here, $\overset{+}{<}$ denotes inequality by more than the additive constant permitted due to choice of the universal Turing machine $\mathcal{U}$. Equivalently, $K(P(X)) \overset{+}{\nleq} K(P(X, Z))$.

Hypothesis 2.1 (Confounders Are Likely Generically Recoverable)**.** *Let $P(X, Z)$ be a computable distribution generated by the causal model involving edges $Z \to X$. Then, for most distributions of this kind, the following hold*

a) The simplest way to compute $P(X)$ is to compute $P(X, Z)$, and then marginalize over Z as $P(X) = \int_Z P(X, Z)dZ$, so that

$$K(P(X)) \stackrel{+}{=} K(P(X, Z)) + K(I)\,,$$

where I is the program for computing the marginalization.

b) The algorithm I for provably approximating the integral uniformly up to a desired accuracy is not of constant complexity in that it depends on the underlying distribution $P(X, Z)$ itself so that

$$K(P(X)) \stackrel{+}{\not\leq} K(P(X, Z))\,.$$

Hence, $P(X)$ will, in general, be strictly more complex than $P(X, Z)$.

While we cannot provide any proof of the hypothesis in this very general form, we will see it borne out throughout this thesis both in the theoretical analysis of the specific scores we employ to evaluate competing causal models, as well as in our empirical results. In particular, we will see that as we include more observed variables X, cases in which the unconfounded model requires fewer parameters than the confounded model become more exceptional.

Overall, we have seen that by using the algorithmic Markov condition on an extended set of variables, we can determine the most likely causal model by simply comparing the scores for different $P(X, Z)$ and choosing the one with the lowest Kolmogorov complexity. This approach does not suffer from the problems we saw in Section 2.1.3 since we now explicitly take the complexity $P(Z)$ of the latent confounder into account. Moreover, this formulation allows us to consider *any* distribution $P(X, Z)$ with *any* type of latent factor Z.

Next, we turn to two aspects of this framework that we have not considered in our analysis so far. First, we do not know the true distribution $P(X, Z)$, nor even distributions $P(X)$ or $P(Z)$. Instead, we only have a sample x from $P(X)$ from which we can approximate $\hat{P}(X)$, but which gives us no explicit information about Z, $P(Z)$ or the joint $P(X, Z)$. Second, Kolmogorov complexity is not computable and the criterion therefore not directly applicable in practice (Li and Vitányi, 2009). We will next show that both of these aspects can be addressed by employing the Minimum Description Length (MDL) principle (Grünwald, 2007) with an appropriate model class.

2.4 Minimum Description Length

Out of the two issues outlined above, we start with the first: instead of access to $P(X)$, we are given only a sample x. So even if we *could* compute the Kolmogorov complexity $K(x)$ of the sample, how would we compute $K(P(X))$? Fortunately, there is a correspondence between $K(P(X))$ and $K(x)$ in expectation, and thus for generic samples from $P(X)$, via (Marx and Vreeken, 2022)

$$E_{x\sim P(X)}[K(x)] \stackrel{+}{=} K(P(X)) + H(X)\,,$$

where $H(X)$ is the entropy of $P(X)$. Considering the bivariate case for X, Y with samples x, y, in the direction $X \to Y$ we obtain the approximation

$$\begin{aligned} K(x) + K(y \mid x) &\stackrel{+}{=} K(P(X)) + H(X) + K(P(Y \mid X)) + H(Y \mid X) \\ &= K(P(X)) + K(P(Y \mid X)) + H(X, Y)\,, \end{aligned}$$

and analogously for the reverse direction $Y \to X$. Since $H(X, Y)$ is the same in both directions, we can cancel the term out and obtain

$$\begin{aligned} K(P(X)) + K(P(Y \mid X)) &\stackrel{+}{\leq} K(P(Y)) + K(P(X \mid Y)) \\ &\stackrel{\sim}{\longleftrightarrow} \\ K(x) + K(y \mid x) &\stackrel{+}{\leq} K(y) + K(x \mid y)\,, \end{aligned}$$

which holds approximately for generic samples from the joint distribution $P(X, Y)$. When we are interested in determining whether a set of variables X is confounded, we can similarly compare the values $K(x)$ and $K(x, z)$ instead of $K(P(X))$ and $K(P(X, Z))$, which we have no access to.

Of course, since Kolmogorov complexity is not computable (Li and Vitányi, 2009), even this does not suffice by itself. To make further progress, we use upper bounds on $K(x)$ and $K(x, z)$ by restricting the class of programs over which the minimum is taken. This is precisely what the Minimum Description Length (MDL) principle (Rissanen, 1978) does: it provides a statistically well-founded approach to approximate $K(\cdot)$ from above. To achieve this, rather than considering all programs as in the definition of K, in MDL, we consider a model class $\mathcal{M}$ for which we *know* that every model $M \in \mathcal{M}$ will halt, and identify the best model $M^* \in \mathcal{M}$ as the one that describes the data most succinctly, i.e., the one providing the best lossless compression of the data.

This is also known as *two-part*, or crude MDL, where we score models $M \in \mathcal{M}$ by first encoding the model and then the data given that model,

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

where $L(M)$ and $L(x \mid M)$ are code length functions for the model and the data conditional on the model, respectively. When $\mathcal{M}$ is a parametric model class with parameter space Θ, then we can write this in the form

$$L(x, M) = -\log p(x \mid \theta) - \log p(\theta)\,, \tag{2.3}$$

where θ is the parameter corresponding to M and the distributions are given by the density $p(\cdot) \propto e^{-L(\cdot)}$ (Grünwald, 2007). This is essentially the familiar maximum a posteriori (MAP) estimate of the parameters θ (Bishop and Nasrabadi, 2006), although there are philosophical differences in the interpretations and desirable properties of the estimates (Grünwald, 2007).
The two-part MDL approximation of Kolmogorov complexity is an adequate approximation in two senses. First, if we endow the class $\mathcal{M}$ of all programs that output x and halt with the Solomonoff prior $\mu(a) \propto 2^{-K(a)}$ (Solomonoff, 1964a,b), then the MDL-optimal score for the model is (Li and Vitányi, 2009)

$$\arg\min_{p \in \mathcal{M}} L(x, p) \stackrel{+}{=} K(x \mid p^*) + K(p^*) \stackrel{+}{=} K(x)\,,$$

where p^* is the minimizer of the left-hand side. This approach is known as *Ideal MDL*, since it is based on all possible models (Li and Vitányi, 2009).
Equivalently, we can formulate MDL as an upper bound of K as follows,

$$\begin{aligned} K(x) &= \min_p K(x \mid p) + K(p) \\ &\leq \min_{p \in \mathcal{M}} K(x \mid p) + K(p) \\ &\stackrel{+}{\leq} \min_{p \in \mathcal{M}} L(x \mid p) + K(p) \\ &\stackrel{+}{\leq} \min_{p \in \mathcal{M}} L(x \mid p) + L(p)\,, \end{aligned}$$

where in the last line, we used the fact that the Solomonoff prior μ is a universal approximation for every other positive distribution (Li and Vitányi, 2009).
Second, when $x = x^n \sim P$ is generated from a distribution P corresponding to the model $M \in \mathcal{M}$, then we have (Li and Vitányi, 2009)

$$\frac{1}{n} E_{x \sim P}\left[K(x)\right] = H(P) + o(1) = \min_{M \in \mathcal{M}} \frac{1}{n}\left(L(M) + L(x \mid M)\right) + o(1)\,. \tag{2.4}$$

That is, when we have a large (and thus generic) sample from P, the average MDL score and average Kolmogorov complexity of a representative sample from P are almost identical. Since all the structure that can be consistently extracted from a sample $x \sim P$ is contained in P, Kolmogorov complexity cannot do any better than a score that extracts precisely this structure.

Two-part MDL often works well in practice, and as we have seen, it has a solid theoretical relationship to Kolmogorov complexity. Under the assumption of causal sufficiency, it has also been used to discover causal directions in both the bivariate (Budhathoki and Vreeken, 2017; Marx and Vreeken, 2017) and multivariate cases (Mian et al., 2021; Mameche et al., 2023). However, by encoding the model separately, it introduces additional choices into the modeling process. That is, by encoding the model separately from the data, we impose an essentially arbitrary interpretation of "structure" and "noise" on the data, which may not be desirable. These choices can furthermore lead to problems in the finite sample regime (Grünwald, 2007). In one-part MDL—also known as *refined* MDL—we avoid these choices by encoding the data using the entire model class jointly, for example as in the Bayesian MDL score introduced below. The reason to choose such encodings is that they are pointwise asymptotically min-max optimal (Grünwald, 2007). That is, no matter which data point x we obtain, the (average) refined score for x is always within an additive constant of the optimal score we could obtain by using the (unknown) optimal model $M^*(x)$. Denoting such a refined MDL score by $L(x; M)$, it satisfies

$$\max_{x=x^n} \frac{1}{n} \left(L(x; \mathcal{M}) - L(x, M^*(x)) \right) = o(1) \,.$$

There exist different forms of refined MDL codes (Grünwald, 2007). Since we are interested in whether there exists a confounder for our data, we want to determine whether X is confounded in a broad sense without having to care about the specific optimum of a given model class. As such, we propose to use the Bayesian MDL score for the model class $\mathcal{M}$,

$$L(x; \mathcal{M}) = -\log \int_{M \in \mathcal{M}} P(x \mid M) P(M) dM \,,$$

where $P(M)$ is a prior distribution on $\mathcal{M}$. We can think of $\mathcal{M}$ as one of the model classes "X confounded by some Z" or "X not confounded", and sometimes more specifically as "subset $X_{\mathcal{I}} \subseteq X$ confounded by some Z", depending on the precise task at hand, and will be made clear where relevant. Note that for any $M \in \mathcal{M}$, we have (Grünwald, 2007)

$$L(x; \mathcal{M}) \leq L(x \mid M) + L(M) \,,$$

so that in the idealized case of $\mathcal{M}$ containing all halting programs, we obtain the same relationship with Kolmogorov complexity as above.

When $\mathcal{M}$ is parametric, we can write $L(x; M)$ in the more common form

$$L(x; \mathcal{M}) = -\log \int_{\theta \in \Theta} P(x \mid \theta) P(\theta) d\theta \,, \tag{2.5}$$

where each θ corresponds to a different model $M \in \mathcal{M}$ of X. Just as the two-part MDL score of Equation (2.3) corresponds to MAP, this score corresponds to the evidence, also known as marginal likelihood (de Carvalho et al., 2019). For our specific use case, the relevant model classes are the causal model class $\mathcal{M}_{\text{ca}}$, in which we model data with respect to a causal factorization within X, and the confounded model class $\mathcal{M}_{\text{co}}$, in which the confounder models all correlations. That is,

$$L(x; \mathcal{M}_{\text{ca}}) = -\log \int_{\theta \in \Theta_{\text{ca}}} P(x \mid \theta) P(\theta) d\theta$$
$$L(x; \mathcal{M}_{\text{co}}) = -\log \int_{\theta \in \Theta_{\text{co}}} P(x \mid z, \theta_{Z \to X}) P(z \mid \theta_Z) P(\theta_{Z \to X}, \theta_Z) dz d\theta \,,$$

where by independence of causal mechanisms, we assume that the parameter $\theta \in \Theta_{\text{co}}$ can be split into the parameters θ_Z determining the distribution of the confounder Z, and $\theta_{Z \to X}$ determining the influence of Z on X.
We next describe the two model classes $\mathcal{M}_{\text{ca}}$ and $\mathcal{M}_{\text{co}}$ we use when determining whether the data is causally connected or co-caused by a latent confounder.

2.4.1 Causal Model

We begin by introducing our causal model $\mathcal{M}_{\text{ca}}$. As a proof of principle, and because we will also be using linear latent factor models, we use Bayesian linear regression from Y onto X, which is given by (Bishop and Nasrabadi, 2006)

$$\begin{aligned} X &\sim N(0, \sigma_x^2 I) \\ w &\sim N(0, \sigma_w^2 I) \\ Y \mid X, w &\sim N(w^\top X, \sigma_y^2) \,. \end{aligned} \tag{2.6}$$

We can, therefore, write the corresponding score as

$$L(x, y; \mathcal{M}_{\text{ca}}) = -\log \left(P(x) \int P(y \mid x, w) P(w) dw \right) ,$$

where we assume the parameters σ_x^2 and σ_y^2 to be fixed. Note that while the $P(x)$ term does not affect the integral in any way, we nevertheless need to keep it in order to compare this score to that of the confounded model later on.

2.4.2 Latent Factor Models

Since we do not distinguish between X and Y in this model, we will consider $Y = X_{m+1}$ and talk again about the distributions $P(X)$ and $P(X, Z)$ where

convenient. As we have noted in Section 2.3.1, there are infinitely many possible distributions $P(X, Z)$ which entail the marginal distribution $P(X)$, and hence we have to make further choices to make the problem feasible. In our setting, where we want to determine whether X causes Y or whether the variables are jointly confounded, a natural choice is to use latent factor modeling (Loehlin, 1998). That is, we assume that the distribution over X, Z is of the form

$$P(X, Z) = P(Z) \prod_{i=1}^{m+1} P(X_i \mid Z)\,,$$

where the distribution of Z can, in principle, be arbitrarily complex. Not only does this give us a very clear and interpretable hypothesis, namely that given Z, all X_i should be independent of one another, i.e., $X_{\mathcal{I}} \perp\!\!\!\perp X_{\mathcal{I}'} \mid Z$ for any two disjoint $\mathcal{I}, \mathcal{I}' \subseteq [m] = \{1, \ldots, m\}$. It also corresponds to the notion that Z should explain away *as much* of the information shared within X as possible—in particular, causal mechanisms of different X_i are rendered independent (Equation (2.2)). Moreover, from a more practical perspective, it is also a well-studied problem for which advanced techniques exist, such as Probabilistic PCA (PPCA; Tipping and Bishop, 1999), Factor Analysis (Loehlin, 1998), Gaussian process latent variable models (GPLVM; Lawrence, 2005), and Deep Generative Models (Kingma and Welling, 2014; Rezende and Mohamed, 2015; Ranganath et al., 2015; Bond-Taylor et al., 2022).
For the sake of simplicity, we focus here on using PPCA to find latent confounders for $\mathcal{M}_{\text{co}}$ so that we model our data as being generated by the process

$$\begin{aligned} Z_j &\sim N(0, \sigma_z^2 I) \\ W_j &\sim N(0, \sigma_w^2 I) \\ X \mid Z, W &\sim N(W^\top Z, \sigma_x^2 I)\,, \end{aligned} \tag{2.7}$$

which is appropriate if we deal with real-valued variables with linear relationships and assume Gaussian noise. If the data does not follow these assumptions, one of the abovementioned models may be more appropriate. An appealing aspect of PPCA is that by marginalizing over Z we can rewrite it in terms of only the matrix W (Tipping and Bishop, 1999), so that

$$\begin{aligned} W_i &\sim N(0, \sigma_w^2 I) \\ X \mid W &\sim N(0, \sigma_z^2 W W^\top + \sigma_x^2 I)\,, \end{aligned} \tag{2.8}$$

which dramatically reduces the computational effort and allows us to provide consistency guarantees for our method. Furthermore, since σ_z^2 rescales W, it can be subsumed in σ_w^2, and we will assume that $\sigma_z^2 = 1$.
While in the simple form, PPCA assumes linear relationships, it is possible

to model nonlinear relationships by adding features to conditional distribution $X \mid Z, W$, e.g., using polynomial regression of X on Z. This is similar to using GPLVMs (Lawrence, 2005), which replace the linearity assumption with more complex kernel-based models. While this increases the ability of our model to capture richer relationships between the latent and observed variables, it also comes with both an increase in computation, as well as a decrease in theoretical guarantees, since the simplification of Equation (2.8) no longer holds.

Now, given this model class, what will our score $L(x; \mathcal{M}_{\text{co}})$ be? Given the simplified version of Equation (2.8), we can instantiate Equation (2.5) via

$$L(x; \mathcal{M}_{\text{co}}) = -\log \int P(x \mid W) P(W) dW \,,$$

where $\mathcal{M}_{\text{co}}$ is the PPCA model class with fixed σ_w^2 and σ_x^2. Note in particular that there is no need, or opportunity, to instantiate the confounding factor with specific values z for a given sample x so that the concerns from Section 2.1.3 about overfitting the latent confounder do not apply.

We can now put all the pieces together to determine whether a pair X, Y is more likely causally related or confounded by an unobserved Z.

2.4.3 Confounded or Causal?

With the theory we have developed, testing whether $X \to Y$ or $X \leftarrow Z \to Y$ is a better fit is straightforward. To do so, we consider the two model classes $\mathcal{M}_{\text{ca}}$ and $\mathcal{M}_{\text{co}}$ defined in the previous sections and compare which of the two models obtains a lower score on our data x, y. To ease our notation, we write L_{ca} and L_{co} instead of $L(\cdot; \mathcal{M}_{\text{ca}})$, respectively $L(\cdot; \mathcal{M}_{\text{co}})$.

For the causal model $\mathcal{M}_{\text{ca}}$ we can compute the code length of the data x, y as

$$\begin{aligned} L_{\text{ca}}(x, y) &= -\log \left(P(x) \int P(y \mid x, w) P(w) dw \right) \\ &\approx -\log \left(P(x) N^{-1} \sum_{j=1}^{N} P(y \mid x, \hat{w}_j) \right) \,, \end{aligned}$$

where we approximate the integral by Monte Carlo sampling N weight vectors $\hat{w}_j$ from the distribution defined by Equation (2.6). Of course, the more samples we take, the better the approximation.

Second, we consider the *confounded* model class $\mathcal{M}_{\text{co}}$, where all correlations between X and Y are explained entirely by a hidden confounder modeled by

the (simplified) PPCA model of Equation (2.8), i.e.

$$\begin{aligned} L_{\text{co}}(x, y) &= -\log \int P(x, y \mid W) P(W) dW \\ &\approx -\log N^{-1} \sum_{j=1}^{N} P(x, y \mid \hat{W}_j) \,, \end{aligned}$$

where the N samples for $\hat{W}_j$ are drawn from the PPCA model, i.e., according to Equation (2.8). As for the causal case, the more samples we consider, the better the approximation, but the higher the computation cost.

To determine which model is a better fit for the data, we use the rule

$$L_{\text{co}}(x, y) - L_{\text{ca}}(x, y) \begin{cases} \ll 0 & \text{if the data is likely confounded} \\ \gg 0 & \text{if the data is likely causal} \\ \approx 0 & \text{if both models are roughly equally good} \,. \end{cases}$$

Our decision rule to decide between the confounded and the causal case is to choose "confounded" when the sign of the difference is negative and to choose "causal" when it is positive. We call our method based on fitting the two models and using this decision rule **Co**nfounded or **Ca**usal, CoCa.

A natural question is how large the difference between L_{ca} and L_{co} should be for us to be confident in classifying into either of the two classes. While not directly applicable to our case, we can leverage the intuition provided by the *no hypercompression inequality* from information theory (Cover and Thomas, 1999). To explain this inequality, let x be data generated from $P(X)$ and let $Q(X)$ be another distribution. Then Q compressing the data better is unlikely,

$$P\left(-\log Q(x) < -\log P(x) - k\right) < 2^{-k} \,.$$

That is, compared to the true distribution P, any other distribution can compress the data x better by at least k bits on only a small fraction of the data. Thus, if our data were, in fact, generated from the model $\mathcal{M}_{\text{ca}}$, then the probability of $\mathcal{M}_{\text{co}}$ outperforming it by $k \gg 0$ bits should be small, and vice versa.

In the real world, our data is, of course, unlikely to be generated from either of these distributions exactly, but the intuition is nevertheless useful. Furthermore, since our scores L_{co} and L_{ca} depend on the chosen hyper-parameters σ_x, σ_w, and σ_y, as well as the sample size n, we may not get comparable results for different data sets. To make our scores comparable between different data sets, we therefore introduce the *confidence* score

$$\mathcal{C} = \frac{L(X, Y; \mathcal{M}_{\text{co}}) - L(X, Y; \mathcal{M}_{\text{ca}})}{\max\left\{L(X, Y; \mathcal{M}_{\text{co}}), L(X, Y; \mathcal{M}_{\text{ca}})\right\}} \,, \tag{2.9}$$

which is simply a normalized version of the above difference that accounts for both the intrinsic complexities of the data as well as the number of samples. The confidence C can be interpreted as the relative gain of one model over the other. If the absolute value of $\mathcal{C}$ is small, both model classes explain the data approximately equally well, i.e., we are not very confident in our result and should perhaps refrain from making a decision.
Can we give any theoretical guarantees that this approach will, in fact, work? When $\dim(Z) < \dim(X)$, the answer will generally be yes. That is, we can show that our method is statistically *consistent*, meaning that we will pick the correct model class in the limit of infinitely many samples. This is based on general results of MDL consistency for deciding between model classes when the data is generated by a model in one of these classes (Grünwald, 2007).

Theorem 2.3 (Consistency of CoCa)**.** *Let x^n, y^n be n be samples from the distribution M^* which is contained in $\mathcal{M}_{\mathrm{ca}} \cup \mathcal{M}_{\mathrm{co}}$. Then*

$$\lim_{n\to\infty} n^{-1}\left(L_{\mathrm{co}}(x^n, y^n) - L_{\mathrm{ca}}(x^n, y^n)\right) \begin{cases} \leq 0 & \text{if } M^* \in \mathcal{M}_{co} \\ \geq 0 & \text{if } M^* \in \mathcal{M}_{ca}\,, \end{cases}$$

with strict inequalities if M^* is contained in precisely *one of the two classes.*

Proof sketch. When $M^* \in \mathcal{M}_{\mathrm{ca}}$, then from MDL theory we know that L_{ca} is asymptotically optimal (Grünwald, 2007). Conversely, when $M^* \in \mathcal{M}_{\mathrm{co}}$, then L_{co} is asymptotically optimal. □

Thus, since $\dim(Z) < \dim(X)$ ensures that sets of non-degenerate models in $\mathcal{M}_{\mathrm{ca}}$ and $\mathcal{M}_{\mathrm{co}}$ are mutually exclusive, in the limit, we will infer the correct conclusion if the true model is within the model classes we assume. Here, by degenerate models, we refer to those models in which at least $\dim(X) - \dim(Z)$ parameters are 0. Since we assume that the parameters are randomly sampled from a continuous probability, this happens with probability 0. For example, when X, Y are independent, then all parameters are 0 in both models, and both models will compress the data equally well. Importantly, even when the true model is in neither of our model classes, we can still expect reasonable inferences relative to these model classes; by the minimax property of refined codes we use, we encode every model as efficiently as possible, which suggests reliable performance and confidence scores even in adversarial cases. In the next section, we will see this intuition borne out by various experiments.

2.5 The Origin of Specious Causality: Related Work

Causal inference is arguably one of the most important problems in both statistical inference and also all of the sciences. It hence has attracted a lot of research

attention (Rubin, 1974; Spirtes et al., 2000; Pearl, 2009). Unfortunately, the existence of confounders, selection bias, and other statistical problems make it impossible to infer causality from observational data without making any further assumptions (Pearl, 2009). When their assumptions hold, a large variety of both constraint-based (Spirtes et al., 2000, 1995; Zhang, 2008) and score-based methods (Chickering, 2002a; Scanagatta et al., 2015; Ramsey et al., 2017; Raskutti and Uhler, 2018; Solus et al., 2021; Rashid et al., 2022) for causal discovery can reconstruct causal graphs up to Markov equivalence. However, this means that they are not applicable to determining the causal direction between two variables X and Y, nor when causal sufficiency does not hold.

By making assumptions on the shape of the causal process, Additive Noise Models (ANMs) can determine the causal direction between just X and Y. In particular, ANMs assume independence between the cause and the residual (noise) and infer causation if such a model can be found in one direction but not in the other (Shimizu et al., 2006, 2011; Hoyer et al., 2008a; Zhang and Hyvärinen, 2009). The idea is simple: when $Y = f(X) + \varepsilon$ with $\varepsilon \perp\!\!\!\perp X$, then in general there exists no function g such that $X = g(Y) + \eta$ where $\eta \perp\!\!\!\perp Y$.

A more general framework for inferring causation than the above is given by the Algorithmic Markov Condition (Lemeire and Janzing, 2013; Janzing and Schölkopf, 2010), which we introduced earlier. In this framework, the simplest network over the observed variables—measured in terms of Kolmogorov complexity—is the true causal network from which the data was generated. Thus, in the bivariate case, if $X \to Y$, then the factorizations of the distribution $P(X, Y)$ should satisfy $K(P(X)) + K(P(Y \mid X)) \overset{+}{\leq} K(P(Y)) + K(P(X \mid Y))$. Since K is not computable (Li and Vitányi, 2009), practical instantiations use computable criteria to judge the complexity of the two causal directions, including Rényi-entropies (Kocaoglu et al., 2017), information geometry (Daniusis et al., 2012; Janzing et al., 2012, 2015), coding theory (Figueiredo and Oliveira, 2023), Bayes Factors (Dhir and van der Wilk, 2023), and MDL (Budhathoki and Vreeken, 2017; Marx and Vreeken, 2017, 2019).

When we extend our work from the case of causal sufficiency to that of insufficiency, it is important to note that without further assumptions, many different latent factorizations of the same data are possible, and thus latent confounders are generally not unique without additional assumptions (D'Amour, 2019).

Most similar to our approach is the work by Janzing and Schölkopf on determining the "structural strength of confounding" in a high-dimensional linear regression model for a continuous-valued pair X, Y, which they propose to measure using spectral analysis (Janzing and Schölkopf, 2018), respectively ICA (Janzing and Schölkopf, 2018). Like us, they also focus on linear relationships, but in contrast to us, they define a one-sided significance score rather than a two-sided information theoretic confidence score. More theoretical analysis along these lines shows that under some additional assumptions, L^1-optimal

convergence towards the true causal model can be attained (Ćevid et al., 2020; Rendsburg et al., 2022). Rather than implicitly inferring the existence of latent confounders by measuring the significance of such deviations, we instead explicitly model the hidden confounder Z via probabilistic PCA (Tipping and Bishop, 1999). While this makes our approach linear in nature, too, this approach permits us to fairly compare the scores for the models $X \to Y$ and $X \leftarrow Z \to Y$, allowing us to define a reliable confidence measure. Janzing and Schölkopf (2018) considered the case of determining whether the variables X are confounded by finding deviations of the regression vector from theoretical properties in high-dimensional regression. Kaltenpoth and Vreeken (2019) use the AMC (Janzing and Schölkopf, 2010) to infer whether two sets of variables X and Y are causally related or jointly confounded. In contrast, Wang and Blei (2019) and Ranganath and Perotte (2018) explicitly model latent confounders using factor models to adjust causal estimates for their presence. An information-theoretic similar to our own has also been proposed for finding confounders has also been applied to the discrete case (Kocaoglu et al., 2018) by minimizing $I_\alpha(X, Y|Z) + \beta H_\alpha(Z)$ over a proposed confounder Z, where H_α, I_α are the Rényi entropy and mutual information.

2.6 "Show, Don't Tell": Experiments

We now empirically evaluate CoCa in a variety of experiments. In particular, we consider performance in telling causally generated data from confounded data, both for settings where our model assumptions apply and those where they do not, and using both synthetic and real-world data. We compare against two other methods by Janzing and Schölkopf designed for the same task, which are based on finding deviations from the expected properties of regression vectors in high-dimensional regression problems (Janzing and Schölkopf, 2018; Janzing and Schölkopf, 2018). We implemented CoCa in Python using PyMC3 (Salvatier et al., 2016) for posterior inference with ADVI (Kucukelbir et al., 2017). All code is available for research purposes on our website.[4]

2.6.1 Experiments on Synthetic Data

To see whether CoCa works at all, we start by testing it on synthetic data with known ground truth close to our assumptions. For the confounded case, we generate samples over X, Y from the model

$$\begin{aligned} Z_j &\sim p_z, & W_{ij} &\sim p_w \\ \varepsilon &\sim N(0,1) & X, Y &= W^\top Z + \varepsilon\,, \end{aligned}$$

[4] `https://eda.rg.cispa.io/prj/coca/`

while for the causal case, we generate X, Y as

$$\begin{aligned} X_i &\sim p_x & w_i &\sim p_w \\ \varepsilon &\sim N(0,1) & Y &= w^\top X + \varepsilon\,. \end{aligned}$$

While these models look precisely like the Bayesian regression of Equation (2.6) and the PPCA model of Equation (2.7), we do not assume the distributions of Z, W, or X, w to be Gaussian. To see how much CoCa depends on the precise assumptions of the used models, we consider the source distributions,

$$\begin{aligned} p_z, p_x, p_w \in \{&N(0,1), \text{Laplace}(0,1), \\ &\text{LogNormal}(0,1), \text{Uniform}(0,1)\}\,. \end{aligned} \tag{2.10}$$

We expect CoCa to perform best when all distributions are Gaussian, as this corresponds to our model assumptions made in Equation (2.6) and Equation (2.7). From Theorem 2.3, we further expect CoCa to perform well when $\dim(X) \gg \dim(Z)$, but not so well when this assumption is violated. We therefore begin by generating data with fixed dimensionality $\dim(Z) = 3$, and vary the dimensionality of X, $\dim(X) \in \{1, 3, 6, 9\}$. Afterward, we also study the relationship between $\dim(X), \dim(Z)$ and the performance of CoCa.

To study not only how well CoCa performs in aggregate across all datasets but also how it performs for different levels of its confidence $\mathcal{C}$, we study its performance by looking at its *decision rate plots* as we describe next.

Evaluating Performance with Decision Rate Plots

In Section 2.4.3, we argued based on the no hypercompression inequality that our decisions should be more accurate when the confidence $\mathcal{C}$ of Equation (2.9) is large in absolute terms than when it is small: a large positive value $\mathcal{C} \approx 1$ indicates that the causal model outperforms the confounded model, whereas a large negative value $\mathcal{C} \approx -1$ indicates the reverse situation and neutral confidence $\mathcal{C} \approx 0$ indicates that both models are roughly equally good and no decision should be made. While it is tempting to try to define which values should count as large or small, we take the more empirical approach of ranking the various decisions and checking whether the expected pattern obtains.

To this end, we use decision rate (DR) plots, in which we consider the accuracy of CoCa over different datasets sorted by descending absolute confidence, $|\mathcal{C}|$. That is, for each dataset, we evaluate the confidence $\mathcal{C}$ associated with the classification decision and sort datasets in decreasing order of $|\mathcal{C}|$. On the x axis, we show the percentile ranking of $|\mathcal{C}|$, and on the y axis, our metric of choice, which here is accuracy. A point at coordinates $(0.2, 0.8)$ tells us that for the 20% of data where our method is most confident, it correctly decides

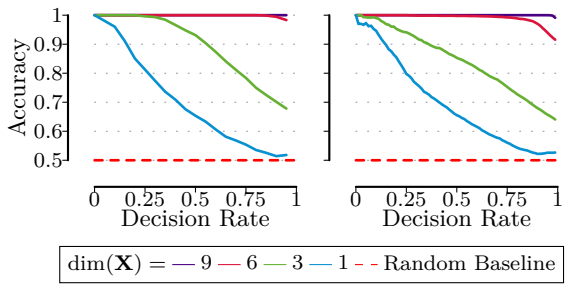


Figure 2.6: [Higher is better] *Accuracy over top-k% pairs sorted by confidence.* We use different generating models of X, Y, and different $\dim(X)$, with $\dim(Z) = 3$ fixed. Left: $p_x, p_z, p_w = N(0, 1)$. Right: All distributions are sampled uniformly from the set of distributions in Equation (2.10)

between the causal and confounded cases 80% of the time. These plots are commonly used in the causal inference literature as they give more information about performance than aggregate accuracy scores, which correspond only to the right-most point in the figure. In particular, if the confidence of a method is to be a useful quantity, the curves plotted should be monotonically decreasing, and the area under the decision rate curve should be as high as possible.

We show our DR plots in Figure 2.6. On the left, we show the case where all p_z, p_x, p_w are Gaussian distributions, and on the right, the case where each of the distributions is uniformly sampled from the distributions in Equation (2.10). Clearly, while CoCa works better when all model assumptions are met, it still performs remarkably well when they are not, and the qualitative picture remains the same. Overall, for all dimensionalities $\dim(X)$, the accuracy of CoCa decreases monotonically in confidence. As we expected, for $\dim(X) \leq \dim(Z)$, there is not enough information in X, Y about the confounder Z to decide between causal and confounded models, so we see relatively steep drops in these cases. Nevertheless, CoCa still determines the correct model for those cases where it is most confident. In contrast, when $\dim(X) > \dim(Z)$, there is enough information to decide between causal and confounded models, and CoCa is both highly confident and accurate.

Importantly, almost all results are significantly better than the 95% confidence interval of a fair coin flip—the exception is $\dim(X) = 1$, which is significant for the 75% where it was most confident. Further, for no combination of distributions and dimensions of X and Z was CoCa biased towards either class.

Performance for Different Dimensions of X and Z

Since the performance of CoCa depends on the relationship between $\dim(X)$ and $\dim(Z)$, we next study how it fares for a larger range of combinations of dimensionalities of X and Z. In Figure 2.7, we plot a heatmap of the area under the decision rate curve (AUDR) of CoCa, which we use to measure its aggregate performance across all levels of confidence. As expected, when

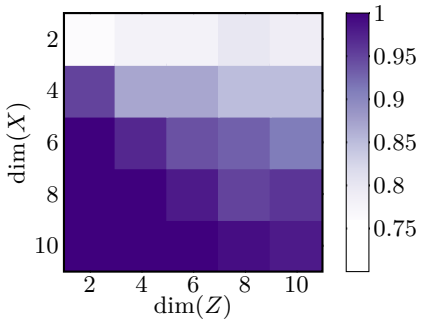


Figure 2.7: [Darker is better] *Area under the Decision Rate Curve for different dimensions of X and Z.* For fixed $\dim(Z)$ performance improves as $\dim(X)$ increases, while for fixed $\dim(X)$ performance degrades as $\dim(Z)$ increases. CoCa scores above 0.75, against a baseline of 0.5.

$\dim(Z)$ is fixed, we become more accurate as $\dim(X)$ increases. Further, as $\dim(Z)$ increases for fixed $\dim(X)$, our performance degrades gracefully—this is because we infer a $\hat{Z}$ of dimensionality one, which deviates further from the true generating process as the dimensionality of the true Z increases. Notably, even in the worst case of $\dim(X) = 2$ and $\dim(Z)$, the AUDR score for CoCa is above 0.7, whereas a random classifier would obtain a score of 0.5.

Comparison with other methods

As the last study on our synthetic data, we compare CoCa to two approaches by Janzing and Schölkopf, both of which are based on the properties of high-dimensional regression vectors (Janzing and Schölkopf, 2010, 2018). In particular, one of them is based on spectral analysis (SA; Janzing and Schölkopf, 2018) and the other on independent component analysis (ICA; Janzing and Schölkopf, 2018). Since both methods require X to be multi-dimensional, we consider the cases $\dim(X) = 3, 6, 9$, and sample p_x, p_y, p_z as before from Equation (2.10). We show the results in Figure 2.8. As both SA and ICA-based methods provide only an estimate $\widehat{\beta} \in [0, 1]$ of the strength of confounding, without any confidence score to distinguish between confounded and causal cases, we used $|\widehat{\beta} - 1/2|$ as substitute. Since the accuracy decreases with less extreme values of $\widehat{\beta}$, this seems to be a reasonable measure.
For both $\dim(X) = 3$ and 6, CoCa outperforms these competitors by a large margin where the respective methods are most confident, but also that the overall accuracies are almost indistinguishable. For $\dim(X) = 9$, the dimensionality of X is large relative to Z so that all approaches obtain close to perfect performances, and consequently, the differences in performance reduce.

2.6.2 Simulated Genetic Networks

Next, we consider more realistic synthetic data. For this, we consider the DREAM 3 data (Prill et al., 2010), originally used to compare different meth-

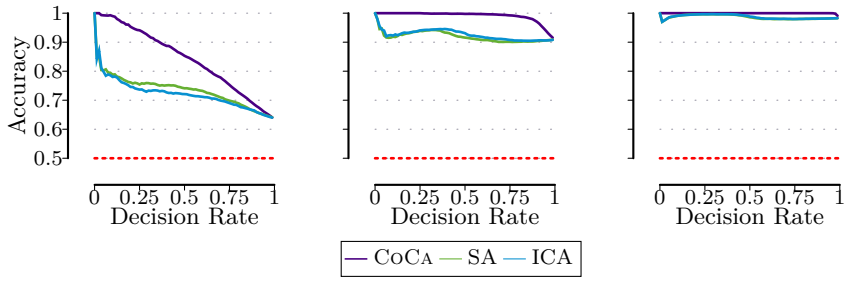


Figure 2.8: [Higher is better.] *Comparison of* CoCa *against the spectral analysis-based (SA; Janzing and Schölkopf, 2018) and ICA-based (ICA; Janzing and Schölkopf, 2018) approaches by Janzing and Schölkopf on synthetic data.* From left to right to right, we use $\dim(X) = 3$, 6, and 9. Baseline accuracy is at 0.5. We see that in all cases, CoCa performs best by a margin, particularly in regions where it is most confident.

ods for inferring biological networks. We use this data both because the underlying generative network is known and because the generative dynamics are biologically plausible (Prill et al., 2010). In particular, the relationships are highly nonlinear and, therefore, pose an interesting challenge to evaluate how CoCa performs when our assumptions do not hold at all. Out of all networks in the dataset, we consider the ten largest networks, those of 50 and 100 nodes, which are associated with time series of lengths 500 and 1000, respectively. Since CoCa was not designed to work with time series data, we treat the data as if it were generated from an i.i.d. source.

For each network, we take pairs (X, Y) of univariate X and Y such that precisely one of the following two cases applies

- X has a causal effect on Y and there exists no common parent Z, or
- X, Y have a common parent Z and there is no causal effect between them

Although we could, in principle, also consider tuples $(X_1, ..., X_m, Y)$ with $m > 1$, there were too few such tuples to provide meaningful comparisons. Further, since the original networks are heavily biased towards causality rather than to common parents, we take all the confounded pairs and then uniformly sample

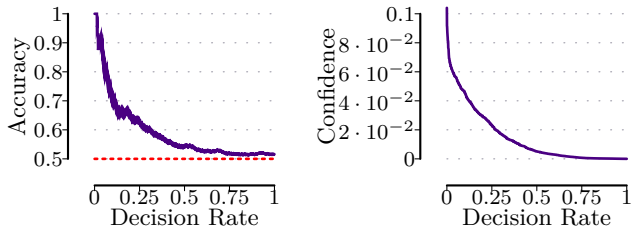


Figure 2.9: *Decision rate and corresponding confidence plots for the genetic networks data.* CoCa is accurate when it is confident, even for this adversarial setting.

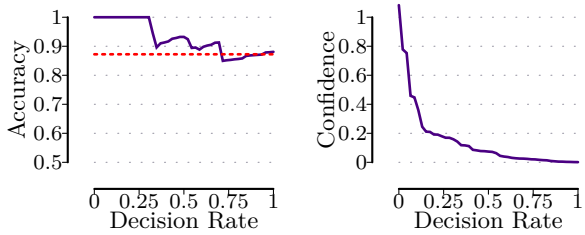


Figure 2.10: *Decision rate plot and its corresponding confidence plot for the Tübingen pairs.* The baseline for the decision rate plot is at 0.87. Note the strong correspondence between high confidence and high accuracy.

an equal number of causal pairs from the set of all such pairs.
We show the decision rate plot when applying CoCa to these pairs after aggregating over all the networks in the left-hand side plot of Figure 2.9. As before, we see that CoCa is highly accurate for those pairs where it is most confident. In comparison to the results for $\dim(X) = 1$ in Figure 2.6, we see that performance drops more quickly, which is readily explained by the fact that the simulated dynamics are both highly nonlinear and form a non-i.i.d. sample. Note, however, that our results are nevertheless still statistically significant with regard to a fair coin flip for the 75% pairs that CoCa is most confident about. To further explain the behavior of CoCa on this dataset, we plot the absolute confidence scores we obtain on the right of Figure 2.9. In particular, we see that for the first approximately 25% of decisions, the confidence we obtain is much larger than for the remaining pairs. This corresponds very nicely to the plot on the left, as the first 25% of our decisions are also those where we compare most favorably to the baseline.

2.6.3 Tübingen Benchmark Pairs

To consider real-world data suited for causal inference, we now consider the Tübingen benchmark pairs dataset (Mooij et al., 2016).This dataset consists of (mostly) pairs (X, Y) of univariate variables for which plausible directions of causality can be decided, assuming no hidden confounders. For many of these, however, it is either known or plausible to posit that they are confounded rather than directly causally related. For example, for pairs 65–67, certain stock returns are supposedly causal, but given the nature of the market, they would be better explained by common influences on the returns of the stocks. We, therefore, code every pair in the benchmark dataset as either causal (if we think the directly causal part to be stronger), confounded (if we expect the common cause to be the main driver), or unclear (if we are not sure which component is more important) and apply CoCa to the pairs in the first two categories. This leaves 47 pairs, of which we judged 41 to be mostly causal and 6 to be mostly confounded. We include the complete list in Appendix A.1.

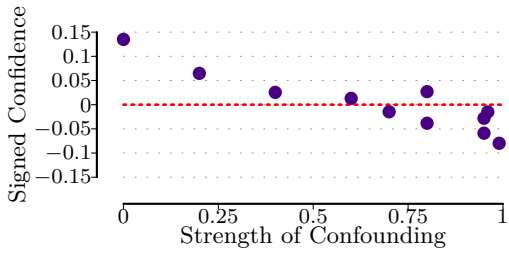


Figure 2.11: *Strength of confounding and confidence of* CoCa *on optical data.* As the data changes from no confounding to pure confounding, the confidence of CoCa tracks the true strength of confounding almost linearly.

In Figure 2.10, we show the decision rate plots over the datasets, weighed according to the benchmark definition. As in the previous cases, CoCa is most accurate where it is most confident while declining to the baseline as we try to classify points about which CoCa is less and less confident. We note that for these cases CoCa was biased towards saying that datasets represented truly causal relationships, even when we judged them to be driven by confounding. This is in large part explained by the fact that X and Y are both univariate, and we have seen that the case of $\dim(X) \leq \dim(Z)$ is particularly problematic when it comes to distinguishing between causal and confounded data. As we see from the right plot in Figure 2.10, CoCa nevertheless does better than the naive baseline of "everything is causal" by assigning more confidence to those datasets, which, according to our judgment, were truly causal.

2.6.4 Optical Data

Finally, we consider real-world optical data (Janzing and Schölkopf, 2018). Here, X is a low-resolution (3×3 pixels) image on a screen, and Y is the brightness measured by a photodiode some distance from the screen. The confounders Z are one LED in front of the photodiode and one in front of the camera, controlled by shared random noise, where the brightness of these LEDs controls the strength of confounding. We define the relative strength of confounding

$$s := b_{\text{LED}}/(b_{\text{LED}} + b_{\text{screen}})\,,$$

where b_{LED} and b_{screen} are the brightness of the respective device.

We evaluate CoCa on each dataset and plot the results in Figure 2.11. The strength of confounding increases from left to right, and values larger (smaller) than zero indicate that CoCa judged the data to be causal (confounded). We see that towards an intermediate confounding strength of 0.5, our method is uncertain about its classification. In contrast, towards the extreme ends of pure causality or pure confounding, it is very confident and correct in being so.

2.7 Beyond Cause and Confound: Toward Causal Discovery

In this chapter, we tackled Problem 1: to what extent is it possible to determine whether a set of correlated variables is causally related or whether they are jointly confounded. That is, we wanted to distinguish whether the data over variables (X, Y) has been generated via a truly causal model or whether the apparent cause and effect are, in fact, confounded by unmeasured variables.
To answer this question, we began by introducing the relevant causal and information-theoretic terminology and, in particular, the algorithmic Model of causality, wherein the typical causal Markov condition (Y is independent of its non-descendants given its parents) is replaced by the algorithmic Markov condition (the causal mechanisms governing each Y are algorithmically independent). We extended this framework to allow for the explicit inclusion of unmeasured variables and showed in Theorem 2.2 that comparing models with and without latent variables can be done fairly. By using the connection between Kolmogorov complexity and MDL, we showed that this framework can be made practical for distinguishing between causal and confounded relationships. In particular, in Theorem 2.3 we showed that by using MDL with latent factor models, we can distinguish them in the linear Gaussian setting.
In our experiments, we showed that CoCa performs well in practice, in cases where the data generation is close to our model assumptions and also where the generating process is different. Importantly, we showed that the confidence—the (normalized) difference in scores between competing models—of our method tracks its accuracy, suggesting that the confidence, which is readily measurable, is a good proxy for its performance, which is not observable.
This raises the question "So what?" Knowing that our data are (likely) confounded, what can we do with this information? If our interest were specifically in causal effect estimation, a natural next step would be to use this knowledge to remove potential bias due to confounding, e.g., as Wang and Blei (2019) do. In addition, we could consider richer model classes, such as richer causal relationships or richer latent variable models. While exploring these avenues would be practically relevant, we do not believe they would prove fruitful for obtaining insights into the nature of the problem we are trying to solve.
Instead, we pursue richer models differently. In this chapter, we considered the case where either all X are independent and jointly cause Y or all variables X, Y are jointly confounded. This is, of course, a highly unrealistic assumption, and we therefore next consider mixed types of causal and confounded relationships. That is, under which conditions can we discover which *subsets* of the observed variables X, Y are jointly confounded, and which are connected by causal relationships? This is precisely Problem 2 of discovering a causal network over X, Y and their potential confounders Z, which we move to next.

Chapter 3

Causal Discovery with Hidden Confounders

"Causa Latet Vis Est Notissima – The cause is hidden, but the result is known."

OVID: METAMORPHOSES IV, 287

While being able to tell whether a set of covariates X and a response variable Y may be jointly confounded is a good start, in many disciplines, we are interested in much more complex networks of causal relationships. In fact, in many biological (Ramb et al., 2013; Parsana et al., 2019), medical (Compton et al., 2023), social (Wunsch, 2007), economic (Angrist and Pischke, 2009), and machine learning systems (Tennenholtz et al., 2021), the observed covariates themselves are causally related, and multiple latent confounders lead to biased causal estimates among different subsets of these variables.

It is, therefore, important not only to discover and correct for latent confounding in a single set of causal estimates $X \to Y$ but to discover the entire causal DAG G governing the causal relationships both between the observed variables and also those between the latent confounders Z and the observed X. Naturally, this is not possible without making a number of assumptions. In this chapter, we will, therefore, stay close to the linear and Gaussian setting introduced in Chapter 2. Interestingly, while this setting is the starting point for a great many statistical methodologies due to its theoretical simplicity (Wasserman, 2004), it turns out to be one of the most challenging cases for causal

discovery (van de Geer and Bühlmann, 2013; Peters and Bühlmann, 2014), and this is only exacerbated by the inclusion of latent confounders.

Intuitively, the issue is as follows. Let X, Y be two jointly Gaussian variables. Then, all of the following three models,

$$\begin{aligned} M_1 : X = \varepsilon_X, \qquad & Y \;\; = \alpha X + \varepsilon_y \\ M_2 : Y = \varepsilon_Y, \qquad & X \;\; = \beta Y + \varepsilon_x \\ M_3 : Z = \varepsilon_Z, \quad & (X, Y) = (\gamma_1, \gamma_2) Z + (\varepsilon_x, \varepsilon_y)\,, \end{aligned}$$

with $\alpha, \beta, \gamma_i \in \mathbb{R}$ can model any possible joint Gaussian distributions over X, Y. That is, these models are indistinguishable based on observational data alone.

One might hope that, as in the previous chapter, the problem disappears as more variables X are observed. Unfortunately, this is not the case. While additional observed variables are indeed *necessary* to distinguish between the different models, they are not, by themselves, *sufficient*. A key ingredient will be the independence of causal mechanisms (Janzing and Schölkopf, 2010). Specifically, we will show that when any subset of at least four variables $X_{\mathcal{I}}$ are jointly confounded by a low-dimensional Z, any proposed causal factorization of X would be parametrized by a set of parameters lying on a low-dimensional manifold, violating this independence. To discover these non-independences of the causal mechanisms, we require additional structural assumptions about the causal DAG so that different factors $Z_i \in Z$ have enough of a signal among observed variables X but do not interfere too much with each other. This ensures that the low-dimensional manifolds for different underlying causal models also have trivial intersections, which is required for identifiability.

To explain our proposed framework and solutions, the chapter will be structured as follows. We start by formalizing the problem (Section 3.1.1) and explaining in more detail why our problem is not solvable in the general case (Section 3.1.2). We then take note of a useful structural property of causal networks discovered when causal sufficiency is violated (Section 3.2.1). Based on this property, we introduce additional assumptions under which we can guarantee the recovery of the causal network *including* the correct number and effects of the latent confounders Z (Section 3.2.2). We then show that under the same assumptions, the graph is not only identifiable but also learnable from data using a score-based method (Section 3.2.3) and provide a practical algorithm for doing so based on the idea of iteratively fitting causal models and extracting sets of confounded nodes from these proposed causal models (Section 3.3). Last, we observe that our approach not only has nice theoretical properties but that it also works well in practice (Section 3.5). As in the previous chapter, we include proofs for all theoretical statements in Appendix A.3.

3.1 What Is To Be Done? And How Not To Do It

We formally introduce the problem we are interested in solving and then study to what extent the commonly used approach of turning the causal discovery problem into a problem of independent component analysis (ICA) can solve it. In particular, we will show why the ICA-based approach can not be used due to non-identifiability issues of the underlying causal graph.

3.1.1 Problem Setting

We assume that the observed variables X and unobserved Z follow a distribution $P(X, Z)$ that factorizes according to the causal DAG $G^* = G^*_{X,Z}$ with nodes $V = (X, Z)$ and edges E^*. Besides the standard conditions for causal discovery, i.e., causal Markov, faithfulness, and sufficiency (over V), we assume that all Z_j are jointly independent and that no reverse causation exists, i.e., $\mathrm{Pa}^*(Z_j) = \emptyset$ for all j. The joint distribution $P(X, Z)$ can then be written

$$P(X, Z) = \prod_{i=1}^{m} P(X_i \mid \mathrm{Pa}_i^*) \prod_{j=1}^{l} P(Z_j)\,,$$

where $\mathrm{Pa}_i^* = \mathrm{Pa}_G^*(X_i)$ are the parents of X_i in G^*. We aim to recover the true graph G^* over both X and Z given information only about the observed variables X. More specifically, we want to solve the following problem.

Problem Statement (Informal)**.** Given a sample x from $P(X)$, discover

- a (small) set of latent variables Z
- a (sparse) network $G_{X,Z}$ over X and Z
- and a (simple) joint distribution $P(X, Z)$ such that

$$P(X, Z) = \prod_{i=1}^{m} P(X_i \mid \mathrm{Pa}_i) \prod_{j=1}^{l} P(Z_j)\,,$$

factorizes according to the discovered G, with $\mathrm{Pa}_i = \mathrm{Pa}_G(X_i)$.

We now explain how discovering the graph *can* be related to the well-known framework of independent component analysis and why this does not solve the problem we are interested in.

3.1.2 ICA Cannot Identify Latent Confounding

For simplicity, let the observed data X be generated by the linear SCM,

$$X = AX + BZ + \varepsilon\,, \tag{3.1}$$

where we only assume that $\mathrm{var}(\varepsilon_i) \neq 0, \mathrm{var}(Z_j) \neq 0$ and independence of Z and ε, but leave the distributions of ε and Z otherwise unspecified. In essence, the inclusion of BZ amounts to an introduction of correlated source of noise affecting the observed variables X. Since A, B describe the acyclic graph G^*, there are no paths of length $\geq m$ among the observed X, so that $A^m = 0$. All eigenvalues of A are therefore 0, so that $(I - A)$ is invertible, and thus we can rewrite the model by writing X as a linear mixing of some source S,

$$\begin{aligned} X &= (I - A)^{-1} (BZ + \varepsilon) \\ &= (CB \quad C) \begin{pmatrix} Z \\ \varepsilon \end{pmatrix} \\ &= QS\,, \end{aligned}$$

where we abbreviate $C = (I - A)^{-1}$, and $Q = (CB \quad C)$ is the mixing matrix and $S = (Z, \varepsilon)^\top$ is the set of all sources generating X.

The problem of recovering the mixing matrix Q of this generating process is referred to as *independent component analysis* (ICA; Comon, 1994). When $\dim(S) \geq \dim(X)$, or equivalently $\dim(Z) > 0$, it is more specifically referred to as *overcomplete* ICA (OICA, Eriksson and Koivunen, 2004).

When the matrix Q has full column rank, no two columns of Q are collinear, and the sources S are all (except for at most one) *non*-Gaussian, it has been shown that the mixing matrix Q is identifiable up to permutation and scaling of its columns (Eriksson and Koivunen, 2004),

$$Q' = QP\Lambda\,.$$

These non-identifiabilities are unavoidable. That is, since there is no unique order for the sources S_i, we cannot distinguish between the models "S_1 affects X_1, S_2 affects X_2" and the same statement with S_1 and S_2 switched. Furthermore, without additional assumptions on $P(S)$ we cannot distinguish between "$X_1 = 5 \cdot 1 S_1$ " and "$X_1 = 1 \cdot 5 S_1$". Without further assumptions, identifiability up to permutations and rescaling is, therefore, the best one can do.

The conditions for identifiability are also natural from the point of view of the generating causal model of Equation (3.1) that we started from. Clearly, if $\mathrm{rank}(Q) < m$, then QS could only generate degenerate distribution $P(X)$ with deterministic collinearities, which could only be consistent with Equation (3.1) if one of the ε_i satisfies $\mathrm{var}(\varepsilon_i) = 0$, in contradiction to our assumption.

The condition that no two columns of Q are collinear also has a natural interpretation in our causal model. First, it entails that the exogenous noise variables ε_i *directly* affect *different* variables X_i. Second, that the confounding factors Z_j each affect *more than one* X_i, and further when Z_j, Z_k have exactly the same children, their relative weights on the children are *not exactly the*

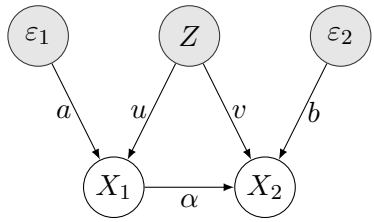


Figure 3.1: *Nonidentifiability of causal effects in OICA.* Even in this simple causal model satisfying the assumptions of OICA identifiability, it is impossible to distinguish between correlations due to the causal path $X_1 \to X_2$ and the impact of the latent confounder Z on the observed variables.

same. That is, for any two latent confounders Z_j, Z_k there exists at least one pair i, i' of observed variables $X_i, X_{i'}$ such that $b_{ij}/b_{i'j} \neq b_{ik}/b_{i'k}$.
To justify the last assumption of non-Gaussianity, note that if the variables S follow a joint independent Gaussian distribution with equal variances,[1] then Q is further non-identifiable due to orthogonal transformations. That is,

$$QS = QU^\top US\,,$$

for any orthogonal matrix $U \in O(n)$. Since $US \sim S$ have the same distribution and are therefore both jointly independent, Q becomes unidentifiable.
Even when all assumptions hold, and Q is identifiable, this does *not* mean that B and C themselves are identifiable. To see this, let us consider an example.

Example 3.1. Consider the following generating model for X (see Figure 3.1),

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} u & a & 0 \\ v & c & b \end{pmatrix} \begin{pmatrix} Z \\ \varepsilon_1 \\ \varepsilon_2 \end{pmatrix},$$

corresponding to causal model shown in Figure 3.1, with $c = \alpha a$. Note that in this case, we can decompose the matrix as

$$C = \begin{pmatrix} a & 0 \\ c & b \end{pmatrix}, \quad B = \begin{pmatrix} u/a \\ (v - cu/a)/b \end{pmatrix}.$$

However, by permuting the variables Z and ε_1, we can also write

$$C' = \begin{pmatrix} u & a \\ v & 0 \end{pmatrix}, \quad B' = \begin{pmatrix} a/u \\ (c - va/u)/b \end{pmatrix},$$

which result in precisely the same observed mixing matrix Q (up to permutations) but different causal effects $X_1 \to X_2$ as well as from Z towards X_1, X_2. In essence, we have here re-derived algebraically that the correlations due to an unmodeled confounding factor Z and due to an edge $X_1 \to X_2$ are not

[1] This is w.l.o.g. since identifiability is only up to rescaling in the first place.

distinguishable from each other in this generating model.

While in this example, we cannot distinguish between the two different decompositions, we will see next that confounding induces specific structures in the marginal graph G_X^* over the observed variables X, which will help us determine the correct decomposition of Q into its causal and confounded parts.

3.2 The Topology and Geometry of Latent Confounding

We now formalize the specific requirements properties and requirements that permit us to discover latent confounding in observational data. We begin by showing that latent confounding leads to cliques in the marginal graph before exploiting the non-independence of implied causal mechanisms to recover the true causal network over both X and Z.

3.2.1 Graphical Structure of Latent Confounding

The key to discovering which variables are jointly confounded lies in the following observation: Whenever a set of variables $X_{\mathcal{I}} = (X_i)_{i\in\mathcal{I}}$, $\mathcal{I} \subseteq [m]$, are confounded by an unmeasured univariate Z, no pair of variables $X_i, X_j \in X_{\mathcal{I}}$ can be d-separated by conditioning on any subset $W \subseteq X_{-\{i,j\}}$, where $X_{-\{i,j\}}$ denotes the set of all variables in X except X_i, X_j (Elidan et al., 2000). To d-separate them, we would have to condition on the unobserved Z itself. Thus, any proposed DAG G_X capturing the conditional independences of $P(X)$ will necessarily mutually connect all pairs of variables in $X_{\mathcal{I}}$. That is, G_X contains a *clique*, a fully connected subgraph, over the variables $X_{\mathcal{I}}$.

Proposition 3.1 (Confounders and Cliques)**.** *Let $P(X, Z)$ be the joint distribution of X, Z where Z is one-dimensional and let $\mathcal{I} = \{i : Z \to X_i\}$ be the set of indices of variables co-caused by Z. Then, any graph G_X capturing the correlations in $P(X)$ contains a clique over $X_{\mathcal{I}}$.*

Proof sketch. When X_i, X_j are jointly confounded by an unobserved Z, then $X_i \perp\!\!\!\perp X_j \mid W$ can happen only if W contains all the information in Z. But this is not possible for any set of observed variables $W \subseteq X_{-\{i,j\}}$. □

When Z is multivariate, each Z_j induces its own clique in G_X, and these cliques may share nodes. We show an example of this in Figure 3.2, where the edges in G_X are left undirected to indicate that *any* (acyclic) ordering would work. Note that if Z is not independent, e.g., $Z_1 \to Z_2$, then this would be graphically indistinguishable from Z_1 being a direct cause of all children of both Z_1 and Z_2. Next, we show that this graphical characterization of confounding is already sufficient to identify the true model in a restricted, sparse, linear setting.

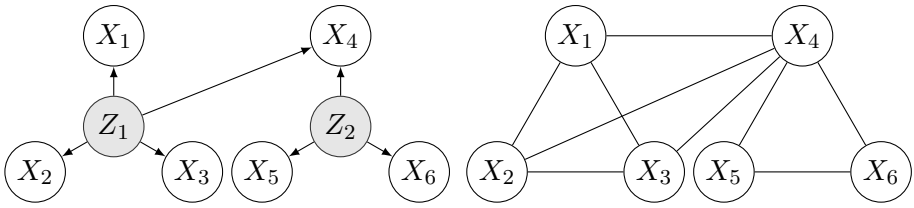


Figure 3.2: *Effects of confounders on marginal graphs.* When the confounders Z_1, Z_2 are not observed, any causal graph capturing the independences among the observed variables X will necessarily form cliques among the jointly confounded nodes (here $\{X_1, \ldots, X_4\}$ and $\{X_4, \ldots, X_6\}$). The cliques overlap at X_4 since it is affected by both Z_1 and Z_2.

3.2.2 Identifiability for a Sparse Linear Confounded Model

Next, to leverage the above relationship between latent confounding and clique structure of causal DAGs over the observed variables, we assume that $P(X, Z)$ is given by the linear structural causal model of Equation (3.1),

$$X = AX + BZ + \varepsilon\,,$$

where A encodes the edges between variables X in the true DAG G^*, and B encodes the edges between Z and X. We further assume the distributions of the confounders, $Z \sim P(Z)$, and of the noise variables $\varepsilon \sim P(\varepsilon)$ to be symmetric $P(Z) = P(-Z)$ and $P(\varepsilon) = P(-\varepsilon)$ so that in particular both have mean 0.

Example 3.2. We begin with three examples to motivate our assumptions.

a) Consider the following simple model: $X = (X_1, \ldots, X_4)$ consists of four variables whose correlations are fully described by the univariate latent confounder $Z \sim N(0, 1)$,

$$X = bZ + \varepsilon\,.$$

Then the correlations are given by $\sigma_{ij} = \mathrm{cov}(X_i, X_j) = b_i b_j$. In particular, the six covariances can be fully described by a set of four parameters, and the relationship between the σ_{ij} is known as *tetrad constraint* (Silva et al., 2006). We will see that, in this case, the observational data gives us sufficient information about the underlying model to determine the parameters b up to trivial invariances similar to ICA solutions.

b) Now consider the more complex model generating $X = (X_1, \ldots, X_4)$, whose correlations are described by both the univariate latent confounder Z, but also an adjacency matrix A describing a complete graph over X,

$$X = AX + bZ + \varepsilon\,.$$

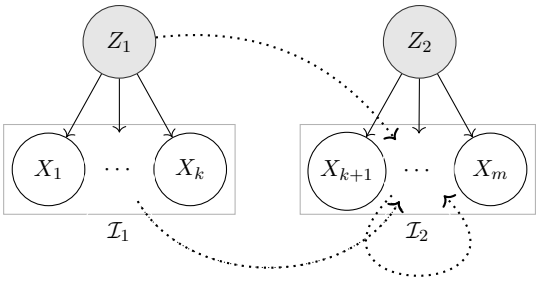


Figure 3.3: *Structural Assumptions A and B of our model.* Each Z_i has an edge towards each variable in $\mathcal{I}_i$ (solid), but there are only few edges incoming to each $\mathcal{I}_i$ from other sources (dotted).

In this case, the six correlations would be parametrized by ten parameters, and any hope of identifying the generating parameters is lost. This is a different view on Example 3.1 in terms of the number of parameters.

c) Now consider the same model as in a), but with only three variables $X = (X_1, \ldots, X_3)$. Then the three correlations σ_{ij} can be described *either* by three parameters describing direct causal effects between the X_i, *or* by three parameters b_i describing pure confounding. That is, we require at least four variables to be able to distinguish between the cases.

Using these examples to guide our intuition about which models are identifiable and which are not, we now introduce the following assumptions.

Assumption A (Sufficient Signal of Confounders)**.** The variables X are split into disjoint sets $\mathcal{I}_1, \ldots, \mathcal{I}_l$ of size $|\mathcal{I}_j| \geq 4$, such that Z_j has non-zero influence on each $X_i \in \mathcal{I}_j$, i.e., $b_{ij} \neq 0$ for all $X_i \in \mathcal{I}_j$.

This assumption guarantees that each confounder has enough children to allow for recovering its effects. In particular, in this framework, we cannot distinguish between causality and confounding for only two or three observed variables.

Assumption B (Sparsity)**.** For each $\mathcal{I}_j$, there are at most $|\mathcal{I}_j| - 4$ edges incoming to vertices in $\mathcal{I}_j$, aside from those starting in Z_j.

By adding this assumption, we ensure that not only does each Z_j influence sufficiently many of the observed variables but also that there are few other influences on these variables. In the second example above, the parameters b cannot be recovered because it is impossible to distinguish between correlations due to the confounder Z and those due to causal relationships between the observed variables. By restricting how many incoming edges each set $\mathcal{I}_j$ has, we ensure that the parameter matrix B can be recovered. This restriction is displayed by dotted edges in Figure 3.3. That is, $\mathcal{I}_2$ must have fewer incoming edges from *all* other sources combined than it has incoming edges from Z_2.
We depict the structural Assumptions A and B we make in Figure 3.3 and show some example graphs satisfying them in Figure 3.4.

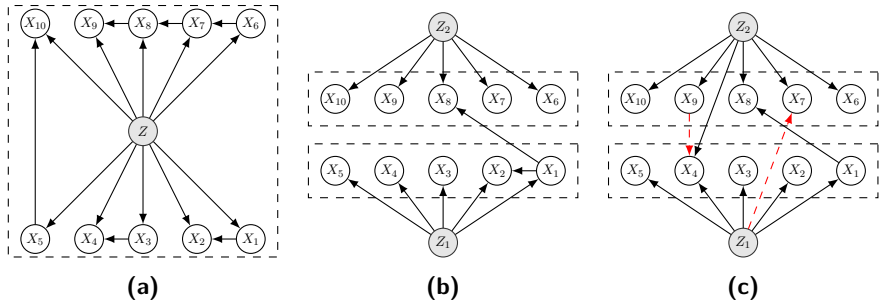


Figure 3.4: *Example graphs illustrating our structural assumptions.* (a) All circle, draw variables are confounded by the same factor Z_1, and 6 edges exist between its children $\mathcal{I}_1$. (b) Two different confounders affecting five nodes each, and one additional edge incoming to each of the sets. (c) Z_2 affects one of the nodes in $\mathcal{I}_1$. Furthermore, if we added either of the dashed red edges, too many edges incoming into $\mathcal{I}_1$, respectively, $\mathcal{I}_2$ would violate our identifiability assumptions.

Assumption C (No False Positives)**.** For all distinct X_i, X_j, X_u, X_v not independent given Z, with covariances σ_{rs}, we have $\sigma_{ij}\sigma_{uv} \neq \sigma_{iu}\sigma_{jv}$.

This assumption ensures that we cannot find a set of causal effects that lead to matching *precisely* the correlations entailed by a confounded model. That is, while we could pick parameters A in a causal model without confounders so that the correlation matrix Σ would match precisely the correlation matrix created by Z and b in the first example above, this matrix A would be *specifically* fine-tuned to satisfy all the constraints and would require the causal mechanisms to not be picked independently. In a sense, this assumption is very similar to the assumption of causal faithfulness in that we assume parameters are not picked precisely so as to "cancel out" and look like perfectly confounded variables.
We call a linear model satisfying these assumptions a *Sparse Linear Confounded* (SLC) model. When all assumptions hold, this causal model is identifiable.

Theorem 3.2 (Identifiability of the SLC Model)**.** *Let Z be of dimension $l \leq m/4$, and let $P(X, Z)$ be described by the linear SCM of Equation* (3.1)

$$X = AX + BZ + \varepsilon\,.$$

Further, let Assumptions A–C hold. Then, both the number l of confounders and its parameters B are identifiable up to trivial indeterminacies (column permutations and rescaling). Furthermore, if either all noise variables ε_i are non-Gaussian or all ε_i have equal variances, then A is also identifiable.

Proof sketch. Due to the imposed structural sparsity of the network, for any target variables X_i, we can find a set of four distinct variables (X_i, X_u, X_v, X_w)

such that $b_i^2 = \sigma_{iu}\sigma_{iv}/\sigma_{uv}$. Furthermore, once we know all the values in B, we can recover $P(X \mid Z)$, for which identifiability follows from known results for non-Gaussian noise (Shimizu et al., 2006), or Gaussian noise with equal variances (Peters and Bühlmann, 2014). □

Note that we do not need to know the number l of confounders Z_j, nor the sets $\mathcal{I}_j$ to determine B. Instead, the correlation structure of the data along with the tetrad constraints of Example 3.2 fully determine both the number of confounders as well as the sets $\mathcal{I}_j$ up to permutations of its indices.
Note that these assumptions are not *necessary* ones but merely sufficient. In fact, we can show that for very large randomly generated DAGs, we do not require any sparsity to identify the causal network.

Theorem 3.3 (Identifiability for Large Dense Graphs)**.** *Let Assumptions A and C hold and let the true causal graph G^* over X, Z be sampled from a directed Erdős-Rényi model $ER(m + l, p)$, with m observed and l latent nodes and edge probability $p < 1$. Then in the limit of infinitely many variables, the matrices A and B are identifiable with probability* 1*,*

$$\lim_{m\to\infty} P(A, B \textit{ identifiable}) = 1\,,$$

where the limit is taken over DAGs with fixed topological order.

Proof sketch. When $p < 1$, for any Z_j with child X_i, for sufficiently large m we are guaranteed to find a suitable quadruple X_i, X_u, X_v, X_w to estimate b_{ij}. Given all b_{ij}, the entries of A corresponding to incoming edges into X_i are identifiable for the same reason as in Theorem 3.2. □

Here, by a directed Erdős-Rényi graph we mean that given a fixed topological order $X_{\pi^{-1}(1)} \preceq \cdots \preceq X_{\pi^{-1}(m)}$ an edge $(X_{\pi^{-1}(i)}, X_{\pi^{-1}(j)})$ exists with probability p if $i < j$, and with probability 0 if $i \geq j$. This is the directed analog of the standard Erdős-Rényi model where an edge between any two X_i, X_j exists with probability p, such that the edges are now directed in accordance with the topological order. Furthermore, when we say that the limit is taken over graphs with a fixed topological order, we mean that when a node X_{m+1} is added, the relative topological order among $X_1, \ldots, X_m$ is unaffected. That is, adding X_{m+1} does not lead to any earlier edges being flipped.
Given that the model is identifiable in theory, can we *learn* the correct network given enough data? The answer turns out to be "Yes".

3.2.3 Learning Confounders

In Section 2.3, we developed our theory of the extended algorithmic Markov condition to determine the existence of confounders. That is, when the observed

variables X are generated from a model involving latent confounders Z, then we would generally expect the full distribution $P(X, Z)$ to satisfy

$$K(P(X, Z)) < K(P(X)) .$$

Conversely, when the set X is causally sufficient, then by Theorem 2.2, no distribution $P(X, Z)$ satisfying this inequality exists.
While the distribution $P(X, Z)$ *should* satisfy this inequality, we do not know this distribution. Further, if we try to find *some* distribution $Q(X, Z)$ which satisfies the inequality, we cannot guarantee that the causal relationships between Z and X are correctly captured. In particular, we cannot guarantee that the sets $\mathcal{I}_j(Q) = \{X_i : Z_j \to X_i\}$ according to Q correctly recover the sets of confounded nodes according to P, given by the index set $\mathcal{I}_j^*$. That is, while such a distribution $Q(X, Z)$ with the correct sets $\mathcal{I}_j(Q) = \mathcal{I}_j^*$ should exist, for any given Q, it is not clear whether it, in fact, recovers the correct sets $\mathcal{I}_j^*$.
While we cannot guarantee that we find the correct model for arbitrary linear models, it fortunately turns out that for the SLC, the Bayesian information criterion (BIC; Schwarz, 1978) is consistent under the additional assumption that the latent confounders Z and noise variables ε are Gaussian.

Theorem 3.4 (Consistency of BIC for Gaussian SLCs)**.** *Let $x = x^n$ be a sample from the SLC of Equation* (3.1) *with Gaussian distributions $P(Z), P(\varepsilon)$ and let Assumptions A–C hold. Let $\mathcal{M}$ be the corresponding model class and $\mathcal{M}_0$ the subset of $\mathcal{M}$ with $B = 0$ fixed. Further, consider the score*

$$L(x^n, M) = -\log P(x^n \mid A, B, \sigma_\varepsilon^2) + \lambda \|A\|_0 + \lambda \|B\|_0 , \tag{3.2}$$

and denote its minimizers by $\widehat{A}, \widehat{B}$. Then, for $\lambda = \log(n)/2$, our score L is the BIC score and is consistent for detecting confounders. That is,

$$\lim_{n\to\infty} P\left(\min_{M\in\mathcal{M}} L(x^n, M) < \min_{M\in\mathcal{M}_0} L(x^n, M) \right) = 1 .$$

Furthermore, $\widehat{A}$ and $\widehat{B}$ converge to the true A, B with probability 1*,*

$$\lim_{n\to\infty} P(\widehat{A} = A, \widehat{B} = B) = 1 .$$

Proof sketch. Due to Assumptions A,B, we know that for any given X_i, there exists a quadruple of variables X_i, X_u, X_v, X_w for which using parameters b_{ij} of the matrix B is the way to parametrize their correlations with the *fewest* number of parameters. Since the penalty $\lambda \|A\|_0 + \lambda \|B\|_0$ precisely counts such parameters, we therefore have $\widehat{B} \to B$, and due to uniqueness of the global minimum (van de Geer and Bühlmann, 2013) also $\widehat{A} \to A$. □

While it is remarkable that in the SLC, the full causal model can be recovered from only a sample x from $P(X)$, we have to solve two more problems before we can put this into practice. First, we know neither B, nor which of the exponentially many subsets $X_{\mathcal{I}} \subseteq X$ are affected by any one Z_j, nor how many Z_j there are precisely in the first place. Second, even knowing B and Z, optimizing Equation (3.2) is still NP-hard (Peters and Bühlmann, 2014). We, therefore, next develop a heuristic as to which subsets $X_{\mathcal{I}}$ are likely affected by Z and show how standard causal discovery algorithms can be leveraged to find a causal network over both the observed X and the latent Z.

3.3 Exploiting Observed Structures: The CDHC Algorithm

With the theory we developed, we can now introduce CDHC, our method for Causal Discovery with Hidden Confounders.

3.3.1 CDHC in a Nutshell

The idea behind CDHC is quite simple. Our goal is to find a network $G_{X,Z}$ over both X and Z and a corresponding joint distribution $P(X, Z)$, which captures the correlations of the observed X. To discover such a graph, we need to determine which correlations are causal and which are due to confounding by Z. To determine whether a particular set of correlations between a subset $X_{\mathcal{I}}$ of the variables is due to one or the other, we further need to evaluate the different potential models, similar to our score from Chapter 2.
Our approach to dealing with these problems can be summarized as follows.

a) Run a causal discovery algorithm $\mathcal{A}$ over $X' = X$ and find approximate cliques in the discovered marginal graph $G_{X'}$ (see Proposition 3.1).
b) Use MDL to evaluate competing causal and confounded explanations for candidate confounded sets.
c) Include the best $\widehat{Z}$ into the variable set, $X' = X \cup \left\{\widehat{Z}\right\}$.
d) Repeat steps a-c) until no more $\widehat{Z}$ can be found.
e) ???
f) Profit. (See Theorem 3.5 and Proposition 3.6.)

We now describe our approach in more detail.

3.3.2 Finding Confounded Variables

To find a causal network over both the observed X and its confounders Z, the first step is to determine which subsets of X are likely to be confounded. In Proposition 3.1, we showed that any (consistent) structure learning algorithm $\mathcal{A}$ will, in the limit, discover a clique over the nodes $X_{\mathcal{I}}$ when $X_{\mathcal{I}}$ are jointly caused

by the same latent variables. With only limited amounts of data, we are unlikely to find exact cliques, but the nodes in $X_{\mathcal{I}}$ should nevertheless form a densely connected approximate clique in the discovered $\widehat{G}$. However, approximating cliques is still a computationally hard problem (Feige et al., 1991), so we use the following simple heuristic: If the $X_{\mathcal{I}}$ are densely connected in a discovered graph $\widehat{G}$, then the Markov blanket for different variables $X_i \in X_{\mathcal{I}}$ should all satisfy the approximate equality $\mathrm{MB}(X_i) \approx X_{\mathcal{I}}$. Therefore, we consider the Markov blanket of each node as the seed sets $\mathcal{I}$ from which to start searching for suitable candidates for sets of confounded nodes, over which we may infer latent confounders. We then iteratively update these sets $\mathcal{I}$ to find those variables that are best compressed by including latent variables, as we describe next.

3.3.3 Learning Latent Confounders

To evaluate a proposed set of confounded nodes and its associated graph G, we introduce a causal model including latent factors as follows. Based on our identifiability results from Theorem 3.2, given a graph G over (X, Z), we assume that the data is generated from a model in a class $\mathcal{M} = \mathcal{M}(G)$ similar to the PPCA model of Equation (2.7), but with added edges between the observed X

$$\begin{aligned} Z_i &\sim N(0,1)\,, & \varepsilon &\sim N(0, \sigma_\varepsilon^2) \\ A_{ij} &\sim N(0, \sigma_a^2)\,, & B_{ij} &\sim N(0, \sigma_b^2) \\ X &= AX + BZ + \varepsilon\,, \end{aligned} \tag{3.3}$$

where entries of A, B are nonzero only when their corresponding edges are in G. As in PPCA, we can marginalize out Z to obtain the reduced form

$$X \mid A, B \sim N(0, C^\top (BB^\top + \sigma_\varepsilon^2) C)\,,$$

where the matrix C is defined as $C = (I - A)^{-1}$. Since the A_{ij} are sampled from a continuous distribution, Assumption C holds almost surely, so that, unlike PPCA, we do not require that $A = 0$ to recover the model.
To evaluate the fit of our model class $\mathcal{M}$ to our data x, we use a similar score $L(x; \mathcal{M})$ as for the PPCA model in Equation (2.7) in the previous chapter,

$$L(x; \mathcal{M}) = -\log \int P(x \mid A, B) P(A, B) dA dB\,,$$

which we estimate using standard methods (Kucukelbir et al., 2017). This score is suitable in that it is consistent for causal discovery with latent confounders.

Algorithm 3.1: CDHC

input : data x sampled from $P(X)$, algorithm $\mathcal{A}$
output : graph G and distribution $P(X, Z)$

$G = (V, E) \leftarrow$ Graph inferred over x using $\mathcal{A}$;
do
  foreach i $\in \{1, \ldots, m\}$ **do**
    $X_{\mathcal{I}} \leftarrow$ Markov blanket of X_i in G;
    $G' \leftarrow (V \cup \{Z\}, E \cup \{Z \to X_{\mathcal{I}}\})$;
    // Forward phase
    do
      $j \leftarrow \arg\min_{j \notin \mathcal{I}} L(x, G' \cup \{Z \to X_j\})$;
      $(\mathcal{I}, G') \leftarrow (\mathcal{I} \cup \{j\}, G' \cup \{Z \to X_j\})$;
    while $L(x, G')$ decreases;
    // Backward phase
    do
      $j \leftarrow \arg\min_{j \in \mathcal{I}} L(x, G' \setminus \{Z \to X_j\})$;
      $(\mathcal{I}, G') \leftarrow (\mathcal{I} \setminus \{j\}, G' \setminus \{Z \to X_j\})$;
    while $L(x, G')$ decreases;
    $z \leftarrow$ sample from $P(Z \mid X)$;
    $G[i] \leftarrow$ Graph inferred over (x, z) using $\mathcal{A}$;
  // Use the model with the best confounder
  $G \leftarrow \arg\min_{G[i]} L((x, z), G[i])$;
while G changes;
return G and the $P(X, Z)$ associated with G

Theorem 3.5 (MDL Consistency for SLCs)**.** *Let the assumptions of Theorem 3.4 hold. Then the minimizer* $\widehat{G}$*,*

$$\widehat{G} = \arg\min_{G} L(x^n; \mathcal{M}(G)) ,$$

converges to the ground truth graph G^* *with probability one,*

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

Proof sketch. Since $L(x; \mathcal{M})$ and $L(x^n, M)$ of Equation (3.2) are asymptotically equivalent, they have the same convergence guarantees (Grünwald, 2007). □

With this guarantee that our score is sound, we now introduce our method for discovering the entire causal network.

3.3.4 Discovering the Causal Network

We can now put all of the above together and present CDHC. We give the pseudo-code in Algorithm 3.1. We first (line 1) discover a graph G over the observed data x using a score-based structure discovery algorithm $\mathcal{A}$, such as GES (Chickering, 2002a), GGSL (Gao et al., 2017) or NOTEARS (Zheng et al., 2018). We then consider every node X_i and initialize the confounded set $X_{\mathcal{I}}$ with the Markov blanket $\mathrm{MB}(X_i)$ and add a node Z and edges $Z \to X_{\mathcal{I}}$ to G. (l. 4–5). We refine $\mathcal{I}$ by greedily adding nodes (l. 6–9), then removing nodes (l. 10–13). After finding the locally optimal set $\mathcal{I}$, we sample z from $P(Z \mid X)$ and fit a network over (x, z) using $\mathcal{A}$ (l. 14–15). Out of all these networks, we update G to be the best of them (l. 16) and iterate until convergence (l. 17). Finally, we return the discovered network G and distribution $P(X, Z)$ over X and its inferred confounders Z (l. 18).
Since our score strictly decreases at each step, our method necessarily converges. Moreover, we can show that in the large sample limit, we are guaranteed to recover the true set of confounded nodes.

Proposition 3.6 (Consistency of CDHC for Discovering Confounded Nodes)**.** *Let x^n be the an i.i.d. sample from $P \in \mathcal{M}(G^*)$ defined in Equation* (3.3)*, let Assumptions A–C hold and let $\mathcal{I}_i^*$ be the set of nodes affected by Z_i. Assume that $\bigcap_{s \in \mathcal{I}_i^*} \mathrm{MB}_{G^*}(X_s) \setminus \{Z_i\} \subsetneq \mathcal{I}_i^*$. Let $\mathcal{A}$ be consistent for recovering the Markov equivalence class of the graph G_X for distribution $P(X)$. Let $\widehat{\mathcal{I}}_i$ be the set of nodes confounded by Z_i discovered by CDHC. Then*

$$\lim_{n \to \infty} P(\widehat{\mathcal{I}}_i = \mathcal{I}_i^*) = 1 \,.$$

Proof sketch. If $\mathcal{A}$ is consistent for recovering the graph, then in particular, it is consistent for finding cliques of size ≥ 4, which can, due to Assumption B exist if and only if there is a joint latent confounder for these variables. □

Note that without further assumptions on $\mathcal{A}$, we cannot guarantee that the entire ground truth DAG G^* will be recovered. That is, so long as $\mathcal{A}$ can recover only the Markov equivalence class of a network, we cannot ensure that the edges among the observed nodes are correctly directed. However, even if we use an algorithm that *can* correctly direct edges in the absence of confounding, we still cannot guarantee that this is true when unobserved confounders are present, even if they are modeled explicitly.

3.3.5 Complexity

Last, we analyze the runtime complexity of CDHC. Let us denote the complexity of the base causal discovery algorithm $\mathcal{A}$ as $C(m, n)$. In general, $C(m, n)$ will

be linear in the number of samples n and (super-)exponential in the number of variables m. For our analysis, however, we use that the complexity of $\mathcal{A}$ can be lower bounded by $C(m, n) = \Omega(m^4 n)$. Then, the complexity of the inside of the loop (l. 2–17) can be decomposed into two parts: fitting latent confounders over different subsets of variables (l. 3–14) and fitting a new graph over the extended variable set (l. 15). The former has complexity $O(m^3 n)$ since for each of m variables, we can do at most m forward and m backward steps, and for each step, we can compute at most m candidates to be added or removed, and fitting a confounder over at most m variables is also of the order $O(m^3 n)$. The complexity of refitting the causal graph is $C(m, n)$. In total, the complexity of the inner loop is therefore $O(C(m, n) + m^4 n) = O(C(m, n))$. Furthermore, since we can find at most $O(m)$ confounded sets (see Assumption A), our worst case runtime is therefore on the order of $O(mC(m, n))$.
In practice, we expect only few confounders, $l \ll m$, to affect our observed data so that our runtime is roughly $O(C(m, n))$—the same as that of $\mathcal{A}$ itself.

3.4 Discovering Some Related Work

Over the years, a large number of methods have been proposed for causal discovery under the assumption of causal sufficiency, such as nonparametric methods on rank correlations (Lin and Peng, 2013; Yu et al., 2023; Keropyan et al., 2023), as well as extensions of ANMs to the multivariate case (Peters et al., 2014; Parida et al., 2018).
Most of these algorithms, however, do not apply when the assumption of causal sufficiency is violated. In that case, a much smaller number of algorithms such as the FCI family (Spirtes et al., 2000; Colombo et al., 2012; Ogarrio et al., 2016), 3OFF2 (Affeldt et al., 2016) and convex optimization-based approaches (Chandrasekaran et al., 2010; Agrawal et al., 2023) can find causal networks in the presence of latent confounding. Specifically, Nested Markov Models (NMMs; Shpitser et al., 2014; Evans, 2016; Shpitser et al., 2018; Richardson et al., 2012; Evans and Richardson, 2019) can sometimes provide identifiability of causal models with latent factors by using Verna constraints. Bhattacharya et al. (2021) directly recover a causal network in which all directly causal edges are directed, and edges corresponding to latent confounding are distinct from causal edges. The problem with these methods is that since they do not model the latent confounder Z directly, they cannot tell us that multiple variables are jointly confounded, leaving us with many different possible causal models over $X \cup Z$. In particular, they are generally difficult to interpret and cannot determine which sets of variables share the same latent confounder.
To discover potential confounders, Tenzer and Elidan (2016) use a copula-based approach to learning ideal parents, which are similar to our proposed approach of latent factor models, but assume that the specific data obtained

should be captured as well as possible, rather than the underlying distribution approximated. Other research controls causal estimates for latent confounders. To do so, Hoyer et al. (2008b) solve the overcomplete ICA problem in the Linear Non-Gaussian Acyclic Model (LiNGAM) setting (Shimizu et al., 2006) to correct the estimated causal effect of X on Y for confounders. Recent work has also extended this approach to larger classes of noise (Salehkaleybar et al., 2020; Adams et al., 2021; Chen et al., 2022), and showing that higher order cumulants can be used to find OICA solutions (Cai et al., 2023). However, besides using OICA to discover latent confounding, related approaches have also been used to deal with measurement error (Ding et al., 2019).
Other work has focused on reducing the requirements for identifiability of the causal model by providing weaker constraints (Kummerfeld and Ramsey, 2016; Cai et al., 2019; Bellot and van der Schaar, 2024), and introducing generalized independence criteria (Xie et al., 2020), and methods for employing proxy variables (Liu et al., 2023). In particular, recent work has extended the discovery of latent confounding to include hierarchical structures (Huang et al., 2022; Xie et al., 2022), as well as other models in which latent variables need not be root variables in the full causal model (Ghassami et al., 2021; Yang et al., 2022).
In this chapter, we explicitly exploited the violations of independence of causal mechanisms. To this end, we build on the work of Silva et al. (2006), who proposed a model based on low-rank correlation structures between observed variables, similar to prior psychometric research (Thurstone, 1934) on the number of latent factors required to describe a correlation matrix. Elidan et al. (2000) proposed an algorithm for replacing semi-cliques in a discovered causal graph with single nodes based on the idea that they are likely confounded. In contrast, we use the low-rank correlation structure between *causal mechanisms* in the inferred marginal graph and leverage this structure to find a causal network including both observed and latent variables.

3.5 Of Graphs and Goodness of Fit: Experiments

In this section, we evaluate CDHC empirically. We are interested in three things: first, how robust it is to the choice of different base causal discovery algorithms $\mathcal{A}$; second, how well it recovers the sets of confounded nodes $\mathcal{I}_j^*$; and third, how well it recovers the entire network. To test its robustness, we instantiate CDHC with three different types of causal discovery algorithms $\mathcal{A}$ and refer to CDHC using $\mathcal{A}$ as CDHC-$\mathcal{A}$. Specifically, we use a score-based method GES (Chickering, 2002a), a local-to-global graph learning method GGSL (Gao et al., 2017), and the NOTEARS approach based on continuous optimization (Zheng et al., 2018). When clear from the context, we write CDHC for CDHC-GES. Note that the omission of constraint-based methods such as the PC algorithm (Spirtes et al., 2000) is due to CDHC requiring scores to add additional latent confounders.

Next, to evaluate it on the other two metrics, we compare CDHC against several graph learning methods. First, as a baseline of a method that does not account for latent confounding, we use NOTEARS (Zheng et al., 2018). To get a better comparison on the front of discovering confounders, we also compare it against the hybrid approach GFCI (Ogarrio et al., 2016),[2] the information-theory based 3OFF2 (Affeldt et al., 2016), and the continuous-optimization based DCD (Bhattacharya et al., 2021).
We implement CDHC in Python. For the base algorithms, for GES, we use the version implemented in CAUSAL-LEARN (Zheng et al., 2024), while for NOTEARS and GGSL, we use the implementations provided by the authors. For comparison with other methods, we use the implementations provided by the respective authors for NOTEARS, 3OFF2, and DCD, and use the version of GFCI implemented in the TETRAD library (Ramsey and Andrews, 2023). All experiments finished within minutes on a commodity laptop. All code and data can be found online.[3]

3.5.1 Experiments on Synthetic Data

We evaluate CDHC on synthetic data by generating random acyclic graphs G of size m from the Erdős-Rényi model $\mathrm{ER}(m+l, p)$ with m observed and l latent variables, and edge density $p = 0.3$. We model the causal relationships via a linear SCM as in Equation (3.1), $X = AX + \alpha BZ + \varepsilon$ where $A_{ij}, B_{ij} \neq 0$ if and only the corresponding edges are in G. Nonzero causal effects of A, B are all sampled independently from $\sim N(0, 3)$, and Z, ε are sampled independently from distributions $P(Z), P(\varepsilon) \in \{N(0,1), \mathrm{Laplace}(0,1), \mathrm{LogNormal}(0,1), \mathrm{Uniform}(0,1)\}$. The parameter $\alpha \sim U[1, 8]$ determines the relative strength of confounding, allowing us to study the effects of the strength of confounding on our ability to recover the causal graph. Before we move to the general case, we begin by studying how CDHC improves over the outputs of the base algorithms $\mathcal{A}$ on a small illustrative example with $\dim(Z) = 1$.

Comparison with Base Algorithms

To see the issues that the base algorithms $\mathcal{A}$ have when a latent variable affects multiple observed variables and how CDHC improves on those results, we consider the network shown in Figure 3.5a containing nodes $X_0, \ldots, X_6$, of which X_1, X_2, X_3 are confounded by X_0. When X_0 is withheld, none of the base methods find the correct structure over X_1, X_2, X_3 (Figures 3.5b,d,f). Furthermore, while GFCI (Figure 3.5h) and DCD (Figure 3.5i) find the variables to

[2] GFCI is part of a group of methods, including FCI and RFCI. Preliminary experiments corroborated previous research (Ogarrio et al., 2016) that GFCI performs better than its relatives.

[3] `https://eda.rg.cispa.io/prj/cdhc/`

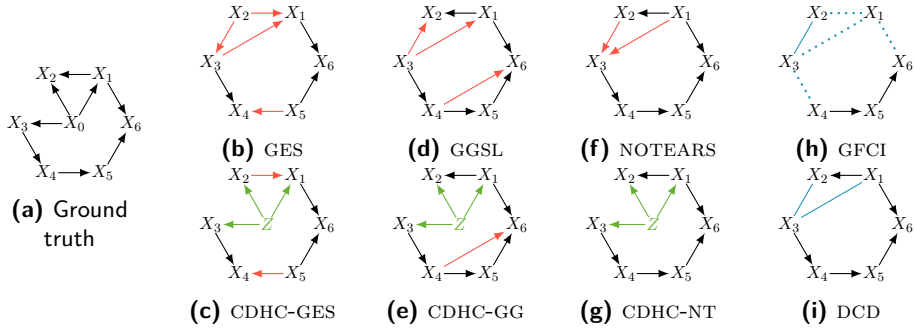


Figure 3.5: *Application of* CDHC *to synthetic data generated from the network shown in (a).* When X_0 is withheld, all base algorithms $\mathcal{A}$ (b,d,f) find a clique of spurious edges on X_1, X_2, X_3 (red). GFCI (h) and DCD (i) indicate that some pairs from X_1, X_2 and X_3 are confounded (blue) but cannot tell that they share the same confounder. In contrast, by applying CDHC (c, e, g), we discover a confounder capturing the effect of X_0 (green) and obtain higher quality networks in all cases.

be confounded, they cannot tell that all variables share the same confounder. In contrast, by applying CDHC on top of different base algorithms $\mathcal{A}$, we consistently find X_1, X_2, X_3 to be confounded while the quality of the remaining edges unaffected by the confounder is maintained (Figures 3.5c,e,g).

Confidence and Performance

To test our approach more generally, we generate 1000 datasets over $m = 50$ variables, of which ten nodes are confounded by a univariate Z. As described in Section 2.6, we care not only about the overall performance of CDHC but also how well it correlates with its confidence. Thus, we again consider decision rate (DR) plots. We compute the confidence for CDHC as

$$\mathcal{C} \coloneqq \frac{L_{\mathcal{A}} - L_{\text{CDHC-}\mathcal{A}}}{\max\{L_{\mathcal{A}}, L_{\text{CDHC-}\mathcal{A}}\}} \geq 0\,,$$

measuring how much value CDHC adds on top of simply running the base algorithm $\mathcal{A}$. For our competitor score-based methods NOTEARS and DCD, we do not have access to such a gain—in the former case because NOTEARS does not model latent confounding, and in the latter case because DCD cannot compute a causal graph without latent confounding—so that we use the score differences compared to the empty network,

$$\mathcal{C} = \frac{L_{\emptyset} - L_{\min}}{L_{\emptyset}}\,.$$

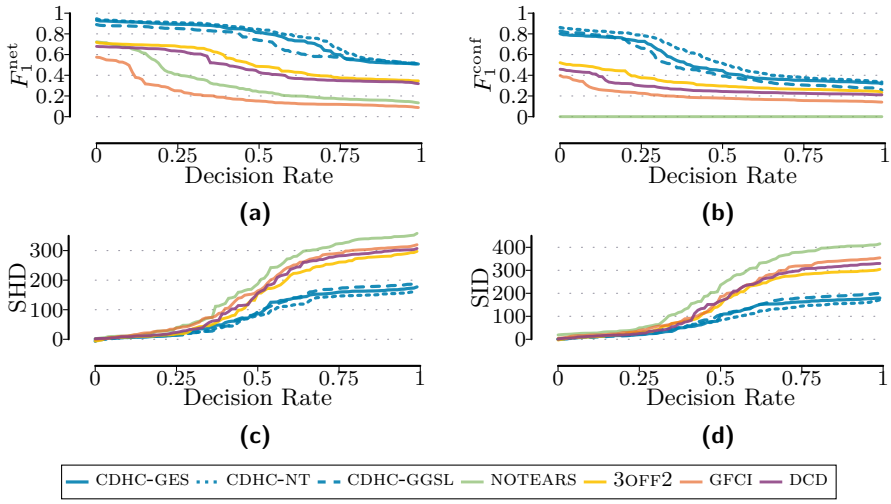


Figure 3.6: *Evaluation on synthetic data.* F_1 scores for (a) network recovery and (b) confounded set recovery (higher is better), and (c) Structural Hamming Distance and (d) Structural Intervention Distance (lower is better). Each figure shows the average score over increasing fractions of datasets, sorted in descending order by the confidence of each method. We see that CDHC clearly outperforms its competitors regardless of algorithm $\mathcal{A}$ used, and that different algorithms $\mathcal{A}$ deliver approximately equally good performance.

While this choice is not ideal, we see that both NOTEARS and DCD show the desired behavior of decreasing performance as their confidence decreases, so it seems to be a reasonable choice nonetheless. Furthermore, since neither 3OFF2 nor GFCI are score-based methods, we cannot easily define a confidence measure for them. Instead, in every evaluation, we always *order the results most favorably* for these two methods. That is, 3OFF2 and GFCI will, by construction, show monotonically decreasing performance as their "confidence" decreases.

We evaluate each method based on four different criteria chosen to capture different aspects: (1) the F_1 score for network recovery, including the confounder (F_1^{net}) measures how well we discovered the overall network, (2) the F_1 score for the recovery of the set of confounded nodes (F_1^{conf}) measures how well the set of confounded nodes specifically has been recovered, (3) the Structural Hamming Distance (SHD) between the discovered and the true networks is another measure of the structural similarity, and (4) the Structural Intervention Distance (SID) to measure differences in causal interpretations between the recovered and the true causal networks (Peters and Bühlmann, 2015).

We show the results in Figure 3.6. As in the previous chapter, the left side is where each method is most confident, and as we go toward the right, confidence decreases. We see that CDHC both outperforms its competitors by a large mar-

gin and that the relationship between its confidence and its evaluation metrics is as we would like it to be. Interestingly, the choice of the base algorithm $\mathcal{A}$ has little influence on the performance of CDHC.

Higher-dimensional Z

Next, we consider the effect of multiple confounders Z_i in our model, each influencing non-overlapping sets of 5 variables in a network of $m = 50$ variables. We show the F_1^{conf} scores for one to five confounders in Table 3.1. We omit NOTEARS, GGSL, and GES since none of them can recover confounded nodes. We see that for one to three confounders, CDHC performs at a consistently high level, but for four and five confounders, its performance decreases due to the difficulty of finding additional sets of confounded nodes. In contrast, while DCD, 3OFF2, and GFCI perform well for single-dimensional confounders, their performance precipitously drops the moment additional confounders are introduced. This behavior highlights an important difference between their approaches and ours: since they do not model the confounder but only indicate which pairs of variables are confounded, they cannot distinguish between the effects of many different confounders influencing different pairs of variables, and a single confounder affecting many variables.

How Significant Are Our Results?

To verify whether CDHC significantly outperforms its competitors, we use the Bayesian signed rank test (Benavoli et al., 2014). It explicitly models the probability that one model is significantly better than the other *in practice* by introducing a *region of practical equivalence* (rope) specified by parameter r. Two methods are considered to perform equally well if the difference in scores for the methods lies in $[-r, r]$. We pick $r = 0.05$ (Benavoli et al., 2014) but the conclusion remains the same for values $r \in (0, 0.15]$. Since the test was designed for *two* competing methods, for we compare CDHC with the dataset-wise best competitor, which we refer to as OPT. For each dataset k, we compute the F_1^{net} scores for both CDHC and OPT and compute their differences.

	Number of confounders				
Method	1	2	3	4	5
CDHC	**0.43**	**0.38**	**0.35**	**0.23**	**0.15**
DCD	0.35	0.18	0.11	0.07	0.03
3OFF2	0.36	0.2	0.14	0.11	0.04
GFCI	0.22	0.11	0.05	0.02	0.01

Table 3.1: *Comparison of* CDHC, DCD, 3OFF2 *and* GFCI *for varying numbers of latent confounders.* Only CDHC performs well as the number of latent factors increases.

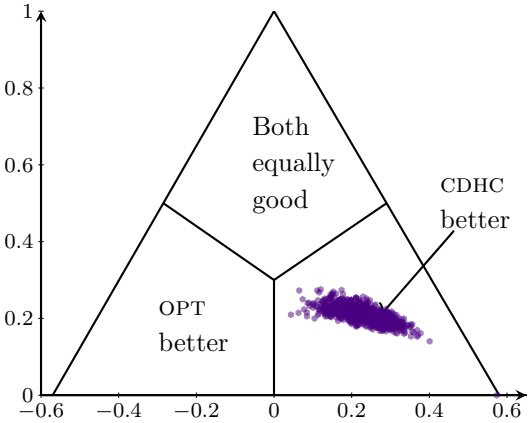


Figure 3.7: *Significance assessment of the improvement of* CDHC *over its competitors, in barycentric coordinates.* Points in the bottom left and bottom right regions indicate OPT, respectively, CDHC performing significantly better, while points at the top indicate practical equivalence. Since all points lie in the bottom right region, we conclude that CDHC practically significantly outperforms its competitors.

To compare the two methods over all samples, we aggregate the F_1 scores for both CDHC and OPT and take their differences $z_k = F_{1,k}^{\text{OPT}} - F_{1,k}^{\text{CDHC}}$, $k \in \{1, \ldots, q\}$. To include the prior assumption that both methods are equally good, we include a pseudo-observation $z_0 = 0$, i.e., that both methods are precisely equally good. We take weights $w = (w_0, \ldots, w_q) \sim \text{Dirichlet}(s, 1, \ldots, 1)$ where s corresponds to the number of times we obtained z_0. This is commonly set to be $s = 0.5$, but due to our large number of experiments, its influence on the posterior is minor. The posterior probabilities are computed as

$$\theta_{\text{OPT}} = \sum_{i,j=0}^{q} w_i w_j I_{(2r,\infty)}(z_i + z_j)$$

$$\theta_{\text{rope}} = \sum_{i,j=0}^{q} w_i w_j I_{[-2r,2r]}(z_i + z_j)$$

$$\theta_{\text{CDHC}} = \sum_{i,j=0}^{q} w_i w_j I_{(-\infty,-2r)}(z_i + z_j)\,,$$

where $\theta_{\text{OPT}}, \theta_{\text{CDHC}}$ are the posterior probabilities that OPT, respectively CDHC are better by at least a margin r, while θ_{rope} is the posterior probability that they perform practically equally well. The distribution of θ is not analytically tractable, but we can evaluate it empirically by sampling values for w. We depict the result of such a sample in Figure 3.7. Points in the bottom left and bottom right areas correspond to OPT outperforming CDHC, respectively CDHC outperforming OPT, while points in the upper area indicate both methods performing equally well. Since all points lie in the bottom right corner, we see that CDHC performs significantly better than OPT across the board.

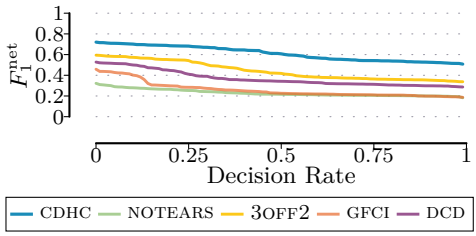


Figure 3.8: [Higher is better] *Decision rate plot for* CDHC *and its competitors on the REGED dataset.* Overall, CDHC outperforms all other methods across the board, both for points where they are confident as well as those where they are not.

3.5.2 Realistic Data: REGED

Next, we consider realistic synthetic data from REGED (Guyon et al., 2008) based on human lung-cancer microarray gene expression data. Since the available $m = 1000$ samples are non-i.i.d., the causal relationships are nonlinear, and the ground truth is known from gene intervention studies, it provides a good test for CDHC when its assumptions are violated.

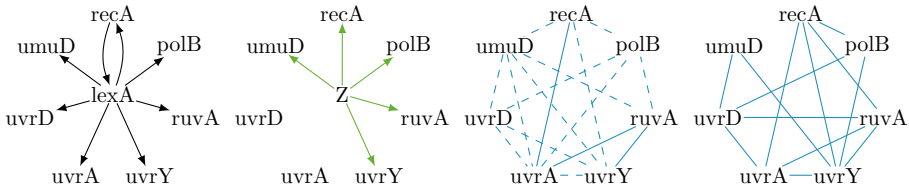
To make CDHC applicable to the REGED dataset, we consider the following setup. For each node X_i in the ground truth graph G^* with $k \geq 5$ children, the set of which we denote by $K = K_i$, we select a random subset $R = R_i$ also consisting of k nodes of G^* which do not have a common parent in G^*. We then consider the induced subgraph G_i over the nodes $K_i \cup R_i \cup \{i\}$. However, the data given to each method is only over the variables $X_{K \cup R}$, so we should recover the nodes in $K = K_i$ to be jointly confounded.

We show the results for F_1^{net} for different methods in a DR plot in Figure 3.8. Even though the data violates our assumptions, CDHC outperforms its competitors by a large margin. Moreover, even for those sets of variables where CDHC is only moderately confident, it still performs better than its competitors at their *most* confident. This suggests that CDHC works reliably even when the true model deviates from our assumptions.

3.5.3 Case Study: Cellular Signaling

Last, we consider real-world data to investigate the interpretability of the results returned by CDHC. In particular, we consider the SOS DNA repair network in E. coli (Ronen et al., 2002). This data consists of protein levels of eight genes measured every five minutes for five hours, resulting in 60 samples. Since the governing relationships in gene regulation are highly nonlinear, this further tests the applicability of CDHC when our assumptions do not hold.

Since the ground truth network has been established (Perrin et al., 2003), we can test CDHC by excluding a gene known to have a downstream causal effect on other genes. An excellent candidate is *lexA* as it has a causal influence on *all* of the other genes: it is upstream of six genes and has a bidirectional



(a) Ground Truth **(b)** CDHC **(c)** GFCI **(d)** DCD

Figure 3.9: *Results on SOS DNA repair network in E. coli.* CDHC (b) discovers a confounder Z capturing five out of the seven edges (green) in the ground truth network (a). In contrast, GFCI (c) and DCD (d) find many pairs of nodes that are confounded (solid blue), and in the case of GFCI, more pairs yet which *might* be confounded (dotted). However, they do not discover that all nodes share the same confounder, making the resulting networks challenging to interpret.

relationship with the seventh (Figure 3.9a). Since 3OFF2 provides very similar results to GFCI, and NOTEARS cannot model confounders, we omit their results. We show the results in Figure 3.9. For clarity, we focus only on discovering which nodes are confounded, omitting causal edges for all methods. In Figure 3.9b, we find a striking similarity between the Z discovered by CDHC and the true common parent *lexA*. CDHC correctly identifies five out of seven relationships: four out of six downstream effects and one of the two edges between recA and lexA—which is the most a DAG can do, given that the two edges are mutually exclusive. Next, for GFCI (Figure 3.9c) and DCD (Figure 3.9d), we indicate definite confounding by solid edges and pairwise correlations which could be due to either confounding or causation by dotted edges. GFCI indicates definite confounding for only three out of 16 pairs, while we cannot be certain for the other pairs. The resulting network of DCD indicates definite confounding for many pairs of variables. However, neither method can determine that all variables share the *same* latent confounder. The results of both GFCI and DCD are consistent with many different confounder structures and provide no well-founded way of choosing one over the other. Overall, despite the low sample size and violations of our model assumptions, CDHC finds a more readily interpretable network that is close to the ground truth.

3.6 Limits of Linearity: Navigating Nonlinear Networks

In this chapter, we tackled Problem 2a) and studied to what extent we can discover the causal network underlying a distribution $P(X)$ when the observed variables X are affected by latent confounders Z. That is, we wanted to find out whether, given a sample x from only $P(X)$, we can recover the true causal graph G^* governing the factorization of $P(X, Z)$.

We began by relating our problem of recovering the causal graph to the commonly used framework of turning causal discovery with linear effects into a problem of finding the mixture matrix in ICA. We noted that while the overall mixture matrix can be recovered under general conditions, it is impossible to determine which correlations are due to direct causal effects between the observed variables and which are due to latent confounding.

In order to obtain better identifiability results, we studied the graphical structures induced by latent confounding (Proposition 3.1), and from these structures, we derived sufficient conditions for the identifiability of the causal model. In particular, when the causal model is a sparse linear causal (SLC) model, the causal structure is identifiable up to trivial transformations. Unlike the standard ICA approaches, we can uniquely determine the *full* graphical structure between observed and unobserved variables up to relabeling of the exogenous variables (Theorem 3.2). Furthermore, in the limit of very large graphs, we showed that sparsity is no longer required (Theorem 3.3).

We then showed that under the same assumptions, both BIC and MDL scores are consistent for recovering the true model (Theorem 3.4 and Theorem 3.5). Furthermore, we developed a general framework for discovering the causal network by combining the output of *any* causal discovery algorithm relying on causal sufficiency with standard latent factor modeling. We showed that so long as the used causal discovery algorithm is consistent, this will indeed recover the correct sets of latent variables (Proposition 3.6).

In several experiments, we showed that CDHC works well in practice on both synthetic and real-world data. In particular, as with CoCa, we saw that its confidence tracks its performance well, providing an excellent observable proxy.

For future work, one vector of improvement could be the theoretical guarantees. For one, given a fixed number l of latent variables, our model constraints imply that there are at most $2m - 4l$ parameters. In contrast, by modeling a full causal graph, we could employ a total of $m(m-1)/2$ parameters. That is to say, from a pure parameter counting point of view, it should be possible to relax Assumptions A and B dramatically and still maintain a sizeable gap between the two kinds of model. As we have seen in Theorem 3.3, this is already borne out for large graphs, but it would be good to have more precise thresholds on what can, or cannot, be done.

Furthermore, while the complexity of CDHC is asymptotically not much worse than running a standard causal discovery algorithm, it nevertheless currently requires multiple passes over the data, learning causal networks repeatedly over only slightly changing datasets. It would, therefore, be interesting to see to what extent we can design a method that directly compares the different options for causal edges and (local) confounders at each step of a graph learning algorithm, reducing the overhead currently required.

Of course, another potentially large issue for the relevance of the method is

its assumption of strict linearity. Unfortunately, even though we have seen CDHC to work well in practice, the real world is rarely so kind as to match this assumption precisely. In the next chapter, we move on to models with specific kinds of nonlinearities, for which we can derive similar identifiability results under some additional assumptions on the nonlinear functions used.

Chapter 4

Nonlinear Causal Discovery with Latent Confounders

"There are more things in heaven and earth, Horatio, than are dreamt of in your philosophy."

William Shakespeare: Hamlet, Act I, Scene V

In the previous chapters, we focused on linear causal models. Unfortunately, while much of the theory of causal inference and causal discovery has been developed for linear SCMs (Angrist and Pischke, 2009; Shimizu et al., 2006, 2011), the real world is rarely so accommodating.
Especially in biological (Gourévitch et al., 2006), social (Nazlioglu, 2011) and economic systems (Nishiyama et al., 2011), causal relationships are often far from linear. Furthermore, our measurements are rarely perfect either, but instead subject to certain nonlinear distortions (Hoyer et al., 2008a; Zhang and Hyvärinen, 2010; Bühlmann et al., 2014). To deal with these domains and issues, it is essential to develop theories that can deal with some, if not all, types of nonlinearities that could affect our causal relationships.
We can write a nonlinear SCM with observed X, latent Z, and noise ε as

$$X = f(X, Z, \varepsilon), \tag{4.1}$$

where f is some nonlinear function, corresponding to the graph G^* where the causal parents Pa_i of each X_i are precisely those variables for which f_i is not constant. Naturally, some nonlinear functions f will make recovering the causal

graph G^* easier than others. By building on the theory we have built in the previous chapters, we will see that by using models of the form

$$X = \tau((I - A)^{-1}(BZ + \varepsilon)),$$

we can extend the analysis in the last chapter to the nonlinear case. Here the nonlinearity τ is given by $\tau = (\tau_1, \ldots, \tau_m)$, where each component τ_i is a univariate function $\tau_i(X_i)$. Then all correlations in the observed data are *solely* due to the matrices A and B defining the causal structure, and the nonlinear functions τ_i act as *post-nonlinearities* distorting the observed variables X_i in a nonlinear manner (Zhang and Hyvärinen, 2010).
To tackle the issue of identifiability in this generating model, as in the previous chapter, we begin by formalizing how nonlinear SCMs relate to linear SCMs (Section 4.1.2) and put our problem in the framework of nonlinear ICA (Section 4.1.3). In Section 4.2, we show that when the functions τ_i behave nicely, the same identifiability results as for the linear case are obtainable.
Despite the similarity of the identifiability results of the causal graph G^* and the matrices A, B, learning these matrices is nevertheless difficult. Unlike in the previous chapter, we need to learn not only A, B but also the nonlinear transformations τ of the variables so that these causal matrices apply. To do this, we use a variational autoencoder (VAE; Kingma and Welling, 2014), which allows us to learn both τ and A, B simultaneously (Section 4.3). We show that when $\tau = \mathrm{id}$, our proposed score is consistent for recovering the true matrices A and B, and therefore the causal graph G^*. In Section 4.5, we then show that NOCADILAC performs well in practice on synthetic and real-world data, even beyond the linear setting for which we showed consistency. In particular, it is robust to the dimensionality of the latent confounder and returns more readily interpretable results than its competitors. As in previous chapters, we postpone proofs for all theoretical statements to Appendix A.4.

4.1 Bending the Rules: Nonlinearity in Causality

We begin by introducing the specific causal model we consider in this chapter and show that while the framework of nonlinear ICA provides relevant intuition on the difficulties, it does not suffice to determine the causal graph.

4.1.1 Notation

As in the previous chapters, we assume that the observed X and unobserved Z follow a joint probability distribution $P(X, Z)$ which factorizes according to the causal DAG G^*. Furthermore, nonlinear functions are written as τ, ν, and we assume that τ denotes element-wise transformation $\tau(y) = \tau(\tau_1(y_1), \ldots, \tau_m(y_m))$

where each τ_i is invertible and three times differentiable. From this differentiability and invertibility, it follows that each τ_i is strictly monotonic.

As before, our goal is to recover the causal graph G^* from a sample x from the observed distribution $P(X)$ without any knowledge about the latent Z, nor the nonlinear transformations τ. To formalize this problem, we define the precise SCMs from which we assume our data is generated.

4.1.2 Beyond Linear SCMs

In the previous chapter, we assumed that our data comes from the linear SCM,

$$X = AX + BZ + \varepsilon\,,$$

where A, B are the matrices parametrizing the graph G^*. A general way to extend this is to consider *additive noise models* (Hoyer et al., 2008a),

$$X = f(X, Z) + \varepsilon\,, \tag{4.2}$$

where the function f takes the places of the matrices A, B. If f is linear in X and Z, then we would simply find that $A = \nabla_X f$ and $B = \nabla_Z f$ where $\nabla_X f = (\partial_{X_1} f, \ldots, \partial_{X_m} f)$ is the gradient of f with respect to X and similarly for ∇_Z. If f is sufficiently well-behaved (Rudin, 1953), then we can "solve for X" and there exists a function g such that we can write

$$X = g(Z, \varepsilon) = g(S)\,, \tag{4.3}$$

where as in the previous chapter we denote by $S = (Z, \varepsilon)$ all exogenous sources of the causal model. As we have seen in Section 3.1.2, the linear SCM specifically can be rewritten as a linear mixture of S as

$$X = (I - A)^{-1}\,(BZ + \varepsilon)\ ,$$

so that a natural path from this description to the model of Equation (4.3) is to add nonlinearities τ and write

$$X = \tau\left((I - A)^{-1}\,(BZ + \varepsilon)\right)\ , \tag{4.4}$$

where τ is an element-wise transformation. In this model, all correlations are purely due to the mixing matrices $(I - A)^{-1}$ and B. This model is known as post-nonlinear mixture model (Taleb and Jutten, 1999), and models nonlinear distortions in the measured variables X.

Next, we will explore why the more general model class of Equation (4.3) is not suitable for our purposes of causal inference with latent confounders.

4.1.3 Nonlinear ICA and Causal Discovery

In Section 3.1.2, we saw that while linear ICA can identify parts of how the data X comes about, it cannot distinguish correlations due to latent confounders from those due to causal mechanisms between the observed variables.
In causal discovery without latent confounding, it is well-known that the addition of nonlinearities can often be beneficial to the identifiability of the underlying causal DAG (Bühlmann et al., 2014; Peters et al., 2014). In this section, we investigate whether nonlinearities here serve a similar purpose.
To answer this question, let us study the model of Equation (4.3),

$$X = g(S)\,,$$

for some independent sources S with $\dim(S) = s \geq m = \dim(X)$ and nonlinear functions $g = (g_1, \ldots, g_m) : \mathbb{R}^s \to \mathbb{R}^m$. The goal is then to find a nonlinear function h such that $h \circ g \approx \mathrm{id}$, where we will specify $\approx$ more precisely below.
For linear ICA, we saw that under reasonable assumptions of non-Gaussianity and distinguishability of sources, the mixing matrix can be identified up to trivial indeterminacies. However, the problem of nonlinear ICA is generally highly underdetermined, allowing for a great many indeterminacies.
Naturally, as in the linear case, we will again have indeterminacies of joint permutations of the functions and sources, $g_1 \leftrightarrow g_2, S_1 \leftrightarrow S_2$. Furthermore, similar to the linear case, element-wise rescaling $g_i \mapsto g_i(as), s_i \mapsto s_i/a$ become element-wise nonlinearities $g_i \mapsto g_i \circ h_i, s_i \mapsto h_i^{-1}(s_i)$ where h_i is any invertible function. Since these correspond to the "trivial" indeterminacies of the linear case, our notation $\approx$ above precisely corresponds to invertibility up to these indeterminacies (Hyvärinen and Pajunen, 1999). That is, when we say that $h \circ g \approx \mathrm{id}$, we mean that $h \circ g = (\nu_1, \ldots, \nu_s)$ for some ν such that each $\nu_i(s_i)$ is an element-wise nonlinearity. Whereas further indeterminacies can be ruled out in the linear case by assuming non-Gaussian sources, this is unfortunately no longer true in the nonlinear case.

Additional Indeterminacies in Nonlinear ICA

We give here two specific kinds (Hyvärinen and Pajunen, 1999; Darmois, 1953). The first type is that of measure-preserving automorphisms (MPAs; Hyvärinen and Pajunen, 1999), which correspond to the orthogonal rotations of Gaussian sources of the previous chapter. The idea is simple: map the sources onto a Gaussian distribution, rotate them, and then invert the map (Figure 4.1). More precisely, let $F(s)$ be the joint cumulative density function (CDF) of S, and let Φ be the CDF of the standard multivariate Gaussian. Then

$$\widetilde{S} \coloneqq F \circ \Phi^{-1} \circ U \circ \Phi \circ F^{-1}(S) \sim S\,,$$

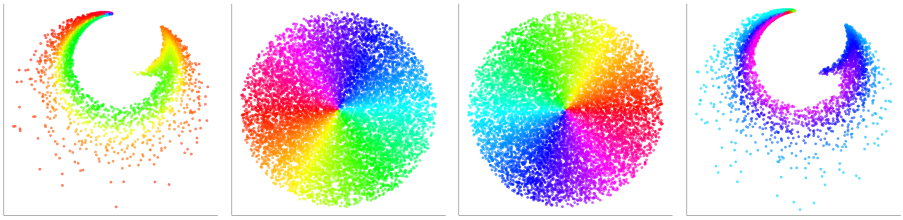


Figure 4.1: *An example MPA of the distribution in (a).* By mapping it first to a Gaussian distribution (b), we can then rotate this distribution (c) and map it back to the original space (d) to obtain an automorphism of the original distribution (a).

has the same distribution as S. That is, since $\Phi \circ F^{-1}$ map S to a standard Gaussian and $F \circ \Phi^{-1}$ map a standard Gaussian precisely to the correct distribution of S, we can apply whatever orthogonal matrices U between these two steps and leave the distribution the same. Therefore

$$\left(g \circ F \circ \Phi^{-1} \circ U \circ \Phi \circ F^{-1}\right)(S) = g(S)\,,$$

would be two different functions mapping S to exactly the same target distribution of $P(X)$. By allowing for nonlinear functions, we implicitly opened the back door to Gaussian source variables, for which MPAs are easy to find. We illustrate this in Figure 4.1. While the distribution in Figure 4.1(a) has no apparent symmetries, we can nevertheless subject it to a transformation as described above to obtain a measure-preserving map of the distribution, making identifiability without further assumptions impossible.

The second non-trivial indeterminacy is given by the Darmois construction (Darmois, 1953; Skitovitch, 1953). The Darmois construction yields sources S^D by

$$s_i^D = f_i^D(x_{1:i}) = P(X_i \leq x_i \mid x_{1:i-1})\,,$$

such that by construction, we have

$$X = (f^D)^{-1}(S^D)\,.$$

Furthermore, each source S_i^D is constructed to be independent of all $S_{1:i-1}^D$ so that all sources S^D are jointly independent. Furthermore, by changing the order of the variables from $X_1, \ldots, X_m$ to $X_{\pi^{-1}(1)}, \ldots, X_{\pi^{-1}(m)}$, we obtain different functions and sources f^D, S^D and f_π^D, S_π^D, giving us another source of non-identifiability of the nonlinear ICA solution.

Can We Save Nonlinear ICA?

In recent years, much progress has been made in finding conditions under which nonlinear ICA models are identifiable, relying on additional structure of the problem. We focus on three different approaches most relevant to our problem here and explain why they do *not* solve the problem we are interested in.

First, Khemakhem et al. (2020) relate the study of nonlinear ICA models to that of variational autoencoders (VAEs; Kingma and Welling, 2014). In their setup, auxiliary variables U modeling varying *contexts* are used, and they assume the generating mechanism to be of the form

$$\begin{aligned} X &= g(S) + \varepsilon \\ S &\sim p(s \mid u) = h(s) \exp(\eta(u)^\top T(s) - A(u)) \,, \end{aligned} \tag{4.5}$$

so that S is sampled from an *exponential family* whose parameters depend on the auxiliary variables U. That is, the mechanism describing how X depends on S is stable across contexts, while the distribution of S changes across contexts. Here, the parameters $\eta(u)$ are called the *natural parameters* of the exponential family, and $T(s)$ is called the *sufficient statistic*. They show that under some technical conditions, such as g being injective, the model is identifiable when data from a sufficiently large number of diverse contexts U is gathered. While this is an interesting approach, it is unsuitable for our task since it requires data from multiple contexts. Furthermore, while at first glance it may look as if this approach is suitable for dealing with confounding, since including both sources S and additional noise variables ε results in more sources total than observed variables, this is not the case. This is because the noise ε in Equation (4.5) is not part of the "mixed" variables introducing correlations between the observed variables X, unlike noise variables ε in the SCM formulation of Equation (4.2).

Second, Gresele et al. (2021) propose a principle called independent mechanism analysis (IMA). That is, starting from the *algorithmic* independence of causal mechanisms (see Section 2.2.3), they propose that the functions g in the nonlinear ICA equation should all satisfy that at every point s the columns of the Jacobian matrix $Dg(s)$ are orthogonal to each other. That is, they require

$$Dg(s) = \lambda(s) U(s) \,,$$

where $\lambda(s)$ is a diagonal matrix and $U(s)$ an orthogonal matrix. They show that, in this case, the Darmois construction is no longer a non-identifiability of nonlinear ICA. Furthermore, when $\lambda(s)$ is a scalar, MPAs are also no longer non-identifiabilities of nonlinear ICA (Gresele et al., 2021). In general, even when the theoretical guarantees do not apply, empirical follow-up work shows that their proposed approach recovers the correct mixing function for randomly initialized multi-layer neural networks (Sliwa et al., 2022). However, by requir-

	S_1	S_2	S_3	S_4
X_1	•		•	
X_2		•		•
X_3	•	•		
X_4			•	•

Figure 4.2: *Depiction of structural sparsity.* Each of the source variables S_i has a subset of variables $X_\mathcal{I} \subseteq X$ that is *uniquely* jointly affected by S_i. This increases identifiability of the nonlinear ICA model by ensuring that all sources have distinguishable effects on the observed variables.

ing $Dg(s)$ to have orthogonal columns, they require that $\dim(S) \leq \dim(X)$, making it impossible to model latent confounding with this approach.

In contrast, Zheng et al. (2022) show that the model can be identified under assumptions on the *structural sparsity* of g. More precisely, when g is smooth and invertible, and for each source S_i, there exists a subset of variables $X_\mathcal{I}$ of the variables X such that S_i is the only source affecting *all* of the variables $X_j \in X_\mathcal{I}$. See Figure 4.2 for a graphical depiction of structural sparsity. However, once more, this approach has only been shown to work when there are as many sources as observed variables, $\dim(S) = \dim(X)$.

In addition to the problems of dimensionality of the source space with all of the above approaches, the latter two approaches differ from our goal at a more fundamental level. That is, both approaches make assumptions about the structure of the function g relating the sources S to the observed variables X rather than assumptions about the function f relating the observed variables X to each other. Since it is this latter function that we care about, it appears pertinent to make assumptions about f rather than g. To see this more clearly, let us consider the linear case again for a moment. By writing

$$X = C\varepsilon\,,$$

and making assumptions about C, we do not obtain any interesting or interpretable constraints for the matrix A of the equivalent parametrization

$$X = AX + \varepsilon\,.$$

That is, since inverses of sparse matrices are not generally themselves sparse, the sparsity of C as proposed by Zheng et al. (2022) does not imply sparsity or any other interesting structures in $A = I - C^{-1}$. Similarly, the orthogonality of the columns of C proposed by Gresele et al. (2021) does not correspond to orthogonality of columns in A, nor to any constraints about the matrix A that can be readily interpreted or expected to hold in a real causal system. Conversely, more natural assumptions in the *causal parametrization* of Equation (4.1), such as sparsity or independence of causal mechanisms in the nonlinear function f, do not readily translate into assumptions on the mixing function g under which the nonlinear ICA problem is known to be identifiable.

Next, we show that in contrast to the general nonlinear ICA problem described in Equation (4.3), the problem of finding a post-nonlinear model over both observed and latent variables as described in Equation (4.4) is possible under slightly stricter assumptions than those we developed in the previous chapter.

4.2 Nonlinearity? No Problem: Identifiability

As we have seen in the previous chapter, when the effects of latent confounders are sufficiently distinct from each other and from causal effects between observed variables, we have a chance at recovering the ground truth causal network. More precisely, in addition to standard causal discovery assumptions, we made the following structural assumptions in the previous chapter, which we restate here for ease of readability.

Assumption A (Sufficient Signal of Confounders)**.** The variables X are split into disjoint sets $\mathcal{I}_1, \dots, \mathcal{I}_l$ of size $|\mathcal{I}_j| \geq 4$, such that Z_j has non-zero influence on each $X_i \in \mathcal{I}_j$, i.e., $b_{ij} \neq 0$ for all $X_i \in \mathcal{I}_j$.

Assumption B (Sparsity)**.** For each $\mathcal{I}_j$, there are at most $|\mathcal{I}_j| - 4$ edges incoming to vertices in $\mathcal{I}_j$, aside from those starting in Z_j.

Assumption C (No False Positives)**.** For all distinct X_i, X_j, X_u, X_v not independent given Z, with covariances σ_{rs}, we have $\sigma_{ij}\sigma_{uv} \neq \sigma_{iu}\sigma_{jv}$.

These assumptions guarantee that the true network structure is sufficiently sparse in a suitable manner and that the matrices A, B are not picked in an adversarial manner. However, in addition to these structural assumptions, we need to ensure that the nonlinearities τ_i are sufficiently well-behaved.

Assumption D (Super Linear or Superlinear)**.** The confounder Z is Gaussian, $Z \sim N(0, \operatorname{diag}(\sigma_Z^2))$. Further, precisely one of the following holds:

a) $\tau = \mathrm{id}$ is the identity function and $\varepsilon \sim N(0, \sigma_x^2 I)$, or
b) τ_i is three times differentiable and strictly nonlinear for all i.

While this assumption is not strictly necessary to recover the effects of the latent confounders Z, we require it to identify the matrix A determining causal relationships between the observed variables X. The first case corresponds to the well-known identifiability of causal models for linear Gaussian models with equal noise variances (Peters and Bühlmann, 2014). The second case corresponds to the identifiability of PNL models (Zhang and Hyvärinen, 2009; Peters et al., 2014). By adding this last assumption, we ensure that the identifiability result of Theorem 3.2 translates into our current setting.

Theorem 4.1 (Identifiability in the Sparse PNL Model)**.** *Let the distribution $P(X, Z)$ be described by the nonlinear SCM of Equation* (4.4)*, i.e.,*

$$X = \tau\left((I - A)^{-1}(BZ + \varepsilon)\right),$$

for some Z of dimension $l \leq m/4$. Further, let Assumptions A–C hold. Then both the number l of confounders Z and the causal effects of the confounder, B, are identifiable up to trivial indeterminacies (column permutations and rescaling). Furthermore, if Assumption D also holds, then A is also identifiable.

Proof sketch. Due to the way the model is generated, all correlations between observed variables X can be described entirely by linear relationships between transformed variables $\tau^{-1}(X)$, for which Theorem 3.2 applies. □

As with Theorem 3.2, we do not need to know the number of confounders, nor which subset of the variables X is confounded by each of the confounders Z_j. Instead, all of these can be recovered solely from the observed data. If we further assume that the variances of Z are known, that the τ_i are normalized, and that we have some domain knowledge, we can get more precise identifiability.

Corollary 4.2. *Let the Assumptions of Theorem 4.1 hold and let all τ_i be strictly increasing and standardized to satisfy $\tau_i(1) = 1$. Further, let the variances $\sigma^2_{Z_j}$ of each Z_j be known, and for each j let the sign of b_{ij} be known for at least one $X_i \in S_j$. Then B is identifiable up to permutations of its columns.*

Proof sketch. The additional assumptions that $\sigma^2_{Z_j}$ are known and that $\tau_i(1) = 1$ fix the scale of B. The assumption that for each Z_j, the value of at least one $b_{ij} \neq 0$ is known fix the sign of B. □

The required domain knowledge here is the knowledge of the signs of some b_{ij} and the variance of Z. While this assumption may seem unreasonable at first glance, in many scientific fields, we have a good idea of the potential latent confounders and how they would affect some of the observed variables. Examples of cases where these assumptions hold include various psychometric constructs (Gerber et al., 2011; Grosse and Zhou, 2021), socio-economic status in microeconomic or epidemiological analyses (Hajat et al., 2021), and GDP in cross-country macroeconomic analyses (Hu et al., 2015).
Note that Theorem 4.1 makes no statement about the identifiability of the nonlinear functions τ. In fact, we are not interested in nonlinearities τ_i themselves, but instead in discovering the underlying generating DAG G and the effects of the latent confounder Z, so that this is not an issue for our purposes. As long as we can find *any* element-wise nonlinearity ν such that $\nu(X) \sim N(0, \Sigma)$ for some Σ, we have achieved our goal. Therefore, we next develop a method to find such a ν along with the graph G^*.

4.3 Learning with Variational Autoencoders

In order to learn a nonlinear function ν such that $\nu(X)$ is normally distributed, we make use of variational autoencoders (VAEs; Kingma and Welling, 2014, 2019). We begin by giving a short introduction to VAEs and variational methods. Our goal is to estimate the evidence or marginal log-likelihood,

$$\log p(X) = \log \int p(X \mid H) p(H) dH \,,$$

of a model, where we use H to denote whatever parameters or auxiliary variables we do not care about. For example, in a Gaussian mixture model, we might want to know how probable the observed data is after marginalizing over the "cluster assignments" for each data point. While the typical approach is to evaluate the joint probability at the MAP cluster assignment to approximate $\log p(X) \approx \max_H \log p(X \mid H) p(H)$, this is a strict underestimate of the true $\log p(X)$, especially when many values of H obtain roughly equally good scores, such as when the joint distribution is multi-modal or when the modes are flat. To obtain better lower bounds, we write (Kingma and Welling, 2014)

$$\begin{aligned}
\log p(X) &= \log \int p(X \mid H) p(H) dH \qquad (4.6)\\
&= \log \int q(H \mid X) p(X \mid H) \frac{p(H)}{q(H \mid X)} dH \\
&\geq \int q(H \mid X) \log \left(p(X \mid H) \frac{p(H)}{q(H \mid X)} \right) dH \\
&= \int q(H \mid X) \log p(X \mid H) dH - \int q(H \mid X) \log \left(\frac{q(H \mid X)}{p(H)} \right) dH \\
&= E_{q(H|X)} \left[\log p(X \mid H) \right] - \mathrm{KL} \left(q(H \mid X) \mid p(H) \right) \,,
\end{aligned}$$

where $\mathrm{KL}(p, q) = \int p(u) \log \left(p(u)/q(u) \right) du$ is the Kullback-Leibler divergence, and we used Jensen's inequality in the third line (Durrett, 2019). The two conditional distributions $q(H \mid X), p(X \mid H)$ are called *encoder*, respectively, *decoder*, corresponding to probabilistic maps $X \mapsto H$ and $H \mapsto X$ such that their composition approximates the identity function.

The right-hand side of the inequality is referred to by the eminently sensible

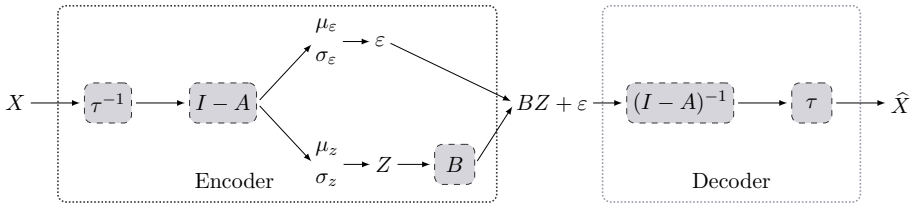


Figure 4.3: *Proposed model architecture.* The encoder outputs both noise ε and confounder Z. Shaded nodes are the learned mechanisms. Learnable parameters include the adjacency matrix A of the causal model over the observed variables as well as the matrix B containing the influence of Z on the observed nodes.

name *evidence lower bound* (ELBO; Kingma and Welling, 2014),

$$L_{\text{ELBO}} \coloneqq E_{q(H|X)} \left[\log p(X \mid H)\right] - \text{KL}\left(q(H \mid X) \mid p(H)\right) .$$

The two terms ELBO are the *reconstruction error* $E_{q(H|X)} \left[\log p(X \mid H)\right]$, measuring the capacity of the "channel" defined by the encoder $q(H \mid X)$ and the decoder $p(X \mid H)$ to recover the original data, and the *divergence* measuring the similarity between the prior $p(H)$ and the encoder $q(H \mid X)$. Note that the inequality in Equation (4.6) holds for *any* encoder $q(H \mid X)$ and decoder $p(H \mid X)$. Hence, in order to approximate the evidence $\log p(X)$ as well as possible, we can maximize this lower bound,

$$\max_{q(H|X),\, p(X|H)} L_{\text{ELBO}} \leq \log p(X) \,, \tag{4.7}$$

over some class of encoder and decoder distributions, also known as *proposal distributions*. Clearly, the larger the set of proposal distributions, the better the approximation. Historically, due to computational limitations, variational approaches were restricted to simple proposal distributions such as mean field approximations, for which solutions were analytically tractable (Girardeau and Mazo, 1973; Tanaka, 1998; Wainwright et al., 2008). With today's abundance of computational power, more modern variational approaches instead use more expressive, but also more computationally expensive classes of distributions, which therefore provide tighter lower bounds (Kingma and Welling, 2019). Specifically, both the encoder $q_\theta(H \mid X)$ and the decoder $p_\lambda(X \mid H)$ are generally chosen to be Gaussian distributions parameterized by neural networks so that the optimization in Equation (4.7) is over some parameter vectors θ, λ parametrizing the elements of these larger classes of distributions.

To apply this to our problem, we set $H = BZ + \varepsilon$. In order to estimate our model evidence, what would be good choices for the encoder and decoder?

Fortunately the causal model of Equation (4.4) already describes a decoder,

$$X = \tau\left((I - A)^{-1}(BZ + \varepsilon)\right),$$

so that we only need to determine a suitable corresponding encoder. By rewriting this equation, the unobserved source variables can be written as

$$BZ + \varepsilon = (I - A)\tau^{-1}(X),$$

so that we can write a corresponding encoder as

$$H \sim N((I - A)\tau^{-1}(X), \sigma_z^2(X)BB^\top + \sigma_\varepsilon^2(X)I).$$

Equivalently, we can split H into the effect of the latent confounding variables Z and the independent noise variables ε by writing it in the form

$$\begin{aligned} H &= BZ + \varepsilon \quad \text{where} \\ Z &\sim N(\mu_z(X), \sigma_z^2(X)I) \\ \varepsilon &\sim N(\mu_\varepsilon(X), \sigma_\varepsilon^2(X)I) \\ \text{s.t.} \quad B\mu_z &+ \mu_\varepsilon = (I - A)\,\tau^{-1}(X). \end{aligned}$$

We show the full model in graphical form in Fig. 4.3. Note that absent any penalties associated with μ_z, μ_ε, we can simply set $\mu_z = 0$ and $\mu_\varepsilon(X) = (I - A)\,\tau^{-1}(X)$. However, if we have reason to believe that μ_z should be used to model the mean instead of μ_ε, we can alternatively set

$$\begin{aligned} \mu_z(X) &= \arg\min_\mu \left\|(I - A)\,\tau^{-1}(X) - B\mu\right\|_p^p \\ \mu_\varepsilon(X) &= (I - A)\,\tau^{-1}(X) - B\mu_z(X), \end{aligned}$$

which would by construction minimize the p-norm $\|\mu_\varepsilon(X)\|_p$ of $\mu_\varepsilon(X)$ among all valid solutions. Note, however, that since both solutions result in exactly equal model scores, we have no principled reason to prefer one solution over the other. Therefore, we will assume $\mu_z = 0$ to ease computation.

For this process to learn well, the nonlinearity τ and its inverse τ^{-1} need to be simple. In particular, τ needs to have a closed-form solution for its inverse and allow for ease of gradient computation in the shared parameters θ. While tanh activation functions satisfy these conditions, years of research in neural network learning dynamics have established that rectified linear units (ReLU; Fukushima, 1975) show better convergence in practice (Krizhevsky et al., 2012). However, since ReLU activations are not invertible, we instead use parametric ReLU (PReLU) functions (Maas et al., 2013), whose inverse are again PReLU

functions, and which are defined as

$$\phi(x;\gamma) = \begin{cases} \gamma x & \text{if } x < 0\,, \\ x & \text{if } x \geq 0 \end{cases}$$
$$\phi^{-1}(x;\gamma) = \begin{cases} x/\gamma & \text{if } x < 0\,, \\ x & \text{if } x \geq 0\,. \end{cases}$$

To create a more expressive nonlinear function, we can then stack such PReLU functions ϕ to create our nonlinearities τ_i,

$$\tau_i(x_i;\theta_i) = \phi(\lambda_{i2}\phi(\lambda_{i1}x + \beta_{i1};\alpha_{i1}) + \beta_{i2};\alpha_{i2})$$
$$\text{and} \qquad \tau(x;\theta) = (\tau_1(x_1;\theta_1), \ldots, \tau_m(x_m;\theta_m))\ ,$$

where θ_i is the set containing all parameters $\{\lambda_{ij}, \alpha_{ij}, \beta_{ij}\}_{ij}$ (He et al., 2015a). While we could stack the PReLU functions ϕ to arbitrary depth, as a proof of concept for the viability of this approach, we stick with only two layers.

Next, to ensure that our learned model not only reconstructs the data well but also has a valid causal interpretation, we require the causal graph G parametrized by the matrices A and B to be acyclic. Fortunately, by construction, the variables Z have no incoming edges and, therefore, cannot contribute to any cyclic structures in the graph, leaving us to focus only on the matrix A. To this end, we follow the approach introduced by Zheng et al. (2018) to introduce a *differentiable* acyclicity constraint. To explain their approach, note that the matrix A^k counts precisely the number of directed weighted paths of length k in the graph G. More precisely, its entries are given by

$$(A^k)_{ij} = \sum_{p \in P_k(i,j)} \prod_{(u,v)\in p} a_{uv}\,,$$

where $P_k(i,j)$ is the set of all paths of length k leading from X_i to X_j, and the product runs over the edges constituting the path p. In particular, if all entries a_{ij} of A are non-negative, we have the following equivalence

$$i \text{ is not part of any cycle of length } k \longleftrightarrow (A^k)_{ii} = 0\,,$$

so that by summing up all of these non-negative entries, we obtain

$$\text{there are no cycles of length } k \longleftrightarrow \operatorname{trace}(A^k) = 0\,.$$

In practice, not all weights a_{ij} will be positive, but by taking the Hadamard

product $(A \odot A)_{ij} = a_{ij}^2$, this can be resolved in a differentiable manner. Hence,

$$G \text{ is acyclic } \longleftrightarrow \sum_{k=1}^{m} \lambda_k \operatorname{trace}((A \odot A)^k) = 0\,,$$

for *any* set of weights $\lambda_k > 0$. In particular, Zheng et al. (2018) used the weights $\lambda_k = 1/k!$, leading to the matrix exponential

$$\operatorname{trace}(\exp(A \odot A)) = \operatorname{trace}(I) + \sum_{k=1}^{m} \frac{1}{k!} \operatorname{trace}((A \odot A)^k)\,.$$

Since this algorithm is numerically unstable for large graphs, Yu et al. (2019) proposed to use the weights $\lambda_k = \binom{m}{i} m^{-i}$ instead, leading to the score

$$\begin{aligned} h(A) &\coloneqq \operatorname{trace}\left((I + A \odot A/m)^m\right) - m \\ &= \sum_{k=1}^{m} \binom{m}{k} \operatorname{trace}\left((A \odot A/m)^k\right), \end{aligned}$$

which is numerically more stable. This score $h(A)$ is fully differentiable in A and satisfies $h(A) = 0$ if and only if the graph parametrized by A is acyclic. For more details on these scores and other alternatives, see also Wei et al. (2020)

In general, we know not only that A is acyclic, but by assuming causal faithfulness, we know that the true DAG G^* is the sparsest of all DAGs consistent with the observed $P(X)$ (Raskutti and Uhler, 2018). As such, we are interested in finding *sparse* matrices A, B to capture the true causal DAG G^*. Under such a sparsity constraint, we show that our score is consistent for linear SCMs.

Theorem 4.3 (Consistency under Sparsity)**.** *Let x^n be a sample generated from the model in Equation* (4.4) *with $\tau = \mathrm{id}$ and let Assumptions A–D hold. Let L be the L^0-penalized ELBO score given by*

$$L(x^n; A, B) \coloneqq -L_{\mathrm{ELBO}} + \lambda_A \|A\|_0 + \lambda_B \|B\|_0\,,$$

and let $\widehat{A}, \widehat{B}$ be its minimizers subject to acyclicity, i.e.,

$$\begin{aligned} \widehat{A}, \widehat{B} = \operatorname*{arg\,min}_{A,B} & L(x^n; A, B) \\ \text{s.t. } & h(A) = 0\,. \end{aligned}$$

Then for sufficiently small $\sigma_\varepsilon(X), \sigma_z(X)$ the score L is consistent for recovering

the matrices A, B when $\lambda_A = \lambda_B = \log(n)/2$:

$$\lim_{n\to\infty} P(\widehat{A} = A,\, \widehat{B} = B) = 1\,.$$

Proof sketch. For the linear case and sufficiently small values of σ, the ELBO score essentially reduces to modeling the sample correlations between variables, so that this reduces to Theorem 3.5. □

In practice, the regularizers $\|\cdot\|_0$ are not differentiable, so we use the common practice of replacing them with L^1 norms, $\|\cdot\|_1$, instead, and describe a practical way of optimizing the resulting score next.

4.3.1 Optimization under Acyclicity

By replacing the sparsity penalty $\|\cdot\|_0$ of Theorem 4.3 with the more tractable $\|\cdot\|_1$, the overall learning problem, including the nonlinearities $\tau(\cdot;\theta)$ as well as the sparsity and acyclicity constraints is given by

$$\begin{aligned}\min_{A,B,\theta} f(x^n; A, B, \theta) &\coloneqq -L_{\text{ELBO}} + \lambda_A \|A\|_1 + \lambda_B \|B\|_1 \\ &\text{s.t. } h(A) = 0\,.\end{aligned}$$

To obtain a fully differentiable optimization target, we use the augmented Lagrangian approach for constrained optimization problems (Bertsekas, 1997)

$$\mathcal{L}(A, B, \theta, \lambda) = f(A, B, \theta) + \lambda h(A) + \frac{\rho}{2}\,|h(A)|^2\ .$$

Of course, to optimize this objective, we first need to compute L_{ELBO}. Since the expectation term is not fully analytically tractable, we can obtain estimates by using Monte Carlo approximations with K samples from the encoder $q(Z, \varepsilon \mid X)$

$$\begin{aligned}&E_{q(Z,\varepsilon|X)}\left[\log p(X \mid Z, \varepsilon)\right] \\ &\approx \frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{m} \frac{\left(X_i - \mu_x(Z_{(k)}, \varepsilon_{(k)})_i\right)^2}{2\sigma_x(Z_{(k)}, \varepsilon_{(k)})_i^2} \\ &\quad - 2\log\left(\sigma_x(Z_{(k)}, \varepsilon_{(k)})_i\right) - c\,,\end{aligned}$$

where $Z_{(k)}, \varepsilon_{(k)}$ is the k-th sample from the encoder $q(Z, \varepsilon \mid X)$. Next, since both $q(Z, \varepsilon \mid X)$ and $p(Z, \varepsilon)$ are Normal distributions, the KL divergence term

of the ELBO can be analytically computed as

$$\begin{aligned}\mathrm{KL}\left(q(Z,\varepsilon \mid X) \parallel p(Z,\varepsilon)\right) = \frac{1}{2}\Big(&- \log\det\left(\sigma_z^2 BB^\top + \sigma_\varepsilon^2(X)I\right) \\ &+ \operatorname{trace}\left(\sigma_z^2 BB^\top + \sigma_\varepsilon^2(X)I\right) \\ &+ \left\|(I-A)\tau^{-1}(X;\theta)\right\|_2^2\Big)\,.\end{aligned}$$

Writing $\Sigma_{Z,\varepsilon|X} = \sigma_z^2 BB^\top + \sigma_\varepsilon^2(X)I$, the augmented Lagrangian is therefore

$$\begin{aligned}&\mathcal{L}(A,B,\theta;\lambda,\rho) \\ &\approx \frac{1}{K}\sum_{k=1}^{K}\sum_{j=1}^{m}\frac{\left(X_j - \mu_x(Z_{(k)},\varepsilon_{(k)})_j\right)^2}{2\sigma_x(Z_{(k)},\varepsilon_{(k)})_j^2} - 2\log\left(\sigma_x(Z_{(k)},\varepsilon_{(k)})_j\right) \\ &\quad + \frac{1}{2}\left(-\log\det\left(\Sigma_{Z,\varepsilon|X}\right) + \operatorname{trace}(\Sigma_{Z,\varepsilon|X}) + \left\|(I-A)\tau^{-1}(X;\theta)\right\|_2^2\right) \\ &\quad + \lambda_A\|A\|_1 + \lambda_B\|B\|_1 + \lambda h(A) + \frac{\rho}{2}|h(A)|^2\,,\end{aligned}$$

and can be optimized using dual ascent (Bertsekas, 1997). Specifically, we use the common updating schemes for A, B, θ and parameters λ, ρ given by

$$\begin{aligned}A^k, B^k, \theta^k &= \operatorname*{arg\,min}_{A,B,\theta}\mathcal{L}(A,B,\theta;\lambda^k,\rho^k) \\ \lambda^{k+1} &= \lambda^k + \rho h(A^k) \\ \rho^{k+1} &= \begin{cases}\alpha\rho^k & \text{if } h(A^k) > \gamma h(A^{k-1}) \\ \rho^k & \text{otherwise}\,,\end{cases}\end{aligned}$$

where $\alpha > 1, \gamma < 1$ determine how quickly ρ increases (Yu et al., 2019).

Optimization of the first line can be done using any black box stochastic optimization algorithm readily available in machine learning toolboxes, such as Tensorflow (Abadi et al., 2016) or PyTorch (Paszke et al., 2019). For the other two equations, we use the values $\alpha = 10, \gamma = 0.25$ suggested by Yu et al. (2019) and found that the precise values did not significantly impact our results.

Note that the augmented Lagrangian $\mathcal{L}$ no longer forces the matrix A to be strictly acyclic. Following Zheng et al. (2018), we prune the weights a_{ij} in increasing order until the discovered graph G is acyclic.

Overall, we refer to our approach of learning the matrices A, B as NOCADILAC, short for **No**nlinear **Ca**usal **Di**scovery with **La**tent **C**onfounders.

4.4 From ReLUs to Related Work

While most classical causal discovery methods have been based on combinatorial optimization, these algorithms do not benefit from a large number of recent advances in automatic differentiation and differentiable programming (Abadi et al., 2016; Baydin et al., 2018; Blondel and Roulet, 2024).

One class of approaches to leveraging these advances has been the use of generative flows (Deleu et al., 2022; Li et al., 2022), which parametrize a flow over the space of DAGs on the data, and thereby find a locally optimal topological older over the observed variables.

A much larger class of approaches has been ushered in with the introduction of NOTEARS (Zheng et al., 2018), which reformulates network inference as a continuous optimization problem by introducing a differentiable constraint measuring how many cycles a matrix contains. While this approach was initially designed for purely linear relationships, it has been generalized to permit nonlinear relationships directly (Yu et al., 2019; Zheng et al., 2020). Furthermore, the proposed differentiable constraint has been used widely in a large number of other causal discovery methods, including autoregressive causal flows (Khemakhem et al., 2021; Monti et al., 2020), reinforcement learning with or without graph attention (Zhu et al., 2019; Yang et al., 2023), diffusion-based methods (Chao et al., 2023), and others (Lachapelle et al., 2019; Zhang et al., 2023a; Kertel and Klein, 2024; Bello et al., 2022; Waxman et al., 2024).

While this approach allows for the full suite of differentiable programming tools to be leveraged, it should be noted that when care is not taken, simply optimizing the objective leads often leads to trivial results. In particular, on simulated networks, due to the way that data generation processes are generally picked, it is *trivial* to obtain results comparable to the state of the art *simply by sorting the variables* in order of increasing variance (Chen et al., 2019; Reisach et al., 2021), and that this can be used to manipulate the inferred graph by adversarially rescaling some of the variables (Seng et al., 2022). It is, therefore, essential to account for this fact, for example, by normalizing the data. Furthermore, it has been shown that using the absolute value $|A|$ instead of the Hadamard product $A \odot A$, one can obtain better theoretical guarantees, albeit at the cost of differentiability around 0 entries of the matrix (Wei et al., 2020).

As described in Section 4.1.3, other approaches focus on identifiability of nonlinear causal networks by way of assumptions on the mixing functions in nonlinear ICA (Khemakhem et al., 2020; Gresele et al., 2021; Zheng et al., 2022).

However, as with more classical approaches, all of the methods above rely heavily on the assumption of causal sufficiency. Only more recently have methods been proposed to either estimate causal effects under latent confounding (Soleymani et al., 2020; Hu et al., 2021), as well as diffusion-based methods for causal discovery under confounding (Shimizu, 2023). More interesting to us is

again the DCD method and its nonlinear extension (Bhattacharya et al., 2021; Ashman et al., 2023), which combine NMMs with the differentiable constraint by Zheng et al. (2018) to discover a partially directed causal network indicating which nodes are likely confounded. As we have seen in the previous chapter, the resulting networks are nevertheless still difficult to interpret since they do not model the confounder directly and, therefore, cannot determine if a set of variables shares the same latent confounder.

4.5 Nonsense or Nuance: Experiments

In this section, we evaluate our method NOCADILAC empirically. As in the previous chapter, we are interested in how well it recovers the set of nodes affected by each Z_j and how well it recovers the entire causal network. We compare with other state-of-the-art methods for network discovery, including those that permit latent confounding, GFCI (Ogarrio et al., 2016), 3OFF2 (Affeldt et al., 2016), and DCD (Bhattacharya et al., 2021), and those assuming causal sufficiency, NOTEARS (Zheng et al., 2018) and DAG-GNN (Yu et al., 2019). We implemented NOCADILAC in Python using Tensorflow (Abadi et al., 2016) and perform optimization using Adam (Kingma and Ba, 2014). For our competitors, we use the same implementations for GFCI, 3OFF2, DCD, and NOTEARS as in the previous chapter and use the implementation provided by the authors for DAG-GNN (Yu et al., 2019). We make all code and data available online.[1]

4.5.1 Evaluation on Synthetic Data

We begin by giving an overview of our data generation process. As before, we generate a random DAG from the Erdős-Rényi model with $p = 0.3$. We then generate a corresponding adjacency matrix with $A_{ji} \sim U[-1, 1]$ when $(i, j) \in G$ and $A_{ji} = 0$ otherwise. Our causal generating model is given by $v = Ag(v)+\varepsilon$ where $g(v) = \mu+\alpha\odot f(v)$ where f is an element-wise function with entries uniformly sampled from linear, quadratic, cubic, exponential, logistic or sinusoidal functions, and $(\mu, \alpha) \sim U[-1, 1]^{2m}$ are i.i.d. The noise is $\varepsilon \sim N(0, 1)$. To generate the observed x, we sample v from the above model and remove some source nodes $z = (v_{k_1}, \ldots, v_{k_l})$. We consider first the performance when the confounder is univariate, and study the effects of varying dimensionality of Z later. For each experiment, we used $n = 2500$ samples and 500 repetitions. We evaluate each method as in the previous chapter using F_1^{net}, F_1^{conf}, SHD, and SID to measure its performance on different aspects. Furthermore, as explained in Section 2.6, we use decision rate plots to show the relationship between each method's performance and confidence. The confidence for NOCADILAC

[1] `https://eda.rg.cispa.io/prj/nocadilac/`

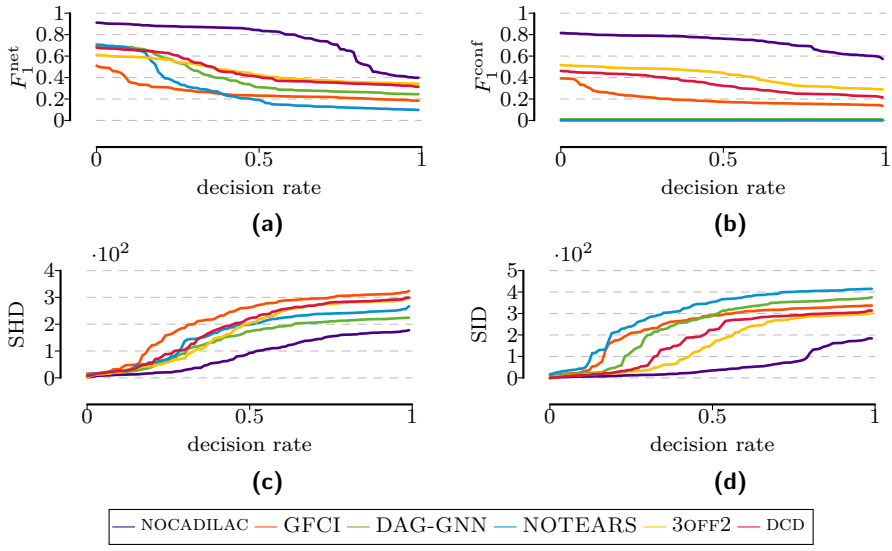


Figure 4.4: *Evaluation on synthetic networks of size* 25. We show F_1 scores for (a) network discovery and (b) confounded set recovery (higher is better), as well as (c) structural Hamming distance and (d) structural intervention distance (lower is better). Each figure shows the average score over increasing fractions of all datasets, sorted for each method from most confident to least. We see that NOCADILAC outperforms its competitors at all levels. In particular, in confounded set recovery (b), NOCADILAC performs better *at its worst* than its competitors do *at their best*.

is computed as in the previous chapter, using the relative improvement by including confounders over the unconfounded model

$$C = \frac{\mathcal{L}(A, B = 0, \theta, \lambda) - \mathcal{L}(A, B, \theta^*, \lambda)}{\mathcal{L}(A, B = 0, \theta, \lambda)} ,$$

whereas for all other methods, we use the same method to compute their confidence as in the previous chapter. That is, for DAG-GNN, NOTEARS, DCD, we compare their scores against the scores under the null model, while for GFCI and 3OFF2, we order their decisions in the way most favorable to them.

Confidence and Performance

We now study the performance of NOCADILAC compared to its competitors on networks of $m = 25$ observed variables. We show the decision rate plots in Figure 4.4. For NOCADILAC, all metrics are monotonic in the confidence C, suggesting that it can be used to determine which network inferences are more

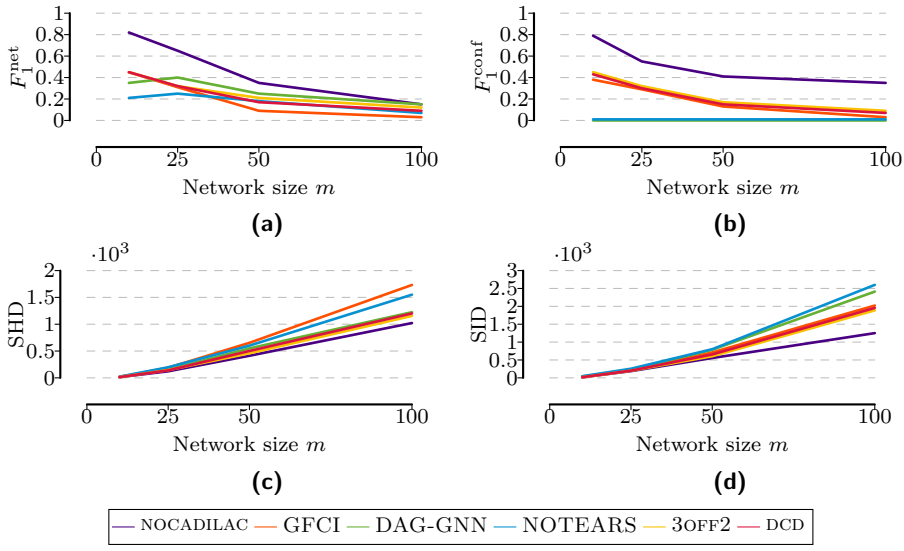


Figure 4.5: *Evaluation on synthetic networks of different sizes.* We show F_1 scores for (a) network discovery and (b) confounded set recovery (higher is better), as well as (c) structural Hamming distance and (d) structural intervention distance (lower is better).

reliable than others. For NOTEARS and DAG-GNN, their confidence correlates with F_1^{net}, SID, and SHD, but since they are not designed to find confounders, their F_1^{conf} score is zero throughout. Overall, NOCADILAC outperforms its competitors by a large margin for all metrics at almost all levels of confidence.

Performance for different network sizes

To see how well NOCADILAC performs for larger networks, we next test all methods for networks of sizes $m \in \{10, 25, 50, 100\}$ and show the results in Figure 4.5. We see that NOCADILAC outperforms its competitors by a large margin across the board for almost all network sizes. While all methods show decreasing performance for increasing network size m, NOCADILAC performs best for increasing network sizes. As the network size increases, the gap becomes smaller for the F_1 scores. This is due primarily to the latent confounder affecting only a smaller fraction of variables as we increase the network size. Meanwhile, especially for SID, the gap becomes more prominent as every incorrect edge between nodes simultaneously affects many other pairs of nodes.

Method	Number of confounders 1	2	3	4	5
NoCa-DiLaC	**0.35**	**0.33**	**0.29**	**0.26**	**0.19**
DCD	0.23	0.14	0.11	0.07	0.03
3off2	0.22	0.12	0.08	0.05	0.03
GFCI	0.15	0.07	0.03	0.02	0.01

Table 4.1: *Comparison on graphs with varying numbers of latent confounders.* While all methods perform well for a one-dimensional confounder, only NOCADILAC maintains its performance as the number of latent factors is increased.

Higher-dimensional Z

To study the effect of including multiple confounders Z_i, we show the overall F_1^{conf} scores for one to five confounders in Table 4.1, with $m = 50$ observed variables. As in the previous chapter, we omit NOTEARS and DAG-GNN since they cannot discover confounded nodes. We see that NOCADILAC performs at a consistent level for one to four confounders, with a slightly larger drop in performance for five latent variables. In contrast, all of DCD, 3OFF2, and GFCI perform less well from the start but also display a dramatic drop in performance upon adding a second latent confounder. The reason for this is as in the previous chapter: by modeling only pairwise confounding, they cannot determine which variables share the same confounder. This remains true even when we try to cluster nodes using spectral clustering of confounded variables.

4.5.2 REGED Benchmark Data

We again use the REGED data introduced in Section 3.5.2 and use the same setup to obtain confounded and unconfounded subsets of data.
We show decision rate plots for all metrics in Figure 4.6. While all methods perform worse than on synthetic data, NOCADILAC nevertheless performs best by a large margin, and the overall pattern of relative performances between methods is just as it was for the synthetic data above.

4.5.3 Application to a Protein Signaling Network

Last, to see how well NOCADILAC works on real-world data, we evaluate it on the widely used Sachs dataset (Sachs et al., 2005) for protein signaling. It contains $n = 7466$ continuous measurements for $m = 11$ phosphorylated proteins and phospholipids in human immune system cells. The consensus network contains 20 edges, which we show in Figure 4.7a. Since the graph contains cycles, some mistakes are inevitable. To make the data appropriate for our setting, we remove the node PKC with out-degree four from the network. Note that the edge from PIP2 to PKC violates our assumption that latent confounders have

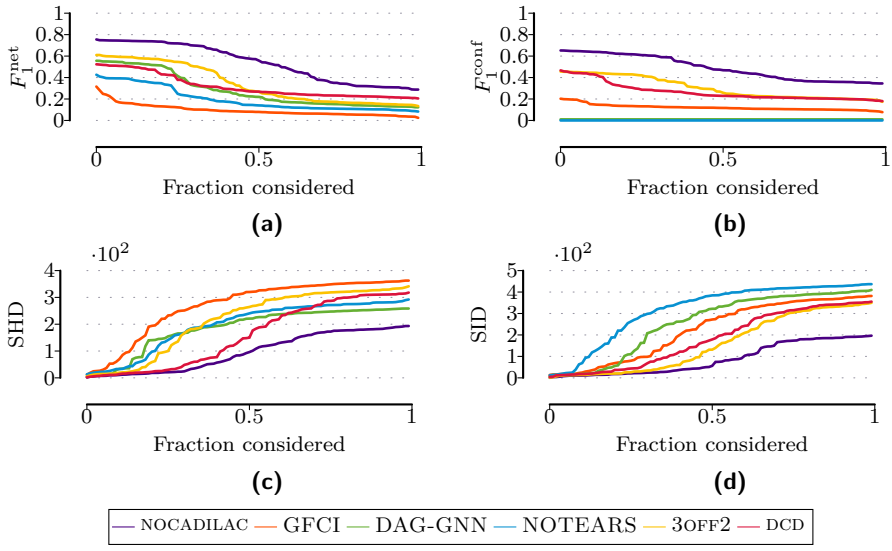
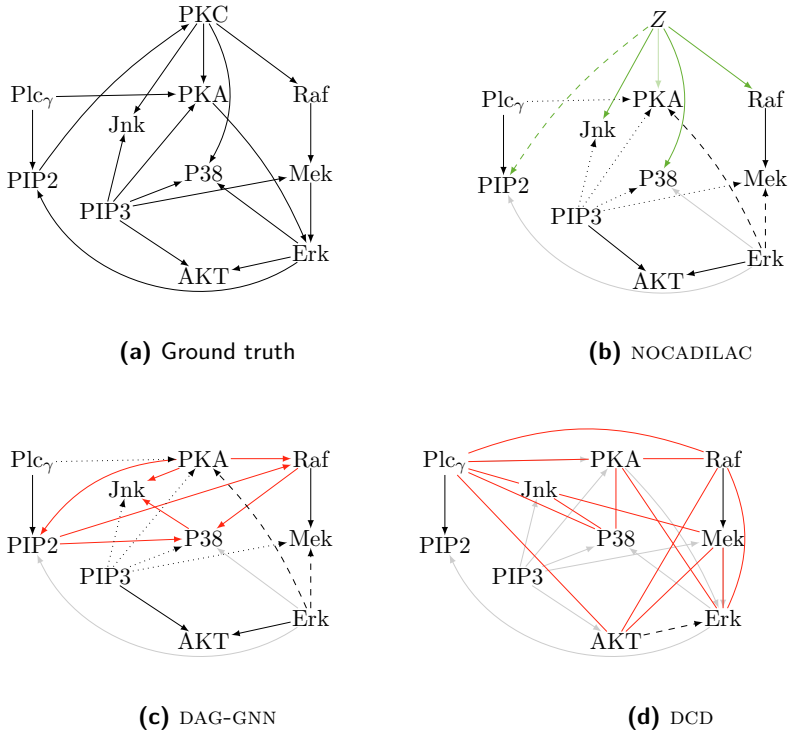


Figure 4.6: *Evaluation on REGED.* We show F_1 scores for (a) network discovery and (b) confounded set recovery (higher is better), as well as (c) structural Hamming distance and (d) structural intervention distance (lower is better).

no incoming edges. We show the benefit of explicitly modeling confounders by comparing NOCADILAC with DAG-GNN and DCD.

We show the results of this experiment in Figure 4.7. In Figure 4.7b, we see that NOCADILAC automatically discovers a substitute latent factor Z connected to the correct variables (green edges) and thereby takes the place of PKC. For the overall network, only three edges are missing entirely (gray), while two are reversed (dashed), and another three are instead contained as paths of length 2 (dotted). This performance is similar to the result of other state-of-the-art methods on *fully* observed data (Yu et al., 2019). In Figure 4.7c, we see the result of DAG-GNN. The absence of PKC from the observed data leads to DAG-GNN inferring a large number of edges (red) between nodes initially connected to PKC while making more mistakes over the remaining variables. We see a similar pattern in the results of DCD in Figure 4.7d. Many pairs of children of PKC are considered to be potentially confounded (red), but so are many other pairs of variables that do *not* have PKC as a parent. It is also unclear which pairs share the *same* latent parent. Furthermore, DCD misses many more edges than either NOCADILAC or DAG-GNN. Overall, NOCADILAC produces more accurate and interpretable results than its competitors.



(a) Ground truth **(b)** NOCADILAC

(c) DAG-GNN **(d)** DCD

Figure 4.7: *Results on the Sachs dataset.* NOCADILAC discovers a confounder Z capturing the influence of PKC (green edges). In contrast, DAG-GNN finds many edges between nodes influenced by PKC (red) and DCD contains indications of confounding for many pairs of nodes, but neither method can determine that nodes share a confounder. All methods make roughly the same number of errors regarding reversed edges (dashed), including some edges only as indirect paths (dotted), and missing edges (gray).

4.6 Between Confounding Charybdis and Selective Scylla

In this chapter, we tackled Problem 2b) by building on the theory of identifiability under a suitable type of sparsity we developed in Chapter 3. In particular, we developed a framework for recovering the causal network, including latent variables Z, when the causal mechanisms are not purely linear.

By choosing our causal model to be a post-nonlinear model, we related our problem to the theory of nonlinear ICA, highlighting the difficulties of the problem, and showing that recent advances in identifiability for nonlinear ICA are not sufficient for solving the task we are interested in. While ideas of sparsity and algorithmic independence have been proposed for this task of nonlinear ICA, it is unclear under which conditions *overcomplete* nonlinear ICA

allows for identifiability, which is necessary to deal with latent confounding.
To tackle this problem, we leveraged our insights from the previous chapter about both the graphical structure entailed by latent confounding, as well as the benefits of a certain type of sparsity. We further leveraged the structure of the PNL model to generalize the results we derived there to this setting. By adapting the ELBO score, we obtained a consistent score for recovering the true causal network in the linear setting. Furthermore, by leveraging recent approaches to learning causal networks in a fully differentiable manner, we can optimize this ELBO score by adapting the commonly used VAE architecture to encode our exogenous noise and confounding variables separately.
In several experiments we showed that NOCADILAC works well in practice on synthetic and real-world data. In particular, as with CDHC, we saw that its confidence tracks its performance well and that it returned more readily interpretable outputs than competing methods.
For future work, we are interested in improved theoretical guarantees. As discussed in the previous chapter, the gap between how many parameters we can currently deal with and how many we *should* be able to deal with appears to be large. For NOCADILAC, however, another difference is that we currently cannot prove identifiability of the nonlinearity τ. While recovering τ is not necessary to recover the causal structure G^*, it is nevertheless of interest for the purpose of end-to-end learning of causal estimates.
Furthermore, another highly relevant avenue would be to provide theoretical results for the more commonly used class of additive noise models (ANMs). We can, in fact, write our PNL model in the equivalent form

$$X = \nu(X) + BZ + \varepsilon\,,$$

where the function ν is given by $\nu = \mathrm{id} - \big(\tau \circ (I - A)^{-1}\big)^{-1}$. Unlike τ, the nonlinearity ν is not an element-wise transformation of the vector X. By rewriting the PNL in this manner, all our identifiability results apply to this specific kind of ANM with latent confounding. This approach might, therefore, provide us with a path towards extending our identifiability results to general ANMs with less restrictive forms on the nonlinearity ν in the future.
Nonetheless, in the next chapter we turn instead to a different major potential source of bias in causal learning: selection bias.

Chapter 5

Identifying Selection Bias from Observational Data

> *"It was a good answer that was made by one who when they showed him hanging in a temple a picture of those who had paid their vows as having escaped shipwreck, and would have him say whether he did not now acknowledge the power of the gods, — 'Aye,' asked he again, 'but where are they painted that were drowned after their vows?' And such is the way of all superstition, whether in astrology, dreams, omens, divine judgments, or the like; wherein men, having a delight in such vanities, mark the events where they are fulfilled, but where they fail, though this happens much oftener, neglect and pass them by."*
>
> Novum Organon, Francis Bacon

In the previous chapters, we have considered the effects of latent confounding on causal discovery and inference. Another related problem leading to biased estimates of causal effects is that of *selection bias*. Selection bias is due to preferential inclusion of some subjects over others based on unknown factors causally downstream of the observed variables (Bareinboim et al., 2014).
As an example, consider the study by Kovács and Sharkey (2014) on Goodreads book ratings. Employing a regression discontinuity design in a sample of 32 books, they studied the effects of winning an award on book ratings by comparing books that won an award and matched runner-ups that did not. They found that the average book ratings declined after winning an award (see Figure 5.1). This pattern can be explained by the fact that there are two kinds of readers.

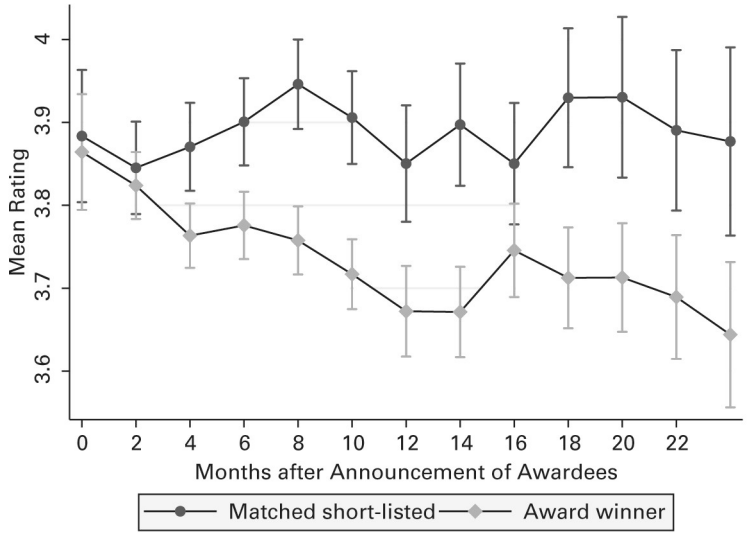


Figure 5.1: *Goodreads ratings for books that won an award (gray) and similar books that did not (black) from Kovács and Sharkey (2014).* The decline of the ratings for the former is explained by an extended readership that is less predisposed to like the book.

Those who read a book before it won an award did so because they were likely looking for books within a given genre and were, therefore, predisposed to like it and thus more likely to give the book a good rating. Meanwhile, those who read a book after it won an award were not predisposed to like it and are likely to be more representative of the population as a whole. This kind of selection is referred to as *self-selection* because subjects themselves join groups of readers based on their literary (or other) preferences.

Similar and less benign issues occur everywhere in empirical sciences, such as case-control studies in epidemiology (Glymour and Greenland, 2008), studies using hospital-admission data (Berkson, 1946; Herbert et al., 2020), genetics (Mefford and Witte, 2012), economics (Angrist, 1997), statistics (Kuroki and Cai, 2006), psychology (Kaźmierczak et al., 2023), and many more.

In machine learning systems, too, predictive performance suffers from selection bias. When training samples are collected preferentially from some subpopulation, such as those who self-select into feedback programs, or Mechanical Turk workers, or data from some hospitals but not others, covariate shift occurs between training and test samples, leading to worse performance on the test data (Bickel et al., 2009). Even when data from multiple contexts is collected, such as multiple hospitals, care must be taken as naive combination of these

datasets can lead to degradation of performance (Compton et al., 2023). Methods that adjust for such covariate shift (Sugiyama et al., 2007; Gretton et al., 2009; Mallick et al., 2022) rely on the availability of independent training and test datasets, in which the available test data is assumed to be *unbiased*. This makes them unsuited for determining whether selection bias might be affecting our data when we only have access to a single dataset.

Therefore, in this chapter, we study conditions under which selection bias is identifiable from only a single biased dataset. We show that when the selection effect is linear, the underlying distribution $P(X)$ can be recovered in two cases,

a) parametrically, when $P = P_\theta$ lies in a known exponential family, e.g., when $P = N(\mu, \Sigma)$ is a Normal distribution (Section 5.2.1), and
b) non-parametrically, when P is known to satisfy certain invariances, such that for some map j we have $P = P \circ j$ For example rotational invariance, $P = P \circ U$, for a standard t-distribution (Section 5.2.2).

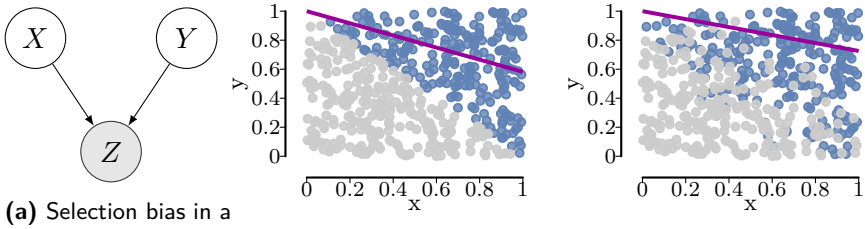
More specifically, we provide results for exponential families when selection is a deterministic function of observed covariates and for the normal family when selection is influenced by additive Gaussian noise. For the non-parametric case, we show identifiability under the general assumption that the underlying ground truth distribution is subject to some set of invariances, such as the above-mentioned rotational invariances or more general measure-preserving automorphisms mentioned in Section 4.1.3.

Based on these theoretical results, we propose two practical methods to tell, based on a single dataset, whether this data is subject to selection bias, as well as how strong this bias is. By studying the behavior of a distribution conditioned on selection bias, for parametric families we motivate the use of alternate optimization between the parameters of the distribution and the selection boundary (Section 5.3.1). In contrast, for non-parametric families with a known set of potential (rotational) invariances, we propose to learn an invariance of the distribution and use violations of the invariance to discover which parts of the distribution may be subject to selection bias (Section 5.3.2)

Last, through an extensive set of experiments, including case studies on penguin and exoplanet data, we show that our methods can provide valuable and novel insight and significantly outperform baselines that try to control for distribution shifts or latent confounding factors. As with previous chapters, proofs for all theoretical statements are included in Appendix A.5.

5.1 Setting up for Selection

As before, we denote observed variables by X. Then, *selection bias* is the act of conditioning on an unobserved variable Z causally downstream of X. This causes a shift from the population distribution $P(X)$ to a distribution $P(X \mid Z)$, resulting in potentially false inferences. We consider the most general case,



(a) Selection bias in a graphical model.

(b) Noiseless selection.

(c) Noisy selection.

Figure 5.2: *Selection bias.* Left: A graphical representation of selection bias as the act of conditioning on a common child Z of multiple observed variables. Middle: The effect of selection on independently uniformly sampled points. Blue points are observed, gray points are excluded due to selection. While X and Y are originally independent, they are negatively correlated in the observed sample, as shown by the regression line (purple). Right: Similar to (b) but the selection is noisy. Included and excluded points are no longer nicely separated, making it more difficult to notice the effect of selection.

where $Z = f(X, \varepsilon)$ can be a function of any or all variables in X and some independent noise ε. We are interested in determining whether a given set of observations x has been sampled from the true distribution $P(X)$ or from a biased distribution $Q(X) = P(X \mid f(X, \varepsilon) \in S)$.

As an example, consider the setup in Figure 5.2. In Figure 5.2(a), we depict the graphical structure of selection bias. In contrast to the case of confounding, in which Z is *causally upstream* of X, Y and not conditioned on, in the case of selection bias, it is *causally downstream* and conditioned on. Specifically, in Figure 5.2(b), we let $P(X, Y) = U[0, 1]^2$, making them independent. We then select samples satisfying $X + Y > 1$, leaving only the data in the top right. We see clearly that there is a spurious negative correlation between X and Y in the distribution $Q(X, Y) = P(X, Y \mid X + Y > 1)$. We call this type of selection above *noiseless* because it depends solely on X, Y but no external source of noise ε. Selection can also be *noisy*, as in the case $X + Y + \varepsilon > 1$ in Figure 5.2(c), where $\varepsilon \sim N(0, 0.05)$ is a small amount of noise. Clearly, the presence of noise makes it more difficult to determine whether selection is occurring and, if so, which parts of the distribution $P(X)$ are affected by it.

In this chapter, we consider linear selection, $Z = f(X, \varepsilon) = a^\top X + \varepsilon$, where we include a variable $X_0 = 1$ to account for affine offsets. Our observed sample x therefore comes from the distribution $Q_a = P(\cdot \mid a^\top X + \varepsilon > 0)$ where $a \in \mathcal{A}$ is unknown. When ε is symmetric around the origin, then $a^\top = 0$ corresponds precisely to the case of no selection bias. Since the condition of $a^\top X + \varepsilon > 0$ is invariant under multiplication with a scalar, we further assume that each non-zero vector $a \in \mathcal{A}$ is normalized in some way, e.g., via $a_1 = 1$ or $\|a\|_2 = 1$. We aim to recover a and P from a sample $x \sim Q_a$. Next, we outline the theoretical

underpinnings of two different approaches to recovering these.

5.2 Of Identifiability and Invariances

In this section, we first show that linear noiseless selection effects are always identifiable for exponential families and noisy selection with Gaussian noise is identifiable for the family of normal distributions. We then show that for non-parametric families of distributions, noiseless linear selection is identifiable under assumptions on the set of potential invariances of $P(X)$.

5.2.1 Identifying Selection Bias in Parametric Models

We assume that our distribution $P(X)$ is described by a parametric distribution P_θ belonging to an exponential family $\mathcal{M}$ parametrized by Θ, with density

$$p_\theta(x) = h(x) \exp(\eta(\theta)^\top T(x) - A(\theta)),$$

where $\eta(\theta)$ are called the natural parameters, $T(x)$ the sufficient statistic and $A(\theta)$ the log-partition function. Further, all P_θ share the same support, $\operatorname{supp}(P_\theta) \coloneqq \{x : p_\theta(x) > 0\}$, which is independent of θ. Many, but not all, common parametric families of distributions are exponential families.

Example 5.1. We give an example of exponential families and classes of distributions which are not exponential families.

a) A typical exponential family is the multivariate normal distribution $N(\mu, \Sigma)$. Its natural parameters, sufficient statistics, and log-partition function are

$$\begin{aligned} \eta(\mu, \Sigma) &= \begin{pmatrix} \Sigma^{-1}\mu \\ -\frac{1}{2}\Sigma^{-1} \end{pmatrix} \\ T(x) &= \begin{pmatrix} x \\ xx^\top \end{pmatrix} \\ A(\mu, \Sigma) &= \frac{1}{2}\mu^\top \Sigma^{-1}\mu + \frac{1}{2}\log\det(\Sigma)\,. \end{aligned}$$

b) The uniform distributions $U[0, \theta]$ on different intervals $[0, \theta]$ do *not* form an exponential family because their supports are not equal.

c) The parametric family of Student's t-distributions, $t_\nu(\mu, \Sigma)$ does *not* constitute an exponential family because its parametrization cannot be factorized in the required manner.

d) In fact, the t-distribution is a special case of the general class of (uncountable) Gaussian mixture distributions, which are generically not exponential families. Other examples include non-centered Laplace distributions.

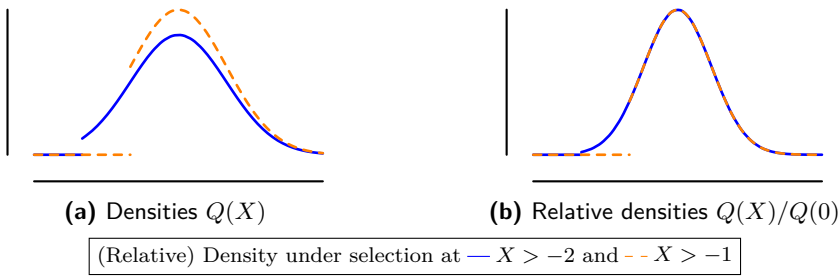


Figure 5.3: *Density ratios under selection.* While the densities (a) for the same distribution affected by different selection criteria are different in the unselected part, the *density ratios* (b) are the same at every point that is left unaffected by the selection mechanism.

When the distribution P_θ is affected by selection bias parametrized by the selection vector $a \in \mathcal{A}$, we obtain the distribution

$$Q_{\theta,a} \coloneqq P_\theta(\cdot \mid a^\top X + \varepsilon > 0)\,,$$

and form the model class $\mathcal{M}_s = \{Q_{\theta,a} : \theta \in \Theta, a \in \mathcal{A}\}$ of models with selection. Note that even if $\mathcal{M}$ is an exponential family, $\mathcal{M}_s$ no longer forms an exponential family since its members do not share the same support.
Next, we move on to identifiability results for the parameters of the distributions $Q_{\theta,a} \in \mathcal{M}_s$. To see why this *should* be possible, we consider the behavior of the distribution $Q_{\theta,a}$ "away" from the (noiseless) selection boundary as follows. Let x, y be two points such that $a^\top x > 0$ and $a^\top y > 0$. Then we have

$$\begin{aligned} \log \frac{Q_{\theta,a}(x)}{Q_{\theta,a}(y)} &= \log \frac{P_\theta(x)/P_\theta(a^\top X > 0)}{P_\theta(y)/P_\theta(a^\top X > 0)} \\ &= \log \frac{P_\theta(x)}{P_\theta(y)} \\ &= \log \frac{h(x)}{h(y)} + \eta^\top(\theta)\,(T(x) - T(y))\,, \end{aligned} \tag{5.1}$$

which tells us that the "remainder" of the distribution P_θ, the part unaffected by the selection mechanism, does not depend on a. See Figure 5.3 for a graphical depiction of this relationship between the density ratios. In particular, note that this is (almost) a linear system in $\eta(\theta)$. As such, if $\eta(\theta)$ is r-dimensional, we expect $r+1$ such density ratios to suffice to recover θ uniquely. The following theorem tells us that this intuition is essentially correct.

Theorem 5.1 (Identifiability under Noiseless Selection for Exponential Families)**.** *Let $\mathcal{M}$ be an exponential family with parameter space Θ and sufficient*

statistics $T(x)$ non-constant on every half-space. Further, let $\mathcal{A}$ be the set of (normalized) selection vectors a such that

$$0 < P_\theta(a^\top X > 0) < 1\,, \tag{5.2}$$

for all θ.[1] *Then the parameters (θ, a) of $Q_{\theta,a}$ are identifiable. In particular, P_θ is fully determined by the distribution $Q_{\theta,a}$.*

Proof sketch. For a, θ and a', θ', from Equation (5.2) we obtain that $a = a'$, and thus from Equation (5.1) that $\theta = \theta'$. □

The assumption of Equation (5.2) that $0 < P_\theta(a^\top X > 0) < 1$ is both natural and necessary. First, if $P_\theta(a^\top X > 0) = 0$ then there is nothing left of P_θ, and the distribution $Q_{\theta,a}$ is not well-defined to begin with. Conversely, if $P_\theta(a^\top X > 0) = 1$, then no selection occurs, and the same would be true for any $a'_0 > a_0$, making the parameter a unidentifiable.
The intuition we developed above also applies approximately in the case of noisy selection. However, when studying the density ratios, one must also study deviations from equality, which depend on the distribution of the noise ε and the distance from the selection boundary. Due to the difficulty of doing this rigorously, we will next prove identifiability in the special case of noisy selection in the Gaussian exponential family.

Theorem 5.2 (Identifiability of Noisy Selection Effects in the Gaussian Family)**.** *Let $\mathcal{M}$ be the Gaussian exponential family with parameter space $\Theta = \{(\mu, \Sigma)\}$ and let $\varepsilon \sim N(0, 1)$. Further, let the biased distribution be*

$$Q_{\mu,\Sigma,a,\zeta}(X) = P_{\mu,\Sigma}(X \mid a^\top X + \zeta\varepsilon > 0)\,.$$

Then the parameters $(\mu, \Sigma) \in \Theta, a \in \mathcal{A}, \zeta > 0$ are jointly identifiable.

Proof sketch. This is simple but annoying algebra. □

The assumption of Equation (5.2) is not necessary here because $\operatorname{supp}(P_\theta) = \mathbb{R}^m$. Next, we consider the case where our distributions are no longer necessarily of a known parametric form but instead satisfy another regularity condition.

5.2.2 Beyond parametric distributions: Invariance

Assume that we have data either from a normal distribution $N(\mu, \Sigma)$ or from a t-distribution $t_\nu(\mu, \Sigma)$ with ν degrees of freedom. Then, while we do not know the model class from which our data comes, we nevertheless know one crucial fact about the underlying distribution: it is the same after reflection across

[1] Since all P_θ share a support, if the assumption holds for any θ, it holds for all θ.

its mean, i.e., X and $-X + 2\mu$ have the same distribution. We call this an *invariance* of P, similar to conservation laws in physics, or measure-preserving automorphisms (MPAs) in the ICA literature (Hyvärinen and Pajunen, 1999).

Definition 5.1 (Invariance)**.** Let P be a probability distribution and j be a measurable bijective function. We say that j is an invariance for the distribution P if $P(j(A)) = P(A)$ for all measurable sets A.

For a spherical normal distribution $N(0, \sigma^2 I)$, the density $p(x)$ depends on x only through its distance from the origin $\|x\|_2^2$, so that any function j leaving this norm the same would form an invariance of P. Therefore, a natural set of invariances for $N(0, \sigma^2 I)$ would be the set orthogonal matrices, $O(m)$. The function j can be arbitrarily complex and fine-tuned to P. For example, despite its lack of obvious symmetry, the 1-dimensional exponential distribution $\text{Exp}(\lambda)$ has an invariance $t \mapsto -\log(1 - e^{-\lambda t})/\lambda$, which is simply the mapping of its q-quantile to its $(1 - q)$-quantile. In fact, by Sklar's theorem (Sklar, 1959; Jaworski et al., 2010), every multi-variate distribution $P(X)$ with connected support has an invariance group generated by mapping the distribution first to a Gaussian distribution and rotating that distribution. These invariances correspond precisely to the MPAs we showed in Figure 4.1 in the previous chapter as one common non-identifiability in nonlinear ICA.
If P has a density p then for any two x_0, x_1 it has trivial invariances $j(x_i) = x_{1-i}$ and $j(x) = x$ everywhere else. Such an invariance is trivial because it differs from the identity only on a set of measure zero. To preclude such degenerate cases, we consider what we call *strongly distinguishable invariances.*

Definition 5.2 (Strongly distinguishable invariances)**.** We call a set J of invariances strongly distinguishable for the distribution P if for all $j, j' \in J$ we have $P(j(x) = j'(x)) > 0$ if and only if $j = j'$.

Consider, for example, the normally distributed $X \sim P = N(0, \sigma^2 I)$. It is invariant under the group of orthogonal matrices $O(m)$. Further, for any two $U \neq U' \in O(m)$, the set $K = \ker(U - U') = \{x : Ux = U'x\}$ is a linear subspace of $\mathbb{R}^m$ with $\dim(K) < m$ so that $P(UX = U'X) = 0$. Therefore, the set of orthogonal matrices U is strongly distinguishable for any spherical normal distribution $N(0, \sigma^2 I)$. Furthermore, any invariance derived on the basis of these, such as the invariance groups based on MPAs derived from mapping first to a Gaussian distribution and then multiplying with an orthogonal matrix (see Section 4.1.3), are also strongly distinguishable for any given P.
The importance of these invariances is that we can use them to detect selection bias. If P is invariant under j, then selection bias will break such an invariance.

For example, when we discard all samples to the left of -1 of a standard normal distribution, the invariance $j : x \mapsto -x$ no longer applies to the new

(a) Symmetry without selection. **(b)** Partial symmetry with selection.

Figure 5.4: *Broken symmetries.* Where in the full distribution P, the symmetry $x \mapsto -x$ applies to the entire range (a), in the presence of selection bias, those sets which are affected by it (red) no longer respect the symmetry (b). Estimating which subsets do, and which do not, gives us information about which parts of P are affected by selection.

distribution Q. However, it still applies to *large parts* of Q, indicating that we can nevertheless obtain helpful information about P. Furthermore, once we have found a good candidate for an invariance of P, we can consider those regions where it does *not* apply. These are precisely where selection bias is likely to be at work! See Figure 5.4 for a graphical depiction of the idea. We formalize this intuition in the following theorem.

Theorem 5.3 (Identifiability of Selection under Invariance)**.** *Let $\mathcal{M}$ be a set of probability distributions and J be strongly distinguishable for each $P \in \mathcal{M}$. Assume that for all $P \in \mathcal{M}$ there is $j \neq \mathrm{id} \in J$ such that $P(X) = P(j(X))$. Let $P \in \mathcal{M}$ and $\mathcal{A}$ be the set of a for which there exists $j \in J$ such that*

$$\left\{a^\top X > 0\right\} \cap j^{-1}\left(\left\{a^\top X < 0\right\}\right) \neq \emptyset\,.$$

Then a is identifiable. Further, if all distributions $P_1, P_2 \in \mathcal{M}$ satisfy $P_1(\cdot \mid a^\top X > 0) = P_2(\cdot \mid a^\top X > 0)$ iff $P_1 = P_2$ then P is identifiable too.

Proof sketch. For an invariance j of P for which the intersection $\left\{a^\top X > 0\right\} \cap j^{-1}(\left\{a^\top X < 0\right\}) \neq \emptyset$ is not empty, we can use this set to determine a by comparing $j\left(\left\{a^\top X > 0\right\} \cap j^{-1}(\left\{a^\top X < 0\right\})\right)$ with $\left\{a^\top X > 0\right\}$. □

The last assumption is true for many classes of distributions. In particular, it holds for all exponential families, unions of multiple exponential families, arbitrary finite mixtures of exponential families, and stationary Gaussian processes (Bishop and Nasrabadi, 2006). Next, we develop methods to find selection effects for both exponential families as well as invariant distributions.

5.3 Manifesting Methods for Making Selection Manifest

In this section, we develop two complementary approaches to discovering selection bias in observational data based on the theoretical results we developed.

The first directly fits an exponential family with selection bias to the data, which we refer to as EXP. The second finds an approximate invariance of the true distribution and derives the selection boundary from it, referred to as INV.

5.3.1 Finding Selection Effects in Exponential Families

When we know that the data x comes from a given exponential family subject to selection bias, based on the intuition we have developed in Equation (5.1), the true parameters (θ^*, a^*) should be such that when we are given either of them, the other can be found by optimizing for it while keeping the other fixed. Consequently, we should be able to recover both the parameter θ and the selection boundary a by alternately optimizing the two parameters. We begin by writing the sample log-likelihood under $Q_{\theta,a}$ in the form

$$\begin{aligned} l_{\theta,a}(x) &\coloneqq \log q_{\theta,a}(x) \\ &= \left(\sum_{k=1}^{n} \log p_\theta(x_k)\right) - n \log p_\theta(a^\top X > 0)\,. \end{aligned}$$

Then, starting from a random initialization θ_0, a_0, instead of doing full optimization at each step, we update our parameters via

$$\begin{aligned} \theta_{t+1} &\leftarrow \theta_t + \lambda_\theta \frac{\partial}{\partial \theta} l_{\theta_t, a_t}(x) \\ a_{t+1} &\leftarrow a_t + \lambda_a \frac{\partial}{\partial a} l_{\theta_{t+1}, a_t}(x)\,, \end{aligned} \tag{5.3}$$

with step sizes $\lambda_\theta, \lambda_a$. As with exponential families in general, we cannot provide convergence guarantees for our approach. However, we can provide a partial characterization of the saddle points of this optimization.

Proposition 5.4. *Let x be a sample from the distribution $Q_{\theta,a}(X)$ with known underlying exponential family $\mathcal{M}$. Then, the saddle points $(\widehat{\theta}, \widehat{a})$ of Equation (5.3) satisfy the following conditions:*

a) $\widehat{a}$ intersects the convex hull of x.

b) Furthermore, if $\eta(\theta) = \theta$, then (θ^*, a^*) is the unique global maximum of the large sample limit of the data log-likelihood $\lim_{n\to\infty} \frac{1}{n} l_{\theta,a}(x)$.

Proof sketch. a) This follows from the fact that in the noiseless setting, in order to maximize the probability of the observed samples, we want to cut away as much of the original distribution as possible.

b) $\widehat{a} = a^*$ follows from the fact that in the large sample limit, all other faces of the convex hull of the sample cut away zero mass of the original

distribution in the limit, combined with Theorem 5.1 and the consistency of maximum likelihood estimates.

□

Further, we will see in the experiments that empirically, we obtain good estimates of the true parameters. Note that here, we assume that the selection effect is noiseless. While it would be straightforward to explicitly model the effect of the noise and include the relevant update step for its parameter, we did not find this to produce significantly better results in practice.

5.3.2 Finding Invariances and Selection Boundaries

Next, we move away from strict parametric model assumptions and instead develop an approach based on invariances of the underlying distribution outlined in Section 5.2.2. For the sake of feasibility, we restrict ourselves to the simple yet expressive class of orthogonal matrices—but remark that some progress on discovering larger classes of symmetries has been made recently (Desai et al., 2022; Gabel et al., 2023; Yang et al., 2023).

To motivate our approach, recall that if P is invariant under U^*, the sample U^*x should look indistinguishable from the sample x. Hence, given the sample x from the distribution $Q_a = P(\cdot \mid a^\top X > 0)$, we would like to maximize some similarity measure of the datasets U^*x and x. Since selection is at play, however, even the true invariance U^* cannot apply to all samples x_i, as we saw in Figure 5.4. To address this issue, we will have to learn an invariance while simultaneously considering that some parts of our obtained data will not be consistent with this learning task, as we shall describe below.

To begin with, we will base our approach of measuring the similarity between the distribution P and $P \circ U$ for a given U on the well-established theory of kernel mean embeddings μ_P of P (Muandet et al., 2017), given by

$$\mu_P = \int k(\cdot, x) P(x) dx \,,$$

where $k(\cdot, \cdot)$ is a kernel function. One can show under general conditions—the kernel k being *generic*—that $\mu_P = \mu_Q$ if and only if $P = Q$ (Gretton et al., 2012). In particular, $\mu_P = \mu_{P \circ U}$ if and only if U is an invariance of P.

In the absence of selection bias, our goal would be to find the matrix U that minimizes the distance between these embeddings, $\|\mu_P - \mu_{P \circ U}\|$. The empirical estimate of this distance for a sample x is given by (Gretton et al., 2012)

$$\frac{1}{N^2} \sum_{i,j} k(x_i, x_j) + \frac{1}{N^2} \sum_{i,j} k(Ux_i, Ux_j) - \frac{2}{N^2} \sum_{i,j} k(Ux_i, x_j) \geq 0 \,.$$

To reduce the number of terms that need to be computed, it will be useful to use an isotropic kernel, $k(Ux, Uy) = k(\|Ux - Uy\|) = k(\|x - y\|)$. Fortunately, the commonly used Gaussian kernel $k(x, y) = \exp(-\lambda \|x - y\|^2)$ satisfies both this property and is also characteristic in the above sense of distinguishing any two distributions. Using such an isotropic kernel, the first two terms above are independent of U so that we can maximize

$$L(U; x) = \frac{1}{N^2} \sum_{i,j} k(Ux_i, x_j) \,.$$

If $k(x, y) = \langle x, y \rangle$, this would measure a weighted average of the cosine similarity between the rotated data Ux_i and the original data x_j.

Unfortunately, due to selection bias, not all samples available to us are "good" samples that respect symmetry. To understand the problem, we should understand what happens for a *true* invariance. That is, if U^* is an invariance of P, how does the selection mechanism $a^\top X > 0$ affect this?

Clearly, those points for which $a^\top U^* x_i > 0$ are unaffected in the above score. Meanwhile, those points for which $a^\top U^* x_i \leq 0$ are far away from every point x_j in the available sample, incurring a large penalty. Hence, the score L would be improved if we recomputed the average without the terms $k(U^* x_i, x_j)$ for these points which lie far away. As such, we propose the following approach to determining which samples are "good" samples. First, we optimize $L(U; x)$ with respect to U, as described below in detail. Then, for each point x_k we check if for $\mathcal{I} = \{1, \ldots, k-1, k+1, \ldots, n\}$ we have

$$L(U; x, \mathcal{I}) = \frac{1}{N \,|\mathcal{I}|} \sum_{i \in \mathcal{I}, j \in [n]} k(Ux_i, x_j) \gg L(U; x) \,.$$

In other words, we check if the set of points $\{Ux_i\}_{i \in \mathcal{I}}$ contains significantly fewer outliers relative to x than the sample $\{Ux_i\}_{i \in [n]}$. We then temporarily remove all "bad" points and set $\mathcal{I} = [n] \setminus \{k_1, \cdots, k_l\}$. We rerun the optimization of U starting at its previously discovered optimum, using the score $L(U; x, \mathcal{I})$ instead. After each optimization step over U, we evaluate $L(U; x, [n])$ and recompute the set $\mathcal{I}$ by removing indices from $[n]$. We do this instead of removing points from $\mathcal{I}$ cumulatively because some of the points we previously considered "bad" might only have appeared thus due to U being poorly optimized. We repeat this process until the pair $(U, \mathcal{I})$ stops changing.

Next, we must deal with the question of how to actually optimize the score $L(U; x, \mathcal{I})$ over U. The trouble is that optimization over elements of the orthogonal group $O(m)$ is difficult due to the requirement that gradient steps

stay within the manifold. In other words, our goal of solving

$$\begin{aligned} &\max_{U} L(U; x, \mathcal{I}) \\ &\text{s.t.} \quad U \in O(m)\,, \end{aligned}$$

is made difficult by the fact that $O(m)$ is not a convex set. One way to deal with this would be to use the Lie group structure of $O(m)$ by using Givens rotations (Shalit and Chechik, 2014) to perform coordinate descent, but we have found this method to be both slow and not result in very good solutions. Instead, we can turn the problem into an optimization problem with a convex constraint. To optimize $L(U; x, \mathcal{I})$ over the set of (special) orthogonal matrices, we use the Cayley transform (Wen and Yin, 2013)

$$U = (I - A)(I + A)^{-1}\,,$$

where A is a skew-symmetric matrix, $A^\top = -A$. Since we can write any skew-symmetric matrix A as the difference $B - B^\top$ where B is strictly upper triagonal, this turns the constrained optimization problem of $L(U)$ with respect to orthogonal matrices into an optimization problem $L(B)$ over the convex set of strictly upper triagonal matrices B.
Note that while the matrices U parametrized in this way are all rotation matrices that lie in the special orthogonal group $\mathrm{SO}(m)$, i.e., $\det(U) = 1$, this is not a concern for us. If P is invariant to U, it is also invariant to $U^2 \in \mathrm{SO}(m)$. Since our primary use for the matrix U is to determine the effects of selection bias, this purpose is equally well-served by working only with matrices in $\mathrm{SO}(m)$.
Once we have obtained an orthogonal matrix U and the index set $\mathcal{I}$, we can use these to estimate the selection boundary. Let $\mathcal{I}^c = [n] \setminus \mathcal{I}$. Then the points x_k for which $k \in \mathcal{I}^c$ are such that Ux_k is far from observed samples x_i and are likely to lie in the region $a^\top U x_k < 0$, i.e., the other side of the selection boundary. We therefore use a linear classifier such as an SVM to separate the two sets of points $\{x_i\}_{i\in[n]}$ and $\{Ux_k\}_{k\in\mathcal{I}^c}$. We will see in the experiments that this simple approach already produces good results.

5.4 Selected Related Works

While a fair amount of research has been done on dealing with the effects of latent confounding in causal discovery and inference, selection bias is much less well-studied. It is well-known that selection bias can have detrimental effects on statistical inferences, especially regarding public health advice (Berkson, 1946; Herbert et al., 2020), and that working with samples subject to selection bias can also reinforce stereotypes, causing issues with regard to the fairness of algorithmic decision-making (Caton and Haas, 2024). Furthermore, work

with such biased data leads to problems regarding the *transferability* of derived results to other populations (Naci and Ioannidis, 2013; Averitt et al., 2020).

Most work done on the topic of selection bias focuses on conditions under which selection bias can be controlled for (Bareinboim and Pearl, 2012; Bareinboim et al., 2014; Bareinboim and Pearl, 2016; Forré and Mooij, 2020; Versteeg et al., 2022; Kundu et al., 2024), generally assuming that the selection variable is measured. In the bivariate setting, previous research has worked on the identifiability of causal directions under selection (Zhang et al., 2016) and to perform linear regression under self-selection bias (Cherapanamjeri et al., 2023).

In contrast, little work has been done to discover whether there is selection bias in the first place. Our concern is precisely under which conditions it is possible to determine whether selection bias is a likely concern for a given dataset.

Some related approaches are those dealing with covariate shift (Gretton et al., 2009; Sugiyama et al., 2007) using kernel-based methods, or propensity scoring (Wang and Kim, 2021; Schoeler et al., 2023). However, they generally require access to multiple datasets, making them unusable when only one dataset subject to selection bias is available. Similar methods use the idea of covariate shift differently and instead attempt to provide *robust* estimates which are valid for any (small) amount of selection effect (Cortes-Gomez et al., 2023), or to combine data from biased and unbiased sources (Elliot, 2009).

Instead, we focus on the question of to what extent we can do this from a single dataset in which it is *not* known beforehand that it is affected by selection bias.

Our approach for exponential families is an EM-like approach (Dempster et al., 1977) and has, since the publication of the paper on which this chapter is based, been extended to arbitrary selection criteria (Lee et al., 2024), although the authors of that paper appear to have little interest in the ensuing biases.

The study of symmetries in probability distributions garnered much attention at the start of the century (Fang et al., 1990; Chikuse, 2003; Kallenberg, 2005). More recent theoretical work has focused chiefly on providing theoretical frameworks to explain the benefits of symmetries for predictive tasks (Lyle et al., 2020; Fortuin, 2022; Chen et al., 2020; Dao et al., 2019). A different line of research has focused on learning models invariant to a given symmetry group T. van der Wilk et al. (2018) developed invariant Gaussian processes f by averaging a Gaussian process g over the orbit of T. Further work also extended this line of work to neural networks (van der Ouderaa and van der Wilk, 2022). Note that these approaches assume that T is known beforehand. Benton et al. (2020) relax this assumption by parametrizing the set of transformations. Other recent work has focused on leveraging the benefits of symmetries, especially in image recognition systems (Ravanbakhsh et al., 2017; Worrall et al., 2017; Immer et al.). However, these approaches focus on exploiting symmetries in the data-generating process to achieve a specific supervised task.

In contrast, SymmetryGAN (Desai et al., 2022) uses a generative adversarial

network (GAN) to learn a linear volume-preserving transformation that leaves the data invariant. Other work has also been done on discovering symmetry groups rather than individual symmetries (Yang et al., 2023).

5.5 A Selection of Experiments

In this section, we perform a comprehensive experimental analysis of our proposed methods. We are interested in seeing how well we recover the selection boundaries and which variables are subject to selection bias. To verify that our methods work, we compare them with two different approaches. First, we use the kernel mean matching (KMM; Gretton et al., 2009) algorithm, designed to tell whether there is a distribution shift between *two* datasets. Second, we use DCD (Bhattacharya et al., 2021), designed to discover confounding. It models non-causal edges, although we will see that it does not perform well at detecting selection bias. We implement our methods in Python using Tensorflow (Abadi et al., 2016) and use the publicly available implementations of KMM and DCD by the respective authors. All code and data can be found online.[2]

5.5.1 Data generation

As before, we start by generating a random directed Erdős-Rényi (ER) network G with probability of an edge being added being p. We define the distribution over $X_1, \ldots, X_m$ via the structural model $X_i = f_i(\mathrm{Pa}_i, \varepsilon_i)$ for appropriate functions f_i and noise variables ε_i, where Pa_i are the parents of X_i in G.
For the multivariate Gaussian distribution, this is $X_i = \beta_i^\top \mathrm{Pa}_i + \varepsilon_i$ where $\beta_i^\top \sim N(0, \sigma_\beta^2 I)$ and $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$. For data generation from the t-distribution, we use the formulation of Kotz and Nadarajah (2004).
We generate samples $x = (x_1, \ldots, x_m)$ from $P(X)$ and then pick a random sink node Z from G and remove all samples for which $Z + a_0 < 0$ where $a_0 \sim N(0, \sigma^2(f_Z))$. For the Gaussian distribution, we pick $\sigma^2(f_Z) = \beta_{i_1}^2$ where i_1 is the first parent of Z in the topological ordering. Note that this is the setting used in Theorem 5.2 if $P(X)$ is Gaussian. For each instantiation of the parameters, we generate data points until a total of 1000 points are included in the observed data. We further run each experiment 1000 times.

5.5.2 Recovering the selection boundary

We start our evaluation by checking how well each method predicts the correct selection boundary in a dataset that is known to be subject to selection bias. To this end, we generate data from three-dimensional Gaussian and t-distributions

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

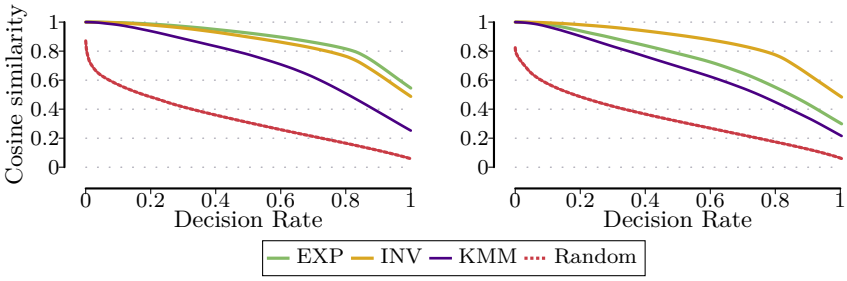


Figure 5.5: [Higher is better.] *Decision rate plots of the cosine similarity between discovered and true selection boundary.* Left: Experiments on Gaussian distributed data. EXP outperforms INV slightly, and both outperform KMM significantly. Right: Experiments on t-distributed data. INV outperforms both EXP and KMM significantly. In both cases, all methods significantly outperform random guessing.

$N(\mu, \Sigma)$ and $t_\nu(\mu, \Sigma)$ subject to selection bias as described above, where we set $p = 0.3$ for our ER network. We then compute the cosine similarity $0 \leq \frac{a^\top a^*}{\|a\|\|a^*\|} \leq 1$ between the true selection boundary a^* and the estimated a. A result closer to 1 corresponds to better performance.

We compare only with KMM here as DCD is not capable of estimating a^*. For KMM, we make some modifications to the data it has access to in order to make it applicable to our setting. Since KMM requires two datasets, besides the one subject to the true selection boundary a^*, we also give it access to a second dataset subject to the selection boundary a', which is a slightly perturbed version of the true a^*. Thus, the original data x and the secondary dataset x' share similar distributions, which are nevertheless different and are therefore amenable to analysis using KMM. In particular, by adding a minor permutation to the selection boundary, the distinction between points that are included in one dataset but not the other should make it possible to detect the selection boundary by distinguishing between assigned weights ≈ 1 by KMM, and assigned weights ≈ 0. More precisely, we use two similar datasets with a^* and a' where $a' = Ua^*$ with an orthogonal matrix U. To this end, we sample random orthogonal matrices via the Cayley transform until the two selection boundaries are similar enough in the sense that $\frac{a'^\top a^*}{\|a'\|\|a^*\|} \geq 0.95$.

We show the results in a decision-rate plot in Figure 5.5. As before, on the x-axis is the decision rate, i.e., the fraction or number of datasets evaluated so far, ordered from most to least confident for each method. On the y-axis, we show the cosine similarity between the recovered and true selection boundary. We see clearly that for all three methods, the confidence strongly correlates with their performance on both datasets. On the left, we see that for Gaussian generated data, EXP with a Gaussian exponential family performs slightly better than INV, although not significantly. Both methods significantly outperform

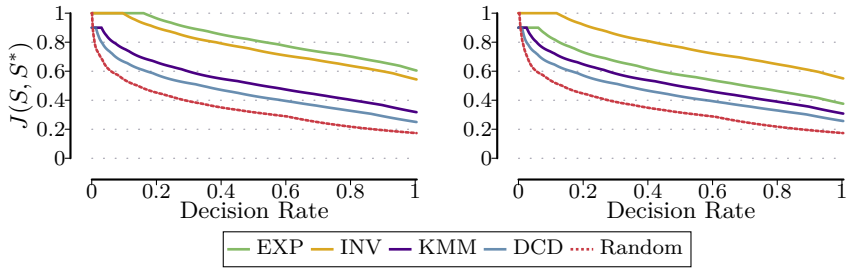


Figure 5.6: [Higher is better.] *Decision rate plots of the Jaccard similarity between recovered and true set of variables affected by selection bias.* Left: Experiments on Gaussian data. EXP outperforms INV slightly, and both outperform KMM and DCD significantly. Right: Experiments on t-distributed data. INV outperforms EXP, which in turn outperforms KMM and DCD significantly. All methods significantly outperform random guessing.

KMM. On the right, for t-distributed data, INV significantly outperforms EXP, which in turn significantly outperforms KMM. Lastly, all methods significantly outperform random guessing on both datasets at all levels.

5.5.3 Recovering variables affected by selection

Next, we consider the task of discovering which variables are affected by selection bias. As described above, we generate data from a ten-dimensional joint distribution with $p = 0.3$ for the ER Graph. Then, the parents of the variable Z we condition on are the variables we would like to recover.

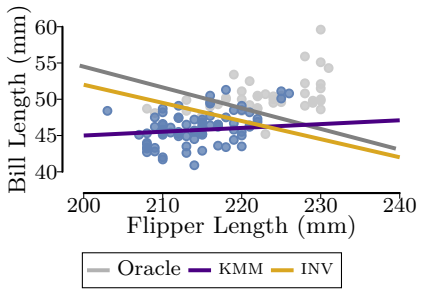
For evaluation, we compute the Jaccard similarity between the true set $S^* = \mathrm{Pa}_G(Z)$ of variables subject to selection and our recovered S,

$$J(S, S^*) = \frac{|S \cap S^*|}{|S \cup S^*|} \in [0, 1],$$

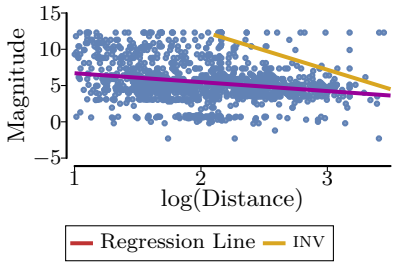
where higher values tell us that S is more similar to S^*.

We compare our methods with KMM and DCD in this setting. For our methods and KMM, we use the discovered selection boundaries a and consider those variables X_i whose a_i is significantly different from zero to be subject to selection. For DCD, we run the method to obtain pairs of variables whose correlations are estimated to be (partially) non-causal. We then estimated the set of variables affected by selection to be all variables included in at least one such pair.

We show the resulting decision rate plots in Figure 5.6. As in the previous section, for Gaussian generated data, EXP outperforms INV slightly. Further, both of our methods outperform both KMM and DCD significantly. For t-distributed data, INV again outperforms EXP significantly, which in turn significantly out-



(a) Result on Palmer penguin dataset. Observed points are displayed in blue, unobserved points in gray. The oracle separator found by LDA is displayed in gray. We see that the selection boundary found by INV (gold) is much closer to the oracle line than the one found by KMM.



(b) Result on exoplanet dataset. The regression line (dark blue) is negative and highly significant, $p < 10^{-20}$. The selection boundary estimated by INV (gold) captures the intuitive lack of points in the top right corner, consistent with selection due to technological limitations.

Figure 5.7: *Results on Palmer penguin (a) and exoplanet datasets (b).* Since we don't know the true selection effect for the exoplanet dataset, we can neither use KMM on this data, nor provide an oracle selection boundary to compare against.

performs both KMM and DCD. Further, all methods significantly outperform random guessing of the set of variables affected by selection.

5.5.4 Real Data

Next, we want to see if our methods can provide novel and interesting insight into real data. We, therefore, evaluate them on two real-world datasets.

Palmer Penguins

We begin by evaluating our methods on the Palmer Penguins dataset (Gorman et al., 2014), collected at Palmer Station, Antarctica. It contains samples from three different species of penguins. Among the measured variables are bill depth, bill length, flipper length, and weight of each penguin.
Clearly, each of the variables should be positively correlated with the weight of the penguin. We can, therefore, pose the hypothetical question, "what if we had only measured the lightest penguins because they were easier to capture and transport?" Since both larger bill size and flipper size lead to larger weight, we expect that conditional on weight, these variables will become negatively (or at least less positively) correlated with each other. To test whether we could, in fact, find such results in this dataset, we preprocess our data as follows. We

split the data by penguin species, and then for each of them, we select the 80% of penguins with the lowest weight from that species, leading to selection bias. We show the results for Adele penguins in Figure 5.7a, where gray points have been removed in the preprocessing step. We see that the selection boundary estimated by our approach are reasonable, while the one estimated by KMM is not. In fact, when compared to a Linear Discriminant Analysis (LDA; Fisher, 1936; Bishop and Nasrabadi, 2006) fit on *all* data with known labels of which data have been excluded, the selection boundaries discovered by INV is almost identical to the optimal linear separator between included and excluded points.

Exoplanet Discovery

Next, we consider data from the Open Exoplanet Catalogue using ExoData (Rein, 2012; Varley, 2016). It contains data about exoplanets and their stars, including the distance d from the earth and their absolute magnitude—a measure of their brightness, measured in terms of $-2.5 \log_{10} B$, where B is the star's brightness. Hence, bright stars have low magnitude, and dim stars high magnitude.
It is generally believed that the universe is uniform at large scales (Liddle, 2015), i.e., the universe should look the same on average independently of which direction we look and far out we look. The distance of stars from us should therefore be independent of their absolute magnitude. However, due to technological constraints, the further away a star is, the brighter it has to be for us to detect exoplanets in its system. As such, we expect selection effects should be expected, making it a good case study for our methods.
We show the data in Figure 5.7b. One thing that stands out from the very first glance is that the top right corner of our dataset (dim points that are very far away) is only sparsely populated. Indeed, the linear correlation between $\log(d)$ and magnitude (dark blue) is negative and significant at the 10^{-21} level. Applying INV, we obtain the selection boundary seen in Figure 5.7b, suggesting that too few points lie in the top right corner (far away dim stars). We see that the selection boundary found by INV is consistent with our speculations of selection effects based on technological limitations.

5.6 From Isolated Insight to Environmental Ensembles

In this chapter, we tackled Problem 3 and studied to what extent we can discover whether a set of variables is affected by selection bias or not.
We began by introducing selection bias as the preferential inclusion of some data points over others due to conditioning on a variable that is causally downstream of the observed variables and described the specific linear selection model we studied. We then studied how such selection effects can be recovered in two different settings. First, we saw that for exponential family models, the

density ratios between different points unaffected by the selection mechanism are the same regardless of the precise selection mechanism, and leveraged this information to show identifiability of the underlying parameters (Section 5.2.1). Second, for nonparametric models for which we know only that a certain kind of invariance exists, we showed that the partial violation of such invariances due to selection can be used to recover the selection boundary (Section 5.2.2). From these characterizations, we then derived two algorithms, one for each approach. For exponential families, we employed a simple alternating optimization scheme and characterized the saddle points of the optimization procedure in Proposition 5.4. Meanwhile, for the nonparametric approach, we introduced an optimization to find an orthogonal matrix that leaves the underlying distribution invariant. By characterizing the difference between points that will and those that will not respect the learned invariance and employing the Cayley transform to turn the optimization into an unconstrained optimization problem, we obtained a fast and reliable solution to this problem.

One direction for future work is to use richer selection models. In this chapter, we considered only linear selection boundaries, but this is often not very realistic in practice. We can readily see two different approaches for generalizing this model of selection. The first way is to replace all occurrences of $a^\top X$ with a kernelized version $k(a, X)$, corresponding to an inner product taken in some higher-dimensional space. This approach may be suitable when selection occurs on the measured variables but is simply nonlinear. In contrast, selection may also occur in some lower-dimensional space, such as when X lies on some lower-dimensional manifold. While the selection boundary in this case is also nonlinear, it is, in fact, still linear in the parametrization of the data manifold. Unfortunately, it is not immediately clear how to learn such a latent representation in a selection-aware manner, such that the representation does not distort the data but can still be used to capture the relevant selection mechanisms.

A second direction for future work is the implementation of richer models of invariance, such as the MPAs we have seen in the previous chapter. That is, the observed variables X could be produced as a nonlinear mixture of some source variables S, where it may be simpler to specify the invariances on S. It is quite clear that this and the previous point on learning representations on which selection occurs are tightly connected, and it would be exciting to study the connections and possibilities of such an approach in more detail. Naturally, with this connection also come problems of how we can learn the nonlinear mixing, invariance, and selection mechanisms on the underlying sources.

The common thread running through these last four chapters has been the use of relatively strong (mostly parametric) assumptions on the true distribution of our data, which permit us to derive our desired quantities from a *single* dataset. Next, we shift our attention to the case of *multiple* environments and study how much such additional data helps us discover latent confounding.

Chapter 6

Discovering Confounders from Independent Mechanism Shifts

One for sorrow,
Two for mirth,
Three for a funeral,
And four for birth.
Five for heaven,
Six for hell,
Seven for the devil, his own self.

One for Sorrow

In the previous chapters, we have seen that in order to discover latent confounding or selection bias and to recover the true causal networks and effects in spite of these influences, we need to make relatively strong assumptions about the shape of the underlying distribution, in terms of its parametric form, its structural sparsity, or its underlying symmetries.
Perhaps the most significant advance in causal discovery has been the principle of invariant causal mechanisms (Peters et al., 2016; Arjovsky et al., 2019; Huang et al., 2020; Mooij et al., 2020) across environments, permitting us to leverage multiple datasets gathered under different conditions to learn additional information about the underlying causal structure. Consider, for example, the case

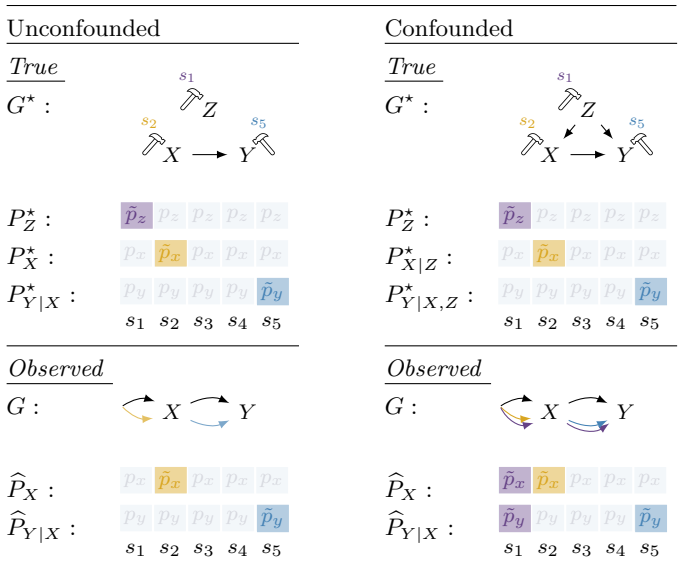


Figure 6.1: *Confounding introduces dependent mechanism shifts.* We consider two systems, one causal (left) and one confounded (right) in five different contexts. The true causal mechanisms $P^\star$ (top) change independently of each other, here due to targeted interventions in certain contexts (colored). If there is an unobserved confounder Z, however, we observe dependencies in the mechanism shifts of X and Y (bottom right).

of gene regulation. Modern tools (Dominguez et al., 2016) allow us to activate or silence specific genes directly. The idea is that if we knew the true (causal) gene regulation networks, the regulation mechanisms corresponding to other edges in the network should be unaffected by such an intervention. That is, the causal mechanisms that are not directly affected by an intervention should remain *invariant* under such changes. Reality does not quite match this idealized scenario, and off-target effects are not unheard off (Zhang et al., 2015). As such, other methods emphasize not the strict invariance of causal mechanisms, but instead the fact that *the number of violations is minimal in the true graph* (Mameche et al., 2022, 2023; Perry et al., 2022).

In this chapter, we investigate to what extent we can leverage similar ideas to discover latent confounding among several observed variables. The basic idea is quite simple. Let X, Y be two observed variables, and Z be unobserved, and assume that the joint distribution $P^s(X, Y, Z)$ can depend on an extra influence s, such as an index telling us which variable has been intervened upon. Then, what evidence would we expect to see in the case that Z influences both X and Y? In Figure 6.1, we show the difference between the two cases where Z

does (right) or does not (left) influence X and Y. The fundamental insight is that in the confounded model, whenever the distribution of the unobserved Z changes, so do the observed distributions of *both* X and Y. In particular, if we assume that interventions performed on X and Y are independent of each other, these joint shifts induced by the confounder Z lead to a testable statistical criterion by measuring whether the shifts in the mechanisms of one variable are predictive of the mechanism shifts of another.
We formalize and explain our setup and assumptions in Section 6.1. In Section 6.2, we then show that under these assumptions, we can recover not only which variables are confounded but also how many latent confounders there are, as well as the true causal graph over both observed and unobserved variables. In Section 6.3 we then provide an efficient algorithm to detect confounders for groups of variables. Finally, in Section 6.5 we show that COCO performs much better at recovering which variables are confounded and also which variables are causally related than its competitors, before wrapping up in Section 6.6. As before, proofs for all statements are postponed to Appendix A.6.

6.1 The More the Merrier

In this section, we present our problem setting and state our assumptions.

6.1.1 Problem Setting

We consider a system of observed variables X and unobserved variables Z, collectively called $V = X \cup Z$. The values of X, Z may be continuous, categorical, or mixed. We assume that the system is observed in multiple *settings* (also known as environments or *contexts*), represented by a categorical variable S taking values $s \in \mathcal{S}$, and denote their number $n_s = |\mathcal{S}|$. We allow the distribution $P^s(V) = P(V \mid S = s)$ to depend on s, as described below.
We assume that the causal relationships between variables V will be described by a *fixed* DAG $G^* = (V, E^*)$, independent of $s \in \mathcal{S}$. W.l.o.g. the indices of X_i and V_i are assumed to be ordered such that, whenever clear from the context, we can write Pa_i^* to denote the parents of X_i. As before, we assume that all Z are jointly independent and that no reverse causation $X \to Z$ exists.
We assume causal sufficiency to hold over all variables $X \cup Z \cup S$, but not over $X \cup S$. We can now state our problem informally as follows.

Problem Statement. Given data over the observed variables X in contexts $s \in \mathcal{S}$, which of the observed variables among X are jointly confounded?

To solve this problem, we next provide a describe in more detail how the causal model varies across contexts, and how these variations are formed.

6.1.2 Causal Mechanism Shifts

We now describe the data-generating process across multiple settings. While we assume that the same causal structure applies in all settings, in many applications such as gene editing experiments (Barrangou and Doudna, 2016), a system is subject to interventions or other causal mechanism changes. That is, the generating process $P^s(V_i \mid \mathrm{Pa}_i^*)$ of each variable V_i may be different across settings. Nevertheless, as interventions typically affect only a small number of causal mechanisms at a time, the causal mechanism governing a specific V_i will generally be the same for most $c \in \mathcal{S}$ and differ only between few. To represent this, for every variable V_i, we partition the settings so that the causal mechanism remains constant within each set. That is, for each V_i we have a partition $\Pi_i^* = \left\{\pi_i^1, \ldots, \pi_i^{k_i}\right\}$ of $\mathcal{S} = \pi_i^1 \cup \ldots \cup \pi_i^{k_i}$ into disjoint π_i^j such that $P^s(V_i \mid \mathrm{Pa}_i^*) = P^{s'}(V_i \mid \mathrm{Pa}_i^*)$ for s, s' in the same $\pi \in \Pi_i^*$. We refer to the set π containing s as $\Pi_i^*(s)$, and call the corresponding mechanism $P_i^{\pi}(V_i \mid \mathrm{Pa}_i^*)$. For example, in Figure 6.1, variables X, Y have partitions $\Pi_X^* = \{\{1,3,4,5\},\{2\}\}$ and $\Pi_Y^* = \{\{1,2,3,4\},\{5\}\}$. If the mechanism of variable V is invariant across *all* environments then $\Pi_V = \{\{1,\ldots,n_s\}\}$. Likewise, if the mechanism of variable V is different for all environments, then $\Pi_V = \{\{1\},\ldots,\{n_s\}\}$.
We allow all partitions Π_i^* to be distinct. More precisely, we regard partition Π_i^* of the contexts $\mathcal{S}$ as random variables and assume that there exists some joint distribution $P(\Pi^*)$ over all partitions Π_i^* of V_i. We hence assume the distribution of the observed V to be as follows.

Assumption A (Markov Property under Mechanism Changes)**.** The distribution $P(V)$ can be written as a mixture

$$\begin{aligned} P(V) &= \int P^S(V) dP(S) \\ &= \int \prod_i P^{\Pi_i^*(S)}(V_i \mid \mathrm{Pa}_i^*) dP(S) \\ &= \int \prod_i P^{\Pi_i^*}(V_i \mid \mathrm{Pa}_i^*) dP(\Pi^*)\,. \end{aligned}$$

In other words, the variables V are assumed to be *conditionally exchangeable*, with the same Causal Bayesian Network G^* applying in every context $s \in \mathcal{S}$ (Guo et al., 2024). Importantly, the distribution $P(V)$ does not depend on $P(S)$ itself, except insofar as $P(S)$ affects the joint distribution of $P(\Pi^*)$.
For an overview, we refer to Figure 6.1, as introduced in the introduction. The causal graph over X, Y, Z is shared across all environments $s \in \mathcal{S}$, and

mechanism shifts are indicated by hammers at the top and multiple edges at the bottom. Each variable is associated with a partition Π_i^*, showing which mechanism applies in which environment (colored boxes). Next, we introduce the properties of causal mechanism shifts relevant to confounder identification.

6.1.3 Independent Mechanism Shifts

In Chapter 2, we introduced the independence of causal mechanisms—causal mechanisms of different variables contain no information about each other—and leveraged it in a single setting. We now extend this principle to *multiple settings*. That is, not only do we want $P^s(V_i \mid \mathrm{Pa}_i^*)$ to be uninformative of $P^s(V_j \mid \mathrm{Pa}_j^*)$ for $i \neq j$, but also a change in mechanisms $P^s(V_i \mid \mathrm{Pa}_i^*) \neq P^{s'}(V_i \mid \mathrm{Pa}_i^*)$ for $s \neq s'$ should not tell us anything about the existence of a change in the mechanism $P^s(V_j \mid \mathrm{Pa}_j^*) \neq P^{s'}(V_j \mid \mathrm{Pa}_j^*)$ for X_j. In particular, we assume that all partitions Π_i^*, Π_j^* are jointly independent.

Assumption B (Independent Mechanism Shifts)**.** We assume that the mechanism changes of $P^s(V_i \mid \mathrm{Pa}_i^*)$ are independent and identically distributed across environments. More precisely, we assume that

$$P(\Pi^*) = \prod_{V_i} P(\Pi_i^*)\,,$$

Note that we assume that all distributions $P(\Pi_i^*)$ are equal. This is not strictly necessary but allows us to simplify notation and exposition. Even so, independence of mechanism shifts is, of course, not a sufficient constraint. The mechanisms of V_i and V_j, both differing across all (or no) environments, would trivially satisfy this condition, but this would reveal no information about the core causal mechanisms we are interested in. Instead, we want to have (a number of) environments s, s' between which precisely one of the mechanisms of V_i and V_j changes. To ensure that such environments exist, we additionally assume that mechanism shifts are sparse so that mechanisms remain the same across most environments, so that joint mechanism changes due to latent confounding can be detected (Guo et al., 2024; Schölkopf et al., 2021).

Assumption C (Sparse Mechanism Shifts)**.** Let S and S' be two i.i.d. samples from the same distribution $P(S)$. We assume that for all variables V_i, the probability of mechanism changes between two contexts is given by

$$p = P(\Pi_i^*(S) \neq \Pi_i^*(S')) < 0.5\,.$$

With this, we assume that mechanism shifts occur infrequently, implying that causal functions persist across the majority of environments. This assumption is valid in many study settings where specific targets are interventions in only

few contexts and has been adopted in the causal discovery literature (Perry et al., 2022; Mameche et al., 2023).

Example 6.1. Let us consider a few examples to gain an intuition for when this assumption does or does not hold.

a) Let each variable X_i have a distribution over $\Pi_i^* = \{\pi_{i,1}^*, \dots, \pi_{i,r}^*\}$, given by $P(S \in \pi_{i,j}^*) = p_j$. Then

$$p = 1 - \sum_j p_j^2 \,,$$

is precisely the Gini index of the distribution. In particular, if all $\pi_{i,j}^*$ have the same probability $1/r$ then $p = 1 - 1/r \geq 0.5$ for any $r \geq 2$.

b) In contrast, if there is a distinguished set $\pi_{i,0}^*$, e.g., corresponding to a purely observational setting, containing most of the datasets, then

$$p \leq 1 - p_0^2 \,,$$

which is ≤ 0.5 so long as $\pi_{i,0}^*$ contains at least a fraction of $1/\sqrt{2}$ of all datasets. This is commonly the case, as observational data is much easier to obtain than interventional data.

c) Note that since we assume independence of mechanism shifts, designs such as diagonal intervention designs in which each variable i is intervened on in precisely one environment,

$$\Pi_i^*(s) = \begin{cases} \pi_{i,1} \text{ if } s = i \\ \pi_{i,0} \text{ otherwise} \,, \end{cases}$$

are not permitted within this framework.

Conversely, we assume that when two settings s, s' are assigned to different sets of the partition Π^*, the corresponding causal mechanisms indeed change.

Assumption D (Π-faithfulness)**.** Let Π_i^* be the partition of V_i. Then for any two s, s', we have the equivalence

$$\Pi_i^*(s) \neq \Pi_i^*(s') \longleftrightarrow P^s(V_i \mid \mathrm{Pa}_i^*) \neq P^{s'}(V_i \mid \mathrm{Pa}_i^*) \,.$$

This faithfulness condition ensures that our partitions capture precisely the changes in causal functions. Next, we show how these assumptions, which we assume to hold when variables in V are measured, are violated when some latent factors Z are not observed.

6.2 Identifying Confounding from Mechanism Shifts

We begin by analyzing the effects of latent confounding on the observed partitions of causal mechanisms. Then, we propose an information-theoretic score for determining whether a given set of variables is jointly confounded and provide consistency guarantees for both the recovery of the sets of jointly confounded variables and the underlying causal network.

6.2.1 Confounding Introduces Dependent Mechanism Shifts

All assumptions we made in the previous section are about the *true* partitions Π_i^* of the *true* causal mechanisms over the *true* causal parents Pa_i^*. Since we are not able to observe all variables, the situation changes. To see this, we can consider the following simple example.

Example 6.2. Let X, Y, Z be related by the following linear relationships

$$\begin{aligned} Z &\sim N(0, \sigma_z^2(s)) \\ X &= \alpha Z + \varepsilon_x \\ Y &= \beta X + \gamma Z + \varepsilon_y \,, \end{aligned}$$

where the only source of mechanism shifts is the non-constant variance $\sigma_z^2(s)$ of the unobserved confounder Z. Then, by regressing Y on X, we obtain

$$\begin{aligned} X &\sim N(0, \sigma_x^2 + \alpha^2 \sigma_z^2(s)) \\ \widehat{\beta}_{Y|X} &= \frac{\mathrm{cov}(X,Y)}{\mathrm{var}(X)} = \frac{\beta \sigma_x^2 + \alpha\gamma\sigma_z^2(s)}{(\sigma_x^2 + \alpha^2\sigma_z^2(s))} \,, \end{aligned} \tag{6.1}$$

so that in general clearly both the distributions $P^s(X)$ and $P^s(Y \mid X)$ change as the variance $\sigma_z^2(s)$ of the latent variable Z changes.

As before, in exceptional circumstances of fine-tuned parameters, the above does not apply. If the parameters are chosen as $\beta = 1$ and $\alpha = \gamma$ in Equation (6.1), then $\widehat{\beta}_{Y|X} = 1$ will not change even if σ_z^2 changes. This kind of fine-tuning of the parameters likely happens only in adversarial cases. As in previous chapters, if the parameters are sampled from a continuous distribution, then the probability of obtaining a set of parameters where a change in the mechanism of the confounder Z does not translate into a change in the mechanisms affecting X and Y is zero. This leads us to the following assumption.

Assumption E (Latent Shift Faithfulness)**.** Let Z be an unobserved common parent of all variables in $X_{\mathcal{I}} \subseteq X$. Then, each mechanism change in Z between s, s' entails a mechanism change between these contexts for each $X_i \in X_{\mathcal{I}}$.

Note that we do not strictly need *all* mechanism shifts of Z to be reflected in X, Y, but only that some (non-zero) fraction is captured. Essentially, we could restrict our analysis to that subset of environments for which these changes are reflected and obtain the same results. Without loss of generality, and to ease the exposition in the following, we therefore work with the above assumption. Hence, changes in the causal mechanism of Z lead to correlations between the observed partitions Π_i of variables affected by Z. We, therefore, now turn to the question of how to measure these correlations.

6.2.2 Mutual Information of Mechanism Shifts

To measure whether the mechanism changes of variables are dependent, we consider the Mutual Information (MI) between partitions. For two partitions Π_1, Π_2 of contexts $\mathcal{S}$ into r, s sets, we consider the contingency table $\mathcal{T}$,

$$\mathcal{T} = \begin{pmatrix} n_{11} & n_{12} & \dots & n_{1s} \\ n_{21} & n_{22} & \dots & n_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ n_{r1} & n_{r2} & \dots & n_{rs} \end{pmatrix},$$

where n_{ij} measures the number of contexts in the intersection $\pi_1^i \cap \pi_2^j$ of $\pi_1^i \in \Pi_1$ and $\pi_2^j \in \Pi_2$, and with row margins $u_i = |\pi_1^i|$ and column margins $v_j = |\pi_2^j|$ counting the size of partition elements.
If the partitions describe causal mechanism shifts of two variables X_i, X_j, then a latent confounder affecting both X_i, X_j leads to correlations between these partitions. To measure these, we consider the mutual information between Π_1 and Π_2. The marginal entropy of Π_1 and joint entropy of Π_1, Π_2 are

$$H(\Pi_1) = -\sum_i \frac{u_i}{N} \log \frac{u_i}{N},$$
$$H(\Pi_1, \Pi_2) = -\sum_{ij} \frac{n_{ij}}{N^2} \log \frac{n_{ij}}{N^2},$$

with $H(\Pi_2)$ similar, and their mutual information is given by

$$\begin{aligned} I(\Pi_1, \Pi_2) &= H(\Pi_1) + H(\Pi_2) - H(\Pi_1, \Pi_2) \\ &= \sum_{ij} \frac{n_{ij}}{N} \log \frac{n_{ij} N}{u_i v_j}. \end{aligned}$$

This is an empirical estimate of the true mutual information based on a sample from the underlying random variables Π_1, Π_2, also known as the *plug-in estimate* of the true mutual information. In general, given data from only a finite

number of contexts, this plug-in estimate of the mutual information between partitions will be positively biased (Vinh et al., 2009). We can correct for this by comparing it against the expected MI for two independent partitions.

Expected Mutual Information under Independent Shifts

We first consider two independent partitions Π_1', Π_2' with contingency table $\mathcal{T}$ with column sums u and row sums v. To define their mutual information, the hypergeometric model of random partitions has been adopted in the literature (Vinh et al., 2009, 2010). That is, given the marginal counts u, v, the joint counts are assumed to follow a hypergeometric distribution $N_{ij} \sim \mathcal{H}(u, v, N)$, with probability mass function given by

$$P(n_{ij} \mid u, v, N) = \frac{\binom{n_{ij}}{v_j}\binom{N-v_j}{u_i-n_{ij}}}{\binom{N}{u_i}} .$$

The expected mutual information between independent partitions is then

$$\begin{aligned} E\left[I(\Pi_1', \Pi_2')\right] &= \sum_{\mathcal{T}} I(\mathcal{T}) P(\mathcal{T}) \\ &= \sum_{ij} \sum_{n_{ij}} I(n_{ij}) P(n_{ij} \mid u, v, N) , \end{aligned}$$

where $I(n_{ij}) = \frac{n_{ij}}{N} \log \frac{n_{ij} N}{u_i v_j}$ and the inner sum runs over the counts $n_{ij} \in [\max\{0, u_i + v_j - N\}, \min\{u_i, v_j\}]$. By replacing the term $I(n_{ij})$ by $I(n_{ij})^2$, one can similarly compute the second moment, and thus the variance

$$\operatorname{var}(I(\Pi_1', \Pi_2')) = E\left[I(\Pi_1', \Pi_2')^2\right] - E\left[I(\Pi_1', \Pi_2')\right]^2 .$$

With this, we can compute the standardized score of our observed mutual information $I(A, B)$, which is given by

$$t = \frac{I(\Pi_1, \Pi_2) - E\left[I(\Pi_1', \Pi_2')\right]}{\sqrt{\operatorname{var}(I(\Pi_1', \Pi_2'))}} , \tag{6.2}$$

and show next that we can use it to find confounded pairs of variables.

6.2.3 Identifying Confounded Variable Pairs

We show that in the bivariate case, when the causal direction between a pair X, Y is known, we indeed obtain the correct results with high probability when using the score above to determine whether the variables X, Y are confounded.

Lemma 6.1 (Significance and Power)**.** *Let X, Y be unconfounded and $X \to Y$. Let Π_X, Π_Y be the corresponding partitions. Then*

$$\lim_{n_s \to \infty} P(t > q_{1-\alpha}) \to \alpha \,,$$

where $q_{1-\alpha}$ is the $1-\alpha$-quantile of the standard normal distribution. Conversely, if X, Y are confounded, then for $\alpha > 0$ in the limit we obtain power

$$\beta = \lim_{n_s \to \infty} P(t > q_{1-\alpha}) \to 1 \,.$$

Proof sketch. The statement about significance follows from the fact that t is asymptotically normal (Vinh et al., 2009). For the statement about power, we need to note only that $EI(\Pi_1, \Pi_2) \propto n_s$ when variable X_1, X_2 are confounded, but that $EI(\Pi_1', \Pi_2') = o(n_s)$ for random partitions. □

This result tells us that with data from enough environments, we are guaranteed to discover which pairs of variables are confounded. Of course, for any fixed α, there will be some false positives, but as $n_s \to \infty$, we should be able to pick decreasing values of α. Unfortunately, determining how to pick the value $\alpha(n_s) \to 0$ such that $\beta \to 1$ nevertheless holds would require a more detailed analysis, which we are unaware of how to do. Instead, we will include an empirical investigation of these values among our experiments in Section 6.5.

6.2.4 Beyond Confounded Pairs

To determine whether a set of variables shares a joint confounder, we extend our score beyond pairs of variables. A natural extension of mutual information for a set of partitions is total correlation (Watanabe, 1960),

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

where $\Pi_{<i} = \{\Pi_1, \ldots, \Pi_{i-1}\}$ and similarly for $\Pi_{>i}$. It is straightforward to correct this score as we did above for the pairwise mutual information score. As both corrected and uncorrected scores are asymptotically equivalent, we will consider T as is in our theoretical analysis.

First, we discuss how to use this score to detect joint confounding. To this end, consider three variables X_1, X_2, X_3. By Assumption E and Lemma 6.1, we know these can only be jointly confounded if and only if all X_i, X_j are pairwise confounded. It could, of course, be that rather than jointly confounded,

there are three disjoint confounders Z_{12}, Z_{13}, Z_{23} affecting each of the individual pairs. Can we distinguish these two cases? Yes, if all three variables share the same latent confounder Z, then knowing about the partition of one variable explains away some of the correlation between the other two partitions, so that we have $I(\Pi_i, \Pi_j \mid \Pi_k) < I(\Pi_i, \Pi_j)$ for any permutation of the variables. Meanwhile, for three pairwise confounders, this is not the case.
In general, for a set of size s to permit such an equivalent explanation in the first place, we would need to add a total of $\binom{s}{2}$ confounders with $s(s-1)$ outgoing edges to obtain the same structure of pairwise confounding. While this may *plausibly* occur for small sets of variables that appear to be pairwise correlated, we assume the true graph G^* to be causally minimal in the following sense.

Assumption F (Confounder Minimality)**.** For every subset $X_{\mathcal{I}}$ of at least $|\mathcal{I}| \geq 4$ variables, there are at most $2\,|\mathcal{I}|$ edges incoming into $X_{\mathcal{I}}$ from latent confounders Z_j with at least three children in $X_{\mathcal{I}}$.

This minimality assumption ensures that variables that appear to be jointly confounded are indeed confounded by the same latent variable. Equivalently, when few latent variables suffice to explain the observed correlations, there should indeed exist only few confounders. With this, we can guarantee that the identification of joint confounding is possible from the total correlation T.

Theorem 6.2. *Let $X_{\mathcal{I}}$ be a set of variables such that all $X_i, X_j \in X_{\mathcal{I}}$ are pairwise confounded. Then $X_{\mathcal{I}}$ is jointly confounded if and only if for each triple $X_i, X_j, X_k \in X_{\mathcal{I}}$ we have*

$$\lim_{n_s \to \infty} P(T(\Pi_i, \Pi_j, \Pi_k) < I(\Pi_i, \Pi_j) + I(\Pi_j ., \Pi_k)) = \begin{cases} 1, & X_i, X_j, X_k \textit{ jointly confounded} \\ 0, & \textit{otherwise}. \end{cases}$$

Proof sketch. This follows from Assumption F because in order to make all variables in $X_{\mathcal{I}}$ triplet-wise mutually confounded, without using one single confounder Z, requires the use of at least $2\,|\mathcal{I}| + 1$ edges. □

With this, we can recover how many latent confounders Z_j there are, and sets of jointly confounded nodes by each Z_j are uniquely identifiable by our score. Due to the large number of tests involved, potential biases in all the involved tests, and a lack of a search criterion to determine which subsets should be tested in the first place, we propose a more efficient and robust method using spectral clustering based on our pairwise scores in Section 6.3.
As we assumed causal directions among all variables to be known up to this point, the remaining question is what happens if this is not the case.

6.2.5 Spurious Spurious Correlations

We now address the case where the true causal structure is unknown and estimate partitions in the presence of misdirected edges. Can we still use our score based on mutual information to determine which variables are confounded? Can we perhaps even use it to recover the true causal network? It turns out that if the underlying causal network is sparse enough, the answer is yes.

First, let us return to the case of two variables X, Y such that in the true graph G^*, the causal direction $X \to Y$ applies. What would happen if we instead considered the partitions obtained by considering the graph G differing from G^* by inverting this edge to be $Y \to X$ instead? To compare the resulting partitions, we write Π_X, Π_Y for the partitions of causal mechanisms in G, and similarly Π^*_X, Π^*_Y for the true partitions corresponding to G^*.

It turns out that, with high probability, the misdirected edge will introduce additional correlations between the inferred partitions Π_X, Π_Y. Intuitively, this is because distribution shifts in $P^s(Y)$ now need to come with matching mechanism shifts of $P^s(X \mid Y)$ to ensure that $P^s(X)$ does *not* change (Huang et al., 2020). This leads to the following asymptotic result.

Proposition 6.3 (Consistency for Pairs of Variables)**.** *If a variable pair X, Y is confounded by a variable Z, then there exists some constant $\rho > 0$ such that*

$$P(I(\Pi^*_X, \Pi^*_Y) < I(\Pi_X, \Pi_Y)) = 1 - O(e^{-\rho n_s}) \,.$$

Proof sketch. This follows from noting that the mutual information terms are determined by the number of joint shifts between the partitions of X and Y and that this number is larger with high probability in the direction $Y \to X$. □

When X, Y are part of a larger graph, the situation becomes more involved. Based on the ideas from Perry et al. (2022), we show that among those graphs of the Markov equivalence class of the marginal $P(X)$, those which correctly capture the relationship between a target variable X_i and its true parents will minimize the mutual information between its partition, and the partitions of its recovered parents. As we have seen in Chapter 3, however, due to the effects of latent confounders, the MEC over $P(X)$ will contain large numbers of additional edges. We, therefore, show that so long as the number of latent confounders affecting and spurious siblings of a given target X_i are not too large, then we can still recover the correct parents of the target.

Proposition 6.4 (Consistency for Recovering Parents)**.** *Let X_i be a target variable and let G and G' be two graphs in the MEC of the marginal distribution $P^s(X)$. Assume that only one of the two graphs correctly recovers the parents*

of X_i, $\mathrm{Pa}_i = \mathrm{Pa}_i^*$ *and* $\mathrm{Pa}_i' \neq \mathrm{Pa}_i^*$, *and further assume that the number of latent confounders affecting* X_i *plus spurious siblings is bounded by* $\frac{\log(0.5)}{\log(1-p)}$. *Then*

$$\begin{aligned} P(I(\Pi_i, \{\Pi_j : j \in \mathrm{Pa}_i\}) < I(\Pi_i', \{\Pi_j' : j \in \mathrm{Pa}_i'\})) \\ = 1 - O(e^{-\rho n_s}) \,. \end{aligned}$$

Proof sketch. The idea is the same as for Proposition 6.3, except that now the presence of spurious neighbors in the marginal graph introduces additional sources of joint mechanism shifts that need to be accounted for. □

To show that we can consistently discover a causal graph in which the causal ordering between all observed variables X is correctly recovered, we can sum over all the scores in the above Proposition.

Theorem 6.5 (Consistency)**.** *Let* G^* *be the true graph over* V *and let* G_x^* *be the induced graph on* X, *and assume that for all* X_i *the number of latent parents plus spurious siblings is at most* $\frac{\log(0.5)}{\log(1-p)}$. *Then with high probability,* G_x^* *and its partitions* $\Pi_1^*, \ldots, \Pi_k^*$ *are the unique minimum of total correlation,*

$$P\left(\underset{G, \Pi_1, \ldots, \Pi_m}{\arg\min}\ T(\Pi_1, \ldots, \Pi_m) = (G_x^*, \Pi_1^*, \ldots, \Pi_m^*)\right) = 1 - O(e^{-\rho n_s}) \,.$$

Proof sketch. By using that $T(\Pi_1, \cdots, \Pi_m) = \sum_i I(\Pi_i, \{\Pi_j : j \in \mathrm{Pa}_i\})$, the result follows from taking a union bound over all the terms in Proposition 6.4. □

With these theoretical guarantees in hand, we now move on to provide an effective algorithm for discovering which variables are indeed confounded.

6.3 Discovering Confounders from Different Contexts

Based on the framework we developed in the previous sections, we now introduce the COCO algorithm for discovering **Co**nfounders from different **Co**ntexts.

Determining Causal Mechanism Shifts

To develop our algorithm, we use existing approaches for discovering causal mechanisms and their changes in multiple contexts. Since it agrees well with our shift testing approach, we build upon the MSS estimator developed by Perry et al. (2022), which starts from the correct MEC and directs edges to minimize the number of mechanism shifts. For each causal mechanism of a target variable

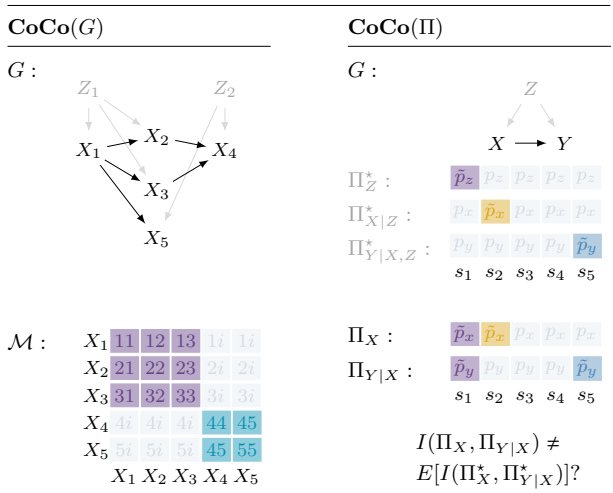


Figure 6.2: *Components of* COCO. In a DAG G with unobserved confounders Z_1, Z_2 (top left), we consider each pair of nodes (top right) discover their partitions, and test them for dependency using MI (bottom right). We obtain an affinity matrix showing which nodes are affected by the same confounder (bottom left).

X_i and each pair of environments, we perform a conditional independence test to detect mechanism changes, resulting in the following p-values,

$$p_{s,s'} = p\text{-val}\left(P^s(X_i \mid \mathrm{Pa}_i) \neq P^{s'}(X_i \mid \mathrm{Pa}_i)\right) \ .$$

We here use the Kernel Conditional Independence test (KCI; Zhang et al., 2011) for all practical purposes, but other instantiations are possible (Park et al., 2021). In case a variable has no parents in G, the above reduces to testing the marginal distributions $P^s(X_i)$ for equality, for which we use the Maximum Mean Discrepancy (MMD; Gretton et al., 2012).

As the pair-wise p-values between pairs of contexts are correlated and hence do not allow a well-defined dependency measure, we convert them to a partition to use our MI-based measure. We obtain a clustering naively from the pair-wise tests by including s_i, s_j in the same group if and only if the pair-wise testing does not indicate that $\Pi(s_i) \neq \Pi(s_j)$. Hence, if there are disagreements between the correlated tests, we resolve these in favor of more mechanism changes, although other options are possible depending on the sensitivity of the test. In the bivariate example shown in Figure 6.1, for instance, we obtain partitions Π_X and $\Pi_{Y|X}$ corresponding to the shown changes in $\widehat{P}_X$ and $\widehat{P}_{Y|X}$, which we test for independence as described in the following.

Discovering Confounding Variables

Next, for every pair of variables X_i, X_j, we determine whether it is confounded by computing the p-values of our MI score based on Equation (6.2),

$$p_{ij} = \Phi^{-1}\left(\frac{I(\Pi_i, \Pi_j) - E\left[I(\Pi_i^\star, \Pi_j^\star)\right]}{\sqrt{\mathrm{var}(I(\Pi_i^*, \Pi_j^*))}}\right),$$

where Φ is the cumulative density function of the standard normal distribution. In the second stage, we aim to discover those subsets of variables affected by the same latent variable. While our theoretical analysis suggests considering the total correlation over variable subsets, performing such a test for every given subset $X_\mathcal{I} \subseteq X$ is both infeasible and results in a multiple testing problem involving enormous numbers of tests. We therefore infer confounders directly from pairwise tests by using Assumption F directly.
If our tests for discovering causal mechanism shifts and confounding were perfect, variables affected by the same confounder would form distinct clusters with high pairwise MI, which could be used as direct estimates of confounded variable sets. In practice, we will find some variable pairs to be incorrectly judged (un-)confounded. While this is the same issue we faced with CoCa in Chapter 2, we cannot use the same solution here. Since we are not explicitly modeling any distributions here, we cannot fit a latent confounder Z to minimize a score function. Instead, we cluster the pairwise mutual information terms directly. More precisely, we consider the affinity matrix J with entries $J_{ij} = I(\Pi_i, \Pi_j)$, using MI as pairwise similarity, and use spectral clustering (Donath and Hoffman, 1972) to discover strongly connected components. The resulting clusters $X_\mathcal{I}$ are then likely subject to the same confounder.

CoCo

To summarize, we present the pseudocode for COCO in Algorithm 6.1, and an illustration in Fig. 6.2. In the first phase, for each variable, we test all pairs of contexts for mechanism shifts (l. 1–4) in order to obtain its partition. In the second phase, we test all pairs of variables for confounding (l. 5–6). Last, we cluster the variables into subsets affected by the same confounder (l. 7–8).
Regarding the complexity of our method, shift testing is in $\mathcal{O}(|G| \cdot |\mathcal{S}|^2)$, testing for confoundedness in $\mathcal{O}(|G|^2)$, plus spectral clustering in $\mathcal{O}(|G|^3)$.

6.4 Independent and Related Work

It has been commonly acknowledged that special care must be taken when datasets from multiple sources are to be combined to derive a common set of

Algorithm 6.1: COCO

input : Data over X, S; causal DAG G
output : Subsets of X that are jointly confounded by a latent variable Z_j

foreach variable X_i **do**
  foreach pair of contexts s, s' **do**
    $p_{s,s'} = p\text{-val}\left(P^s(X_i \mid \mathrm{Pa}_i) \neq P^{s'}(X_i \mid \mathrm{Pa}_i)\right)$
  Convert $\{p_{s,s'}\}$ to a partition Π_i
foreach pair of variables X_i, X_j **do**
  $p_{ij} = \Phi^{-1}\left(\left(I(\Pi_i, \Pi_j) - E\left[I(\Pi'_i, \Pi'_j)\right]\right) / \sqrt{\mathrm{var}(I\left(\Pi'_i, \Pi'_j\right))}\right)$
Construct an affinity matrix J
Discover subsets $X_{\mathcal{I}}$ of X that are connected components in J, using spectral clustering
return Confounded subsets $X_{\mathcal{I}}$

statistical or causal estimates (Dahabreh et al., 2020). When data is combined naively, the resulting causal model can be *worse* than if the data had only been considered in isolation (Compton et al., 2023).

When done correctly, however, combining observational and experimental data can lead to improved causal networks (Kallus et al., 2018; Kocaoglu et al., 2019). However, these methods are generally restricted in their ability to rule out or corroborate the existence of latent variables due to scarcity of experimental data. Furthermore, experimental data is not in fact unbiased (Deaton and Cartwright, 2018; Naci and Ioannidis, 2013; Averitt et al., 2020), leading to the failure of these methods (Statnikov et al., 2015; Colnet et al., 2024; Cheng and Cai, 2021; Kladny et al., 2023) Other approaches attempt to study the stability of parameters under confounding and selection (Oster, 2013) and develop methods that are robust to future mechanism changes (Shen et al., 2023).

There is also a growing literature on relaxing the i.i.d. assumption in causal discovery, showing that one can obtain stronger identifiability results by using the information inherent in distribution shifts of observed variables (Zhang et al., 2017; Rothenhäusler et al., 2019; Huang et al., 2020; Mooij et al., 2020; Gamella et al., 2022; Mey and Castro, 2024). Recent approaches leverage the independent change (Mameche et al., 2023) and sparse shift principles to discover fully directed causal DAGs from multiple environments, such as the Mechanism Shift Score (MSS; Perry et al., 2022).

The aforementioned approaches consider an exogenous context variable, which can be viewed as a special form of confounding (Huang et al., 2020). However,

in practice, not all confounding can be fully explained by the effects of the environment. For example, when confounding effects are genetic, then while differences in the values of the confounder can be partially explained by membership of a subpopulation, the variance within any subpopulation is still large. That is, there may still be a confounder within each context. Most related to our method is the Joint Causal Inference (JCI) framework (Mooij et al., 2020) when instantiated with a discovery algorithm that does not require sufficiency, such as FCI (Spirtes et al., 2000). Other related work includes those which propose mutual information estimators to estimate the similarity of distributions (Reddy et al., 2022), as well as the work by Karlsson and Krijthe (2023) also address violations in exchangeability under latent confounding (Guo et al., 2024) but focus on causal effect estimation under a fixed graph structure.

6.5 A Shift in Focus: Experiments

To conclude, we empirically evaluate COCO on synthetic and real-world data. We implemented COCO in Python and make all code available online.[1]

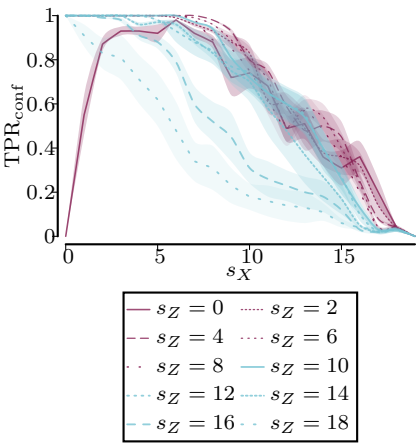
CoCo and Oracles

To separate the effects of discovering latent variables, mechanism changes, and causal directions, we include different oracle versions of COCO. To study our confounding test in isolation, we consider an oracle for the true partitions, named COCO-Π^*. We combine it with mechanism shift testing in COCO-G^*, which takes the causal structure G^* as background knowledge. Finally, we combine our approach with MSS (Perry et al., 2022) using the kernelized conditional independence test (Zhang et al., 2011) to discover a fully directed DAG G. As MSS starts from a Markov Equivalence class, we provide all methods, including all competitors, with the correct MEC as a starting point.

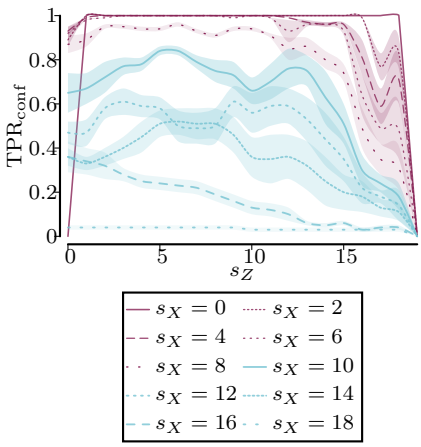
Competitors

Our main competitor 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 causal, confounded, potentially confounded, or none of the above. We also apply FCI to the pooled data from all contexts, FCI-$\mathcal{S}$, and to the data of each context individually, reporting the best such result, FCI-s^*.

[1] `https://eda.rg.cispa.io/prj/coco/`



(a) Effect of increasing the number of shifts of the observed variables.



(b) Effect of increasing the number of shifts of the confounder.

Figure 6.3: *Recovery depends on the number of mechanism shifts of observed and latent variables.* We show the power of our confounding test over pairs of nodes (higher is better) depending on the observed mechanism shifts s_X (left) and latent shifts s_Z (right) over $n_s = 20$ 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 ($s_Z > 15$, blue plots on the left) or does not change at all ($s_Z = 0$).

Synthetic Data Generation

Following Huang et al. (2020), we generate an Erdős-Rényi graph G^* with edge probability p, and generate data from the model,

$$X_i^{(s)} = \sum_{j \in \mathrm{Pa}_i^*} \omega_{ij}^{(s)} f_{ij}\left(X_j^{(s)}\right) + \sigma_j^{(s)} N_j^{(s)}\,, \tag{6.3}$$

with weights $\omega_{ij}^{(c)} \sim \mathcal{U}(0.5, 2.5)$, uniform or Gaussian noise with equal probability, and functions f sampled uniformly from $\{x^2, x^3, \tanh, \mathrm{sinc}\}$. For each mechanism change, we re-sample from Equation (6.3). Finally, for our confounders Z_j, we remove source nodes with edges to at least two variables.

6.5.1 Importance of Sparse Mechanism Shifts (Assumption C)

To begin with, we study the influence of our key assumption, the sparse mechanism shift hypothesis, as we formulated it in Assumption C (Guo et al., 2024; Schölkopf et al., 2021). It states that distribution changes result from only a

small number of changes in causal mechanisms. This is consistent with the view of causal mechanisms as independent modules that do not influence each other and is closely related to the invariance principle whereby causal mechanisms remain the same even in different contexts (Peters et al., 2016; Huang et al., 2020). While sparsity has recently been proposed as a relaxation of the i.i.d. assumption (Perry et al., 2022), it is not easily testable in practice. Hence, we want to empirically investigate how sensitive our confounding test is to an increasing number of causal mechanism changes.
To this end, we we vary the number of changes for the observed (s_X) and latent variables (s_Z) in a fixed set of contexts, here $n_s = 20$. We generate data as in our main experiments and test with COCO-Π^* for confounding between all node pairs in a causal DAG. To show the empirical power of our confounding test, we show the true positive rate (TPR_{conf}) over these decisions in Figure 6.3.

Observed Shifts In Figure 6.3a, we show the effect of increasing s_X. We run the experiment for each s_Z, and color plots red if latent shifts are sparse ($s_Z < 10$) and blue otherwise. We observe a tipping point at $s_X = 9$ where the observed nodes have partitions with ten different groups of 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. Note that in the special case where all variables are identically distributed in all contexts, $s_Z = 0, s_X = 0$, the confounding effect is not measurable using our method.

Latent Shifts In Figure 6.3b, we show the same result when we increase s_Z instead. Sparse shifts with $s_X < 10$ are again colored red, and dense shifts blue. We can see a clear separation between the two cases, which confirms our observations above. In particular, under sparse shifts of s_X, we can tolerate up to $s_Z = 15$ shifts of the confounder.
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. Ideally, both numbers are in an intermediate range.

6.5.2 Empirical Significance and Power (Lemma 6.1)

Next, we revisit Lemma 6.1, which guarantees a power of 1 of our test as we observe more contexts, $n_s \to \infty$. To give a more practical result for fewer contexts, we investigate the power and significance of our test empirically.
We consider COCO-Π^* to study our confounding test in isolation and show true positive rates (TPR_{conf}) and false positive rates (FPR_{conf}) to show the power, respectively significance, of the test. As in our main experiment, we test for confounding between all pairs of nodes in a causal DAG and consider $m = 10$ nodes in $n_s = 10$ contexts, where one confounder influences a random set of

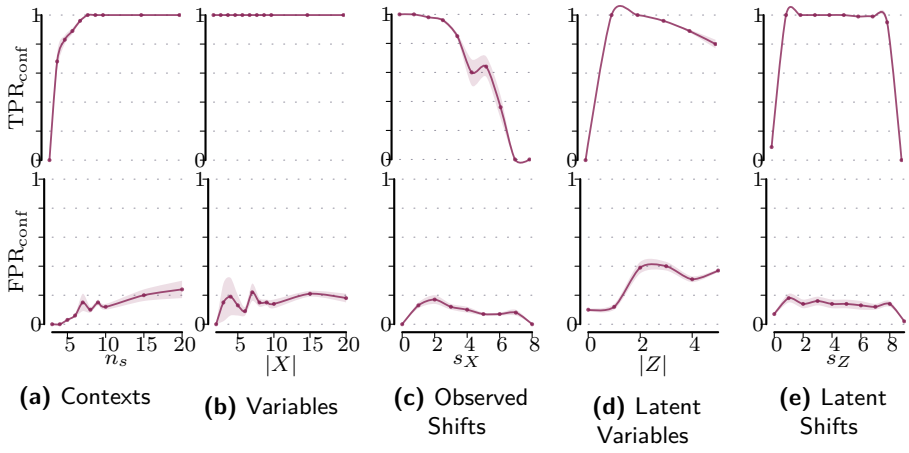
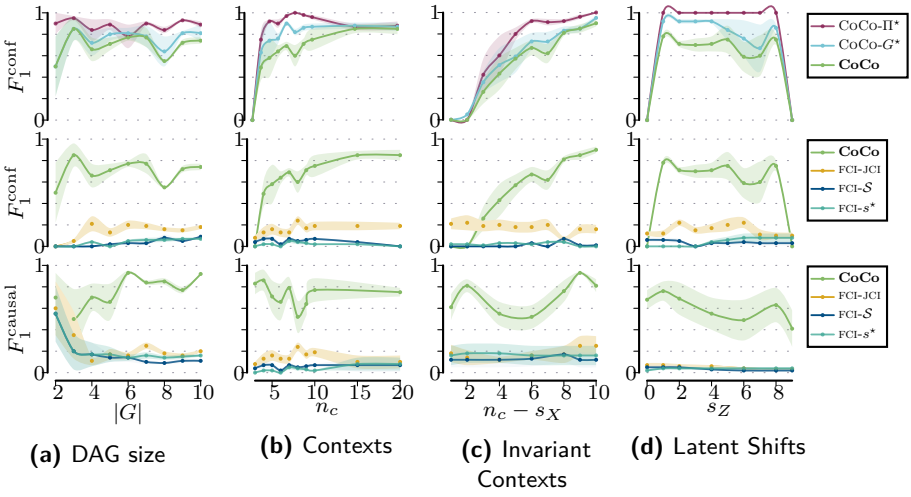


Figure 6.4: *Empirical Power and Significance.* We show the true positive rate (top, higher is better) and false positive rate (bottom, lower is better) of our confounding test depending on the number of contexts, variables, confounders, and mechanism shifts, starting from $n_s = 10, |X| = 10, |Z| = 1, s_X = 1, s_Z = 2$.

between two and m nodes, and where nodes undergo $s_X = 1$ mechanism change and the confounder $s_Z = 2$ changes. We show the results in Figure 6.4. We consider up to $\frac{m}{2} = 5$ confounders because each confounder always affects at least two variables, and up to $n_s - 1 = 9$ mechanism changes because this corresponds to a change in every context.

Power We find that our test already works well with few contexts, with perfect power starting from $n_s = 8$ contexts (Figure 6.4a). We point out the special case $s_Z = n_s - 1$, where the confounder changes in *every* context. In this case, the (adjusted) mutual information of the observed (single-group) is zero, and we cannot detect confounding, as we can see for $n_s = 3$ (Figure 6.4a) and $s_Z = 9$ (Figure 6.4e). Otherwise, the number of latent shifts does not significantly impact our results (Figure 6.4e), and only the shifts of the observed variables do (Figure 6.4c), as we discussed above. The sensitivity of our test is not affected by the number of variables (Figure 6.4b) and decreases slightly when we add more confounders to the system (Figure 6.4d).

Significance As the false positive rates show, our test rarely detects unconfounded variable pairs as confounded, with FPR$_{\text{conf}}$ remaining around 0.1 and below 0.2 in almost all experiments. We notice a change when there is more than one confounder (Figure 6.4d). To explain, in this case, we also check



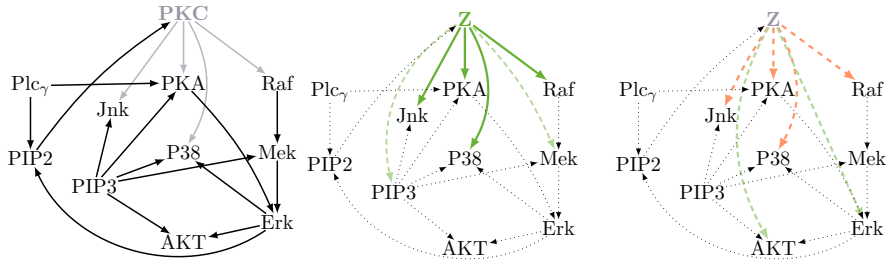
(a) DAG size **(b)** Contexts **(c)** Invariant Contexts **(d)** Latent Shifts

Figure 6.5: *Detecting Confounding and Causality with* COCO. We evaluate COCO on discovering confounding in DAGs G over multiple contexts. We compare (top) COCO with MSS and the KCI test (green) to oracle versions that start from the true partitions $\Pi^\star$ (purple) respectively the fully directed DAG $G^\star$ (blue). We compare (middle, bottom) against JCI-FCI (yellow), FCI-$\mathcal{S}$ on pooled data, and FCI-$s^\star$ per context (blue). We report F_1 scores computed over each pair of nodes, evaluating whether it is confounded (top, middle), respectively causally related (bottom).

whether variables are affected by the *same* confounder, and our method may discover a variable pair X_i, X_j as confounded when they are each affected by a *different* confounder, $Z_k \to X_i, Z_l \to X_j$. In particular, this happens if Z_k, Z_l have joint mechanism shifts coincidentally, in which case the mechanism shifts of X_i, X_j also appear correlated. However, even in this case, FPR_{conf} remains relatively low (Figure 6.4d), suggesting that our method can mostly separate which variables are affected by which latent variable.

6.5.3 Detecting Confounding with CoCo

To begin with, we consider whether COCO discovers confounding in a multi-context DAG G. We do this for varying parameters, including the number of contexts (n_s), the number of observed (m) and latent (n_Z) variables, and the number of observed (s_X) and latent (s_Z) mechanism shifts. Unless otherwise indicated, we use the parameters $n_s = 10, m = 10, n_Z = 1, s_X = 1, s_Z = 2$.
We first perform an ablation study on COCO, based on the kind of oracle it has access to. We show our results in the top row of Figure 6.5. As we expect from our theory in Section 6.2, COCO works best for more contexts, larger numbers of



(a) Ground Truth **(b)** COCO **(c)** JCI-FCI

Figure 6.6: *Confounding effects of PKC.* On the Sachs et al. (2005) data COCO recovers confounding effects of PKC on most variables (solid green), in addition to two extra variables (dashed green). In contrast, JCI-FCI misses all of the truly confounded variables (dashed orange), while finding only two spurious ones instead.

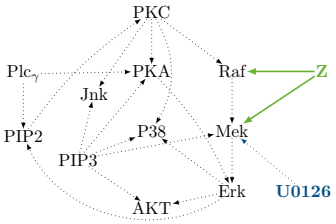
invariant environments for each variable, and an intermediate number of latent mechanism shifts. In contrast, when mechanism shifts are dense, i.e., when $n_s - s_X = 1$ or $s_Z = n_s - 1$, or when there are no shifts, we cannot discover any confounding. Overall, the gap between the oracle versions and the version in which everything needs to be inferred is small in this experiment. This suggests that confounding detection works well even when causal directions are unknown, supporting our results of Section 6.2.5.

Next, in the middle row of Figure 6.5, we see that COCO (green) clearly outperforms JCI-FCI (yellow) by a large margin in discovering confounders. At the same time, JCI-FCI, in turn, has a slight advantage over FCI on best-scoring single-context respectively pooled data (blue). As the FCI variants can only determine potential confounding for pairs of nodes, we evaluate confounding decisions across pairs of nodes in G using F_1-scores.

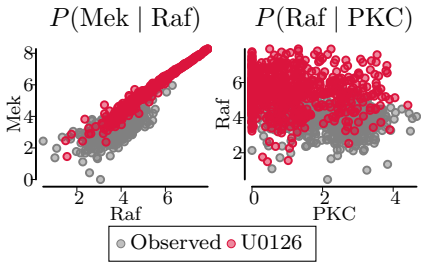
Last, we compare how many of the causal edges are correctly directed in the bottom row of Figure 6.5, starting from a given Markov equivalence class. As expected, we do well under sparse shifts and with more contexts, while all the versions of FCI generally only discover few causal edges.

6.5.4 Real-world Cell Signaling Data

We end with a case study on the flow cytometry dataset by Sachs et al. (2005). It contains samples of eleven protein and phospholipid components in human immune cells, studied under different molecular interventions. To study confounding effects, we start from the consensus causal network in Figure 6.6. As in Section 4.5, we keep PKC hidden and use the data over the remaining variables in the nine different contexts included in the data.



(a) Consensus Network, intervention with U0126, and confounder discovered by COCO



(b) Joint Mechanism Shift of Raf and Mek discovered by COCO

Figure 6.7: *Hidden confounding between Raf and Mek.* In the cell signalling network (Sachs et al., 2005), COCO discovers confounding between the molecules Raf and Mek (a). Although the consensus network only contains the edge Raf $\to$ Mek, many causal discovery methods also report a pathway Mek $\to$ Raf (Perry et al., 2022), suggesting that there may be feedback. We illustrate this in (b), where we show the data in the observational context (gray) and in an interventional context (blue) where the reagent U0126 was added (Mooij et al., 2020). While U0126 is presumed to only directly influence Mek, we see a change in the abundance of Raf. With COCO, we discovered a joint mechanism change of both conditionals $P(\text{Raf} \mid \text{PKC})$ and $P(\text{Mek} \mid \text{Raf})$ in the interventional context, and overall found that partitions for Raf and Mek are correlated.

As we illustrate in Figure 6.6, COCO correctly discovers a confounder Z and all of its outgoing edges (green) as well as two spurious ones (dashed). While the edge between the confounder Z and PIP3 is indeed spurious, we study the relationship between Raf and Mek in more detail below. In contrast, JCI-FCI discovers multiple false positive confounded edges.

Is Mek Really Spuriously Confounded?

To check whether the results returned by COCO are spurious, we repeat the experiment while keeping each node hidden in turn. Overall, COCO returns few confounding effects that disagree with the consensus network. Notably, we always discover Raf and Mek to be confounded, suggesting the possibility of unmeasured confounding even in this highly controlled study.
Mooij et al. (2020) already discussed the relationship between these signaling molecules in detail as an example, suggesting that the consensus network may be incomplete. As shown in Figure 6.7, this network includes the pathway Raf $\to$ Mek, and the only intervention targeting either of the molecules is the Mek inhibitor U0126 (Sachs et al., 2005). Consider the data shown in Figure 6.7b. We show the observational context (gray) and the interventional context where the reagent U0126 was added (red) and can see that there is a distribution shift

of Raf under U0126. This suggests that either U0126 also targets Raf or that there is a feedback loop between Mek and Raf (Mooij et al., 2020).
We found that COCO detects this observation. In the partitions for Mek and Raf, reflecting changes in the conditional distributions $P(\text{Mek} \mid \text{Raf})$ and $P(\text{Raf} \mid \text{PKC})$, we discover a joint mechanism shift of both signaling molecules in the interventional context U0126, and higher than expected mutual information of the partitions, hence deciding that Raf and Mek are confounded. In conclusion, COCO discovers a *dependent* mechanism shift of Raf and Mek under the intervention with U0126, thus pointing to potential hidden confounding between the cells that is consistent with the data.

6.6 E Pluribus Unum?

In this chapter, we tackled Problem 4, determining to what extent we can leverage data from multiple settings, such as observational and interventional contexts, to discover latent confounding non-parametrically.
We began by introducing the necessary framework of causal discovery from multiple environments, in which causal mechanism shifts are assumed to be independent. We then showed that this independence of mechanism shift in the complete set of variables $X \cup Z$ leads to *correlated* mechanism shifts when we observe data only from the distribution $P(X)$. This dependence is precisely what we leveraged in our approach: by measuring the dependence between the partitions induced by mechanism shifts, we showed that we can discover both which variables share the same latent confounder (Theorem 6.2), as well as recover the true causal order among the observed variables X (Theorem 6.5).
From this characterization of the properties of our score, we then derived our algorithm COCO for recovering the sets of jointly confounded variables given the MEC. We showed that it works well on synthetic and real-world data.
For the future, one relevant question is whether can we obtain better results by making better use of the shared mechanism shifts across multiple variables. That is, if l variables all share the same latent confounder Z, then they all jointly share the same mechanism shifts induced by shifts in Z. Right now we cluster variables based on the pairwise structure, but it would be interesting to see whether we can algorithmically extract larger structures directly.
Another interesting avenue is whether we can use what we have learned so far to *deconfound* our observed variables. Once we have determined that a set $X_{\mathcal{I}}$ are jointly confounded by a latent factor Z, can we in some way determine and account for the joint mechanism shifts induced by this latent variable Z, and determine the true causal graph over the observed X?
A related question is whether we can adjust our causal mechanisms to account for such knowledge. Consider the example of observed variables U, X, Y, such that only $X \to Y$ and all three variables are jointly confounded by a latent Z.

Clearly, if we knew both the true causal structure over U, X, Y, and that they are jointly confounded, we should be able to use U as a proxy variable for Z to adjust whatever causal mechanism $X \to Y$ we fit. It is less clear, however, to what extent this is still possible in more complex causal networks.

Chapter 7

Conclusion

Alice: Would you tell me, please, which way I ought to go from here?
Cheshire Cat: That depends a good deal on where you want to get to.
Alice: I don't much care where.
Cheshire Cat: Then it doesn't much matter which way you go.
Alice: ...So long as I get somewhere.
Cheshire Cat: Oh, you're sure to do that, if only you walk long enough.

Lewis Carroll, Alice in Wonderland

In the last five chapters, we have tried to answer the question to what extent we can perform causal discovery in the presence of unobserved latent variables that create biases in the distribution of our observed variables. Such biases can be created either by not conditioning on variables that should be conditioned on—latent confounding—or by conditioning on variables that should not be conditioned on—selection bias. In particular, we tried to answer whether we can discover such biases and studied the conditions under which this is possible and to what extent we can simultaneously learn a causal model over both observed and unobserved variables despite the incompleteness of our data.

In this concluding chapter, we will summarize the contributions of this thesis and relate them to the wider landscape of the field of causality in its current state. In particular, we extrapolate some lines of current development in causality and machine learning and see where our ideas may be of use.

7.1 Biases in Causal Learning

The first problem we tackled was the basic question of to what extent we can distinguish between causally related and jointly confounded variables *at all*.

Problem Statement 1 (Confounded or Causal)**.** Given covariates X and a target variable Y, does X cause Y, or are the correlations best explained due to latent confounding caused by not controlling for an unobserved variable Z?

To answer this question and to provide a generically useful tool with which to think about the effects of latent confounding more generally, we extended the algorithmic model of causality to allow for explicit inclusion of latent variables in the lowest complexity causal factorization of our observed distribution. In particular, we made explicit the reason *why* the algorithmic independence of causal mechanisms is violated in the presence of latent confounding.
We then showed that by constraining the Kolmogorov complexity to a subset of all possible distributions, we can use the MDL principle to form a theoretically and statistically well-founded upper bound for model classes both in- as well as excluding latent confounders. This provides us with a principled approach for deciding whether the data is, in fact, better described by endogenous causal relations or by the effects of latent factors.
We showed empirically that the number of observed variables relative to the number of latent confounders is a crucial factor in distinguishing between these two classes of models: when too few covariates are observed, it becomes impossible to distinguish between causal and confounded models. It turns out that while the common adage "just collect more data" is misguided in its usual sense that controlling for more covariates will somehow provide reasonable causal estimates, it is true in that collecting more covariates permits us to more accurately determine whether latent confounding is a concern in the first place.
The relationship between the number of observed variables and the number of latent confounders, as well as the structure of how the latent confounders affect the covariates, became clearer still as we turned to our second question: to what extent can we learn the causal model over both observed and latent variables when we never observed the latter?

Problem Statement 2a) (Causal Discovery with Hidden Confounders—Linear Case)**.** Given data only for the observed variables X, and assuming that all causal relationships between both the unobserved Z and the observed X are purely linear, can we discover a joint causal network over both the observed variables X as well as the unobserved Z?

By studying the structure of the correlations between observed variables, we determined that in the case where latent confounding affects at least four variables, we can distinguish between joint latent confounding and causal connec-

tions between the observed variables. It turned out that in the case of joint latent confounding and some assumptions of structural sparsity, the correlations of the observed variables co-dependently lie on a low-dimensional manifold, permitting us to exploit violations of the independence of causal mechanisms. Therefore, the joint causal model over observed and unobserved variables is identifiable even when we do not know the true number of latent confounders. We then showed that, perhaps surprisingly, with no need for additional assumptions beyond linearity and structural sparsity of the causal graph, both the BIC as well as MDL-derived scores are consistent for recovering the true causal model over both observed and unobserved variables.
To learn the desired causal model over both observed and latent variables, we propose an elementary approach that can be applied in essentially any score-based causal discovery framework. We start by running an (arbitrary) causal discovery algorithm, resulting in a proposed causal graph over the observed variables. By considering our proposed MDL score, we find latent factors for subsets of the observed variables, which reduce the total score, and add that latent factor, which minimizes the overall score. This yields an extended set of variables, over which we then run the same causal discovery algorithm once more. By iterating this process, we show that if our model assumptions hold, we are guaranteed to find the correct set of confounded nodes. Empirically, we further show that our approach works well on both linear and nonlinear, synthetic and real-world data. As before, we observe that the number of observed and latent variables plays a crucial role in our ability to recover both the confounded nodes and the complete causal graph.
Since the assumption of strict linearity is a strong one, a natural next question was whether we can include nonlinear causal relationships between variables.

Problem Statement 2b) (Causal Discovery with Hidden Confounders—Nonlinear Case)**.** Given data only for the observed variables X but not the unobserved variables Z, do nonlinear causal mechanisms exist for which we can discover a joint causal network over X and Z?

We showed that within the framework of PNL causal models, in which the nonlinear relations between observed variables arise from independent nonlinear transformations of each of the observed variables, this is indeed possible. The essential insight here is that due to the nature of the nonlinearities, there exist element-wise nonlinear transformation ϕ of the observed variables such that the "causal" graph over the variables $\phi(X)$ is fully linear so that the previous chapter's results for identifiability apply under minor additional assumptions on ϕ^{-1}—smoothness and strict nonlinearity.
While identifiability is straightforward, given our preliminary work, learning the true causal network requires a different approach. Since we need to disentangle the nonlinearity ϕ and the causal structure induced by the linear relations

between observed and latent variables, we use a VAE-inspired structure to learn the causal model. We show theoretically that in the linear case, this approach is consistent for network discovery, and empirically that it works both for linear and nonlinear synthetic and real-world data.
Next, we took a detour to consider another source of bias that crops up repeatedly in questions of causal discovery and inference: selection bias. This problem is, in a certain sense, *dual* to that of latent confounding: instead of not conditioning on a variable that is causally upstream of the observed variables, we do condition on a latent variable that is causally downstream of the observed variables. The question of to which extent we can determine whether variables are affected by selection bias was, therefore, a natural next consideration.

Problem Statement 3 (Dealing with Selection Bias)**.** Given data for the observed variables X, are they causally connected, or are the correlations best explained by selection effects due to conditioning on an unobserved collider Z?

We obtained a positive answer to this question both in the setting of parametric families as well as in the setting of non-parametric distributions subject to certain known classes of invariances. We showed that, just as confounding results in specific patterns in the correlation structure of the observed data, selection bias results in specific patterns in the observed data distribution. In a certain sense, selection effects are *local* in that they significantly affect only a part of the distribution. By looking at the relatively unaffected part of the distribution, we can still learn the correct parameters, or invariances, for the underlying distribution and exploit local deviations from the entailed structure to infer which parts of the space are affected by selection.
We showed theoretically that both approaches result in identifiable causal models even in the face of selection bias and empirically that both models work well on synthetic data and can provide us with novel insight into real data.
After this excursion into the land of selection bias, we returned once more to the topic of latent confounding. But this time around, we leveraged ideas from the literature on causal discovery given data from multiple environments.

Problem Statement 4 (Confounding across Contexts)**.** Given data only for the observed variables X across multiple contexts $c \in C$, but not the unobserved variables Z, how readily can we determine which variables are jointly confounded by the same latent factor $Z_i \in Z$?

We show that in this setting, we require no parametric assumptions on the form of the true distribution nor assumptions on the sparsity of the underlying causal graph. Instead, using only relatively weak assumptions on the frequency of mechanism shifts across environments, we show that we can determine with high probability which variables are jointly confounded and recover the true causal graph over both observed and unobserved variables.

The basic idea is that changes in the distribution of a latent variable across environments lead to *correlated* changes in the apparent causal mechanisms between observed variables. This is, once more, another kind of violation of the independence of causal mechanisms, this time across environments. Combined with well-known results indicating that the number of causal mechanism shifts is minimal in the true causal graph, we show that these correlated mechanism shifts suffice both theoretically and empirically to recover the true network.

7.2 Making Observation Great Again

As a deluge of data has brought on a slew of new statistical and machine learning models, with some theoretical as well as empirical results on the performance of such models as data and compute are scaled, along with jobs such as "data scientist", "data engineer", "prompt engineer", and a host of others, our understanding of the applicability and suitability of these methods for any given task is severely lagging behind. Whether it is recommender systems, credit score models, cancer prediction from X-ray images, macroeconomic models, genome-wide association studies, or any other machine learning model, practitioners often *implicitly* treat the predictions made as causal. It is quite common, even for statistically well-trained scientists, to run purely correlational analyses, claim that these are by no means causal, only then proceed in their discussion section to speak of their results as if they were, in fact, causal (Shapiro, 2004; Rutter, 2007).

Practitioners may often attempt to verify the accuracy of such causal interpretations by controlling for additional observed covariates in an attempt to remove the effect of latent confounding. However, as we noted in the introduction, controlling for a variable implies (some) knowledge of the underlying causal graph. That is, the validity of controlling for a variable relies fundamentally on that variable being neither mediator nor collider for the predictors and the outcome (Wysocki et al., 2022). However, these assumptions are rarely verified, nor do most practitioners employing these kinds of controls realize that there are assumptions to verify in the first place.

While we can consider this as practitioners simply making the best of a bad situation, the question remains: how can we do better? One line of answers is given by the classical econometrics literature on estimating treatment effects. This usually involves doing RCTs (such as A/B tests; Kohavi et al., 2020; Austrian et al., 2021), finding instrumental variables or proxies (Angrist et al., 1996), finding apropriate natural experiments (Rosenzweig and Wolpin, 2000), doing regression discontinuity designs (Imbens and Lemieux, 2008), employing differences in differences (Card and Krueger, 1993), or using any one of a number of other approaches (Angrist and Pischke, 2009).

In a second, related line of thinking, one tries to combine data from different sources. The most common approaches in the literature on this are combining observational data with more controlled data, e.g., RCT data or data from other study designs outlined above, and combining the effects obtained from these varying data sources to obtain the true, or at least a less biased, causal estimate. However, one fundamental assumption often made by both this and the previous line of thought is that the data obtained from the RCTs (or other experiment designs) is unbiased (Statnikov et al., 2015; Colnet et al., 2024; Cheng and Cai, 2021; Kladny et al., 2023). In reality, however, this assumption is commonly violated due to stringent selection criteria on the participants of the RCTs (Naci and Ioannidis, 2013; Averitt et al., 2020), or limited variation in experimental setups rendering causal estimates derived from these sources incomparable to effects we would find in the wild. As such, the common motivation to combine *unbiased* RCT data with *biased* observational data in order to obtain lower variance estimates at the cost of larger bias is fundamentally misguided from the start. While other attempts, such as causally valid meta-analyses of RCTs on different populations, have been made (Wiernik and Dahlke, 2020; Dahabreh et al., 2020; Markozannes et al., 2021), they do not leverage the vast amount of available observational data available to us.

The third line of approaches is to understand what makes causal identifiability from purely observational data impossible in general and, conversely, which assumptions *do* make at least partial identifiability possible. Being the line we have followed in this work, we believe it to hold potential both theoretically as well as practically. That is, as we have shown in Chapters 2, 3, and 4, there are distinct patterns to the effects of low-dimensional latent confounders, which are fully generic within the algorithmic model of causality, and which we can detect from observational data in the linear and specific nonlinear cases. We expect that the study of more general patterns will reveal that confounding may be detected under relatively general assumptions for sufficiently rich datasets.

One interesting connection between the second and third lines of thought is to what extent data combination can be leveraged not only between unbiased and biased datasets but also between multiple observational datasets biased *in different ways*. These datasets may be studying the same effects in different settings, across different populations, in different locations, or at different points in time. It has been shown that simply combining datasets from different environments can be detrimental to a model's performance in theory and practice (Compton et al., 2023). The idea is that so long as either the true underlying causal mechanism stays the same and the *sources* of bias change—e.g., observational data and RCT on a subset of the same population—or the sources of bias stay the same but may vary in distribution, while the underlying causal mechanisms can change, then data from such multiple contexts or environments allow us to identify the common underlying structure (Bareinboim and Pearl,

2016). Unfortunately, such approaches are, as of yet, uncommon in the literature, and data merging approaches generally focus on the case where causal sufficiency holds in all, or at least some, of the datasets used to estimate the causal mechanisms. Two exceptions to this dreary lack are the work by Karlsson and Krijthe (2023) and our work in Chapter 6. In particular, Karlsson and Krijthe (2023) focus on violations in exchangeability (corresponding to our Assumption A in Chapter 6), which are entailed in the case of causal sufficiency, but they focus on the effect of a single treatment on a single outcome variable. In contrast, in Chapter 6, we focused on fully non-parametric identifiability of the underlying causal network and showed that under relatively weak assumptions on the mechanism changes, we can recover the true causal network with high probability. In the next section, we outline some more related and different lines of thought as they may affect the field of causality in the future.

7.3 Some Speculation on the Future of Causality

Given how fundamental causality is to our scientific endeavor as a whole, it is heartening to see many sub-fields of causality thriving. However, many of these novel fields are still in their infancy and have yet to grapple with the biases that our analyses are otherwise sensitive to.

7.3.1 Biases in (Causal) Representation Learning

Much has been written about learning good representations, for example, in the context of image recognition. A typical example is the cow on a meadow versus a cow on the beach. Since datasets on which our model is learned will generally contain many images of cows on meadows but few images of cows on beaches, many image recognition models will do a good job detecting the cow in the first but not the second setting. More likely, it will output that the cow is some other kind of animal that is more likely to exist on a beach. In this case, we say that the representations of the cow and its environment are *entangled*. From a purely predictive point of view, this entanglement is only natural. After all, if the presence of a cow is predictive of a meadow, it would be inefficient not to use this information. However, from a perspective of not only human interpretability but also of *carving nature at its joints* (Plato, 1952), a representational cow-meadow mixture is hardly very appealing, even though in many respects, human cognition works the same way (Köhler, 1967). That is, what we desire are *disentangled* representations in which, in this example, foreground (the cow) and background (the meadow or beach) are represented independently of each other (Wang and Jordan, 2021).
Many papers have been written about how to resolve this issue, whether it be via interventions (Ahuja et al., 2022), Hausdorff factorized supports (Roth

et al., 2022), or one of many other methods (Wang et al., 2022), as well as results on the theoretical impossibility of the task without further constraints (Locatello et al., 2019; Träuble et al., 2021). Only few, however, have written about the underlying reasons *why* such entangled representations occur in the first place. The correlation between the presence of cows and meadows is not a bug, it is a feature of the real world. The fact that many images of cows on meadows but few on beaches are included in the data is not a fact to be wished away; it should instead be modeled as part of the data-generating process.

We believe that this can, and should, be fruitfully modeled in terms of selection bias. Of course, this is not the same kind of selection bias we have been concerned about, where preferential inclusion of some images causes biases in the outputs. In contrast, such a preferential selection to remove correlations between the presence of cows and meadows is, in fact, precisely the goal of many researchers in the field of representation learning. That this is misguided is easy to see by analogy with a bivariate linear regression: if we desire to regress Y on X_1, X_2, but these two variables are correlated with each other, then restricting the range of X_1, X_2, Y to a part of the distribution in which X_1, X_2 are not correlated will not solve the problem but instead introduce uncontrolled bias in the regression parameters. Rather, cows are *in fact* more often found on meadows than on beaches, and this would be true even without human intervention (such as animal farms). The selection mechanisms here are natural selection and the ecological niche to which cows are adapted.

The disconnect between disentangled representations and reality is the enforcement of positivity where it does not apply. Even in a world where there are literally zero cows on beaches, perhaps this is how it *should* be. Enforcing positivity on the (cow, beach) probability would fundamentally misjudge the underlying reality of the world that we are trying to represent in the first place. It is, therefore, essential to distinguish between what we are trying to *represent* and what we are trying to *generate*. Instead, we should model the presence of entangled representations by jointly fitting a disentangled representation and the selection mechanisms giving rise to the observed non-positivity of the (cow, beach) pair. This solution has the benefit that we model both the true distribution while also providing us with disentangled representations that can be used to interpret the latent space and generate new samples from it. It might also give us a more readily interpretable model as to *why* the observed entanglements occur in the first place, and thus help us determine whether there is concern about the fairness of our representations (Zemel et al., 2013).

In the last years, instead of disentangled representation learning, another idea cropping up is that of *causal* representation learning (Schölkopf et al., 2021). Given some data, such as an image, can we find the causal factors underlying the data? One of the most promising areas for these kinds of representations is biology. We can display the distribution of RNA across many different cells

in terms of an image collected via RNA-seq (Wang et al., 2009) or for single cells via scRNA-seq (Butler et al., 2018; Stuart et al., 2019). Of course, the individual pixels of the image do not cause each other; instead, underlying all the pixels are genes determining the expression levels of proteins, which in turn regulate other genes. The question that causal representation learning asks in this context is, therefore, to what extent can we recover the gene regulation network from the observed image (Squires et al., 2022b; Sturma et al., 2023)? As stated, this task is underspecified, but theoretical results indicate that given enough types of interventional data, we can recover the underlying causal graph (Squires et al., 2022a; Ahuja et al., 2022, 2023; Liang et al., 2023a; Buchholz et al., 2024; von Kügelgen et al., 2023). This makes the biological domain well-suited for this: by using CRISPR (Barrangou and Doudna, 2016), we can intervene on *specific* genes and observe how the resulting image changes (Squires et al., 2022b; Zhang et al., 2023b). Furthermore, gene expression can be measured in multiple ways, in individual cells or in aggregates, allowing us to leverage multimodal observations to improve the identifiability of the underlying causal network (Trask et al., 2022; Sturma et al., 2023). In many of these identifiability results, however, it is assumed that the number of variables in the latent representation is known. What if this is not the case?

How Many Latent Dimensions Do We Need?

While we can give an upper bound on the number of dimensions of the causal representation in the case of gene regulation, given our knowledge about the number of genes for the organism in question, finding a representation for all genes would require too much data, and biological modularity and sparsity of the gene regulation network suggest that this should not be necessary (Leclerc, 2008; Vattikuti et al., 2014). If our goal is to discover which genes are responsible for a specific disease, by comparing the gene expression levels of healthy and sick patients, we expect differences in expression levels to be causally driven by relatively few genes. But how can we determine the correct number of latent variables, and what happens if we use too few or too many latent factors?
The case of too many latent factors appears to be relatively uninteresting with superfluous dimensions modeled as disjoint from the relevant variables, at least in certain domains (Lippe et al., 2023). While this suggests that we should choose "too many" dimensions in the representation, this is quite difficult in that we generally have little basis on which to judge whether a given number of dimensions is large enough. Therefore, it is still important to consider what happens if we do use too few dimensions in the latent representation. This case of too few latent factors is by far more interesting. For example, what might happen if we try to model a system that contains five underlying causal variables using only four variables instead of the correct five?

While we are unaware of any empirical or theoretical analyses of the behavior of the latent representation when too few dimensions are available for causal representation learning specifically, we can draw parallels to the study of latent factors in psychometrics and consider different cases for the behavior of the latent representations. Unfortunately, the study of what, in fact, *does* happen in causal representation learning is beyond the scope of this section. As such, we will consider different cases of what might happen and how they would affect our ability to obtain usable results. The first, most unfortunate, outcome is that the four-dimensional representation is simply a superposition of the original five dimensions. That is, each factor in the lower-dimensional representation is a (non-)linear combination of the higher-dimensional factors. While this may be optimal as far as reconstruction of the original data goes, and in fact, certain kinds of superpositions seem to permit computations *as if all factors are represented* (Elhage et al., 2022), in terms of modeling any "causal" relations in the discovered representation, we would have to consider such a representation to be a complete failure. A second alternative is that only two of the original factors are collapsed into one dimension, which in the psychometrics literature is considered a desirable indicator that the discovered latent factors are, in fact, meaningful constructs (McCrae et al., 1996; Johnson and Bouchard, 2005; Lee and Ashton, 2010; Condon, 2014). In this case, the causal relations between the remaining factors would remain accurate, and the relations between these and the collapsed dimension *ideally* amount to some combination of the original causal relations, although this may not be the case, especially if the collapsed variables do not occupy similar ranks in the causal ordering. The last alternative is that the four-dimensional representation drops one of the true variables without representing it at all and proceeds to learn the appropriate (marginal) "causal" graph over the remaining variables. Clearly, the structure of the resulting causal representations varies dramatically depending on which of the variables is not modeled. That is, while excluding a sink variable or a mediator between precisely two other variables would have relatively little impact on the discovered causal structure, excluding a common parent of other variables would produce latent confounding between the remaining variables and distort the causal structure. By looking at these different cases, we see that the robustness of the causal representations to misspecified dimensions depends critically on the exact outcome of fitting an insufficiently high-dimensional representation. It is, therefore, critical to investigate these avenues in detail, both theoretically and empirically. Furthermore, it is essential to discover ways that permit us to judge whether enough latent dimensions are used, as well as how to determine the correct number of variables in the causal representation. Unfortunately, unlike with standard methods in latent factor models, such as PCA, we cannot determine the "correct" number of latent dimensions by checking the fraction of variance explained at each number of latent variables included

since the dimensions we care about are *not* orthogonal.

Of course, other sources of bias in causal representation learning may occur just as for any other setup. Whether individuals are selected according to certain strict inclusion criteria, or whether they self-select into the data collection, e.g., via sending their data to 23andMe or other biomarker testing kits (Landeck et al., 2016; Goetz and Schork, 2018), or by self-selecting into surveys (Maul, 2017), selection bias is always a concern that needs to be accounted for lest it leads to incorrect causal conclusions. However, while such selection effects are a concern, they are also an opportunity for study, allowing us to investigate to what extent (self-)selection criteria, operating on a phenotypic level, correspond to selection effects in the discovered causal representation, modeled at a genotypic level, and how we might adjust the latter for the former. In particular, it might give us a better understanding of the relationship between phenotypic and genetic correlations (Crespel et al., 2024). In general, the effects of such selection biases depend strongly on the laws governing the relationship between the variables we care about and the variables that are inadvertently selected. These variables are often of different modalities, and thus, we turn to the causal relationships between modalities next.

7.3.2 Multimodality and Consilience

Within the sciences, the most common way to distinguish between correlation and causation is the use of experiments to intervene on parts of a system and see what effects result from this. As such, one of the most important advances in the field of causal discovery and causal effect estimation is the insight that interventions are simply a specific type of distribution shift, and that other non-interventional shifts often already suffice to permit us to obtain better causal conclusions than we otherwise could. That is, by collecting data from multiple different contexts, such as different hospitals, we can direct more of the causal edges and also obtain more insights into the potential presence of latent confounders or other biasing factors (Chapter 6). Of course, some data sources are more easily accessible than others, and optimally combining across these contexts is an open problem.

More broadly, the relative cost and availability of data exist not only across different contexts—such as observational and experimental—but also across modalities. For example, in the case of causal representation learning for gene regulation networks, bulk transcriptome information across millions of cells is much cheaper to obtain (on a per-cell basis) than single-cell transcriptome information, which, due to their nature of destroying the cells, are necessarily measured in different cells (Sturma et al., 2023). Of course, while bulk data is much cheaper to obtain, single-cell data is much more informative about causal effects. Similarly, satellite images can not only supplement sensor mea-

surements to measure earth's weather and climate (Council, 2000; Bi et al., 2023) but also estimate crop yield (Gallego et al., 2014), as well as economic growth (Ahn et al., 2023; Lehnert et al., 2023).

Regardless of the type of modality in which measurements are made, the causal mechanisms underlying these measurements are fundamentally the same. Any causal model that is not consistent with measurements at *all* levels can be discarded. For example, if the predictions of a dynamical system at the micro scale are not consistent with macro scale behavior, such as the temperature of the system, then the dynamical system is known to be wrong. Furthermore, humans' ability to learn from very few examples may be based on the use of domain-general causal models (Goodman et al., 2011).

The combination of such multimodal data is a fundamental problem across sciences: how can we perform meta-analyses on data in which the used constructs are different (Bun et al., 2020; Dahabreh et al., 2020), how do genetic and phenotypic correlations relate to each other (Crespel et al., 2024), what can genetics tell us about psychometric traits, and vice versa (de la Fuente et al., 2021; Kim et al., 2023), how do we combine micro- and macroeconomic models (Imbens and Lancaster, 1994; Jhun, 2021), or perform systems biology across multiple spatial and temporal scales (Dada and Mendes, 2011)?

The first question, then, is to what extent progress in causality, especially causal representation learning, can help us develop methods that can be widely employed to deal with questions of deriving joint models from disparate modalities. Vice versa, it is also important to ask to what extent insight into these problems can help inspire new methods for causal discovery and representation learning. In other words, how can the supervenience structure of the natural world be used to derive better causal models?

More ambitiously, to what extent can knowledge about causal structures in one field of study be transferred to other fields? The general principle that tools found to be successful in one field can often be successfully employed in different fields is a fact that has previously been called *consilience* (Wilson, 1999). Some examples include the use of tools developed in statistical mechanics to explain the structure of NP-hard problems (Mezard and Montanari, 2009; Marino, 2023), as well as to explain machine learning models (Deshpande and Montanari, 2014; Bahri et al., 2020; Decelle, 2023; Lauditi et al., 2023), the application of information theory to understand the population dynamics induced by natural selection (Adami, 2012; Baez, 2021; Kwessi, 2024), and tools from evolution to understand economic systems (Mirowski, 1983; Hodgson, 1996).

To the extent that such transfer between fields is possible and that understanding in scientific fields is obtained by discovering causal principles underlying the respective phenomena under study, it is clear that the success of transfer between different fields is due to shared features of these causal structures. The question is how the required synthesis can be performed on the level of the

underlying causal patterns and what methods will be required to capture such similarities quantitatively. Some research in this direction has been done in the field of causal meta-learning, in which knowledge of how to learn causal networks in one domain (or for one distribution) is applied to learning causal networks under different conditions (Ton et al., 2021; Chen et al., 2023). However, the field is still in its infancy, and little is known about the use and generalizability of such methods.

Of course, this is a general problem that can be leveraged against the field of causality as a whole: given the lack of commonly used causal benchmark datasets, do we actually know how well *any* method truly works?

7.3.3 (Not) Benchmarking and Causal Research Progress

Much recent progress in machine learning has been enabled by thorough benchmarks. The idea that benchmarks drive progress is not new: in many competitive sports, this is the *main* contributing factor to improved performance: since the 4-minute mark for 1-mile runs—previously thought impossible—was broken, world records have decreased linearly over time. Similarly, scores in Olympic disciplines are benchmarks par excellence, and Go players' ratings and move quality have been rising rapidly since the release of the open source Go program Leela (Choi et al., 2023). Closer to home, computer architecture benchmarks such as SPEC in general, and MLPerf and MLCommons more specifically for machine learning, are used to measure the performance of new architectures (Reddi et al., 2020).

Within machine learning itself, the biggest move towards benchmarks happened with the introduction of ImageNet (Deng et al., 2009) in computer vision, leading to dramatic increases in performance within a few years until the error rate of computer vision models was lower than that of humans themselves.

In the field of NLP, we also have a large number of benchmark datasets such as the Stanford Question Answering Dataset (SQuAD; Rajpurkar et al., 2016), the General Language Understanding Evaluation (GLUE; Wang et al., 2018), and SuperGLUE (Sarlin et al., 2020), and recently the use of tests designed for humans to benchmark Large Language Models (LLMs) such as GPT4 (OpenAI et al., 2023). More benchmark datasets exist for a large number of other tasks and can be found at Papers with Code (PWC),[1] and Hugging Face.[2]

It has been argued that while the majority of our efforts go into developing new models, more value, lies in the preparation of good data both for training purposes (Sambasivan et al., 2021) and to benchmark methods (Koch et al., 2021). Of course, benchmarks are not without their dangers, and it has been

[1] `https://paperswithcode.com/`

[2] `https://huggingface.co/`

shown both that historically, benchmark datasets such as ImageNet were biased (Paullada et al., 2021), but also that the provenance of such benchmarks lies in the hands largely of a few large institutions (Sambasivan et al., 2021). A more general concern is Goodhart's law: "when a measure becomes a target, it ceases to be a good measure" (Strathern, 1997). That is, do benchmark results accurately track progress in a given research field? Given the progress in both CV and NLP, the answer certainly seems to be yes.

In contrast, within causality, there are only relatively few commonly used datasets, such as the Tübingen cause-effect-pairs (Mooij et al., 2016) for bivariate causal discovery, and the Sachs dataset as one of the most commonly used datasets in causal discovery (Sachs et al., 2005). The only larger benchmark dataset for causal network inference that we are aware of is CausalBench (Chevalley et al., 2022). At the time of this writing,[3] the PWC website contains only 39 datasets containing the word "causal", of which only two are tagged as "causal discovery", one as "causal inference", and of which more than half (20) are marked as *text* datasets. Of the 64 datasets on Hugging Face containing the word "causal", every single one of them is either related to LLMs or has no documentation available.

Why is it that are there so few benchmark datasets in the field of causality? We can think of a number of reasons for this. First, causality started out as a highly theoretical field with both Pearl and Rubin (Rubin, 1974; Pearl and Verma, 1995), and a quick look at recent publications shows that not much has changed (of course this thesis is no exception). After all, it is all well and good if things work in practice, but do they work in theory? This is largely in contrast to fields such as computer vision, which, while having a significant theoretical component, are nevertheless primarily empirically driven, and this trend has only become stronger with the rise of deep learning methods.

Second, due to this focus on theory in the causal literature, the goal is rarely to improve at one specific task but rather to find different variations of the problem for which novel results can be proved and novel methods developed. That is, for any one set of assumptions for which the causal parameters are known to be identifiable, there are a handful of different methods to recover these parameters. Instead, more research is being conducted to discover different settings where parameters are not yet identifiable.

Third, on a related note, the risk profile in causal discovery and inference differs from other fields. If a computer vision model tells us that an image is a llama, rather than a cow, it is easy for us to inspect its output and realize that a mistake has been made. As such, while mistakes can be problematic when a faulty model is deployed, the cost of discovering such faults is relatively low. In contrast, in causal discovery, such inspection is *almost* impossible—if humans were good at determining causality, we would not need our models—and it

[3] April 8, 2024

would require experimentation to validate a model's output. Furthermore, since causal models underlying scientific theories guide health, economic, and other policies, mistaken causal estimates are likely to have more severe downstream effects than mistakes in other models.
Fourth, obtaining good data on which to test causal inference and discovery methods is costly and often challenging. Where computer vision data can be collected at a relatively low cost from publicly tagged data (Rios et al., 2021) or by paying Mechanical Turk workers, no such cheap sources of causal ground truth are available in the real world. Unless, of course, they are.

7.3.4 Of Word Models and World Models

Given the time at which this thesis was written, it would be remiss not to touch on the topic of Large Language Models (LLMs). LLMs such as GPT4 have been shown to have a surprisingly large number of capabilities that we would not intuitively expect them to have (Bubeck et al., 2023), and perform at, or close to, human level on many tasks such as summarization (Pu et al., 2023), deception (Park et al., 2024), moral judgments (Dillion et al., 2023), creativity (Koivisto and Grassini, 2023), scientific feedback (Liang et al., 2023b), and chess (Karvonen, 2024). Furthermore, sophisticated prompting techniques such as chain of thought or analogical reasoning (Wei et al., 2022; Meincke et al., 2024; Yasunaga et al., 2023), as well as document retrieval (Andriopoulos and Pouwelse, 2023; Zhang et al., 2024), and feedback from bigger models (Olausson et al., 2023), can significantly improve performance on many tasks.
Even so, these models show a number of biases due to the way they are trained, so that GPT4 is much better, for example, at ROT13 encodings than at ROT2 encodings (McCoy et al., 2023). Similarly, LLMs tend to overestimate rare events since everyday events often go unmentioned (Shwartz and Choi, 2020), and that changing variable names in code also decreases performance (Hooda et al., 2024). Two common failure modes in LLM evaluations, leading to inflated performance, are to neglect how much data has been memorized (Chang et al., 2023) and to include the benchmark questions in the training data, whether this is done deliberately or not (Li and Flanigan, 2024).
More importantly for our purposes here, LLMs have also been claimed to perform well on tasks of causal inference and discovery (Ortega et al., 2021; Jin et al., 2023; Jiang et al., 2023a; Kıcıman et al., 2023; Jiralerspong et al., 2024), and to develop internal representations of space and time (Gurnee and Tegmark, 2023), as well color (Patel and Pavlick, 2022), among others (Marks and Tegmark, 2023; Yildirim and Paul, 2024; Kosinski, 2023). Furthermore, given some measured variables, they are able to suggest latent confounders that might be affecting the results (Sharma and Kiciman, 2020; Blöbaum et al., 2022). While it would be easy to dismiss these results as trivial due to memo-

rization or come up with countless examples where they fail, many such objections could be just as well leveraged against humans. More interesting is why these claims *might* be true, and if so, what we *can* do.

To start with the first question, we need to start with understanding what LLMs are doing. At their heart, LLMs try to solve the unsupervised task of *modeling language* by turning it into the supervised task of *next token prediction*. Despite the simplicity of this approach, it has been shown that modeling the conditional distribution of the next state of a Markov chain suffices to correctly model the entire joint distribution over all states (Shalizi and Crutchfield, 2001). That is, next token prediction is equivalent to learning the structure of the entire language. This joint distribution and the structure it is endowed with is what we will refer to as *word model*. Of course, real models are both constrained by guardrails (or, more euphemistically, safety features), and they are far from modeling the distribution perfectly. While these concerns are important in practice, they are of little interest in answering what LLMs can do in theory.

What we really care about in learning about causality is to develop what has been called *world models* in the literature on reinforcement learning (Ha and Schmidhuber, 2018). That is, what are, or would be, the effects of taking a given action in our current state? This is similar to the mental models humans build as they interact with the world, but unlike LLMs, humans have access to drastically more data sources than mere language.

We can now rephrase our questions as follows: why would building a word model suffice to build a world model, and how would such a world model be represented? We are far from determining any precise answers to these questions, but we can make some attempts at conceptually framing the problem.

The first thing to note is that learning causal relationships from purely textual sources is not so uncommon, even in humans. When we say that gene A regulates gene B, for the vast majority of people, this is not due to their having first-hand experience with the process in such a way that we could draw inferences from data; instead, it is pieced together from our understanding of gene expression, gene regulation, and various other bits of information about biology, many of which are learned from textbooks or similar sources. Of course, this is not to say that language suffices to teach us everything about the world. In fact, many domains are too complex for such descriptions, and we often say that we cannot put our thoughts into words. This is a question of commensurability between the representational capacity of language and the problem we wish to represent. That is, some domains are so complex that language does not suffice to describe them, and we would, therefore, not expect LLMs to perform well at causality in such domains. For example, while the law is formulated mainly in large volumes of text, the domain is intrinsically so much more complex than Newtonian mechanics that we should not be surprised if LLMs perform better at questions of the results of interventions in classical

physics than the results of changes in the law.

Of course, humans do not learn *solely* from text. Instead, large parts of what we learn are due to observing and interacting with the world, obtaining both observational and interventional data in the process of doing so, and deriving causal relationships from observing the results of our actions. Notably, many of the representations by which we learn to understand the world are *not* textual in any way. Learning to catch a ball flying at us is not related to any explicit physics we have learned from textbooks, nor based on any other text we have read. So, despite humans learning (some) things from purely textual data, the world models that we build internally are grounded in perceptual data and active engagement with our surroundings. In contrast, LLMs have no such grounding, but multimodal models may change this.

The answer as to why this may work is that the representation space contained within language has been *pretrained* by humans in two ways. First, language is fundamentally about communication (Fedorenko et al., 2024), and communication is often about causality. In fact, humans can infer the causal structures underlying a sequence events from typical explanations (Kirfel et al., 2022; Beller and Gerstenberg, 2024). However, language is also deeply metaphorically structured out of perceptually grounded experiences (Lakoff and Johnson, 1980), so that large parts even of abstract mathematical concepts are constructed out of embodied metaphors (Lakoff and Núñez, 2000). For example, when we speak of somebody "climbing the ranks", this statement is a metaphor derived from the physical activity of a person climbing, and in fact, it implies agency and deliberate effort of said person. In contrast, the metaphor "rising through the ranks" implies no such deliberate effort. Furthermore, it has been argued that linguistic concept formation leads to concepts mapping onto *convex* regions of representation space (Gärdenfors, 2004, 2014). For example, despite color terms being different the world over (Berlin and Kay, 1969), there exists no language with a name for "red or green but not yellow", which would form a non-convex region in the human color representation space (Gärdenfors, 2014).

While language has acquired a great many purposes over time, two of its most fundamental purposes are the creation of social cohesion and the ability to explain the world. This is reflected by the fact that the oldest stories we know of are myths whose dual purpose was invariably to both explain "who we are and how we got here" and to foster social cohesion through such explanations (Frankfort and Groenewegen-Frankfort, 1946; Barber and Barber, 2006). Another tool to foster social cohesion is common knowledge of the characters of the members of the group. That is, it is essential for us to know if somebody is trustworthy and kind or untrustworthy and unkind. This led to what is now known as the *lexical hypothesis*, the idea first formulated by Galton (1884) that commonly used language contains all the information we need to describe personalities. This was confirmed by Thurstone (1934) and has since

been the basis for most major personality tests such as the HEXACO and Big Five tests (de Raad and Mlačić, 2015). More important to us, however, is the fact that *this personality structure is contained in LLM embeddings* (Cutler and Condon, 2022). That is, by looking at the first five principal components of the embeddings of adjectives used in personality descriptions, we recapture the same constructs as the Big Five personality test, with precisely the same correlations between factors, independently of the precise language model used.

By leveraging the dual purpose of language in human communication, we might expect that causal structures, too, are represented within language, which we may call the *causal lexical hypothesis*. This seems to be further supported by the fact that young children are leveraging their parents' causal language to refine their own causal models (Meltzoff, 2007; Bonawitz et al., 2010). The question, then, would be how the causal structure is represented. Do the embeddings of "X causes Y" and "Y causes X" relate to each other the same way, regardless of what X and Y are? Does "he killed her" relate to "he caused her death" in the same way that "he bankrupted the company" relates to "his decisions lead to the company's bankruptcy", and can we extrapolate this to get stronger or weaker causal statements? Do the embeddings of the correct and incorrect causal directions have anything in common with the linear geometry of truth proposed by Marks and Tegmark (2023)? How complex is the geometry of causality? How large does an LLM have to be for it to be represented? How can we finetune or steer LLMs to better represent it?

Of course, even if language does encode causality in some way, why should an LLM, whose only goal is to predict the next token, "know" about this? Precisely *because* it predicts the next token! An optimal next token predictor is an optimal language compressor in the Kolmogorov sense, so it should learn all computable structure contained in language. The question is how we can access this knowledge. If language is ergodic, then the algorithm of Ziv and Lempel (1978) is also an asymptotically optimal compressor, but we would not be able to easily access the structures it learned. Precisely such a method has been developed using the gzip algorithm, although it is unfortunately not competitive with any large models (Jiang et al., 2023b). In contrast, LLMs appear to represent concepts we are interested in mostly in a linear manner (Cutler and Condon, 2022; Gurnee and Tegmark, 2023; Marks and Tegmark, 2023).

One may object that none of the things contained within the text data are of an interventional nature. Much of what humans learn about the world is learned by interacting with it, thereby intervening on the causal mechanisms affecting the objects around them. However, this is hardly *necessary* to distinguish causal from non-causal explanations of events. Whether one decides to actively test the hypothesis that the crowing rooster causes the sun to rise (by moving elsewhere) or whether the rooster falls ill and does not crow one day is of no importance when the hypothesis is falsified by the sun rising even absent the

rooster's crows. That is, the important ingredient is not the ability to intervene but to observe data from different contexts, which may or may not arise from one's interventions. Given the vast amounts of data that modern LLMs are trained on, it is reasonable to assume that they have seen more data on almost any topic from more different contexts than almost any individual expert in their own domain. Of course, this is simply a restatement of the arguments mentioned in Chapter 6. That is, much progress has been made on the topic of causal learning by recognizing that interventions are simply one form of distribution shift, and other types of distribution shifts can work just as well to give us extra information about the underlying causal structure (Huang et al., 2020; Mooij et al., 2020; Perry et al., 2022). Moreover, there is no reason why the same should not apply in the language domain.

Now, assuming that LLMs can, in fact, model causality, the next question is, of course, *how well*? Unfortunately, the empirical analysis required for this is far beyond this section, but as mentioned above, some work has been done on this (Ortega et al., 2021; Jin et al., 2023; Jiang et al., 2023a; Kıcıman et al., 2023; Jiralerspong et al., 2024). Instead, we can ask what kind of training data would help us *do better*. Based on the idea that distribution shifts across contexts reveal information about causal structures, we suggest that there is only a small gap between learning simple models to make probabilistic predictions and learning the underlying causal structures *in a sufficiently rich setting*. Suppose we can learn to predict the conditional probabilities $P(Y \mid X)$ for widely varying variables X and Y. If we are correct, this permits us to learn something about the causal structure underlying these variables.

While it turns out LLMs are already surprisingly competent at the task of predicting events even with their current training (Schoenegger et al., 2024; Pham and Cunningham, 2024), it seems likely that to do better we would need much larger amounts of data to successfully train models on this task. Of course it may be that no amount of ability to perform probabilistic prediction will allow us to bootstrap true causal understanding in LLMs, but even if that is the case, it will nevertheless be interesting to see just where it fails, and how these shortcomings can be dealt with.

Overall, just like Alice, we live in an exciting time, and wheresoever we may go, we will find more problems in causality so long as we continue walking.

Appendices

A.1 Coding of the Tübingen Pairs

Here we give a full list of which pairs of the Tübingen pairs dataset we considered to be mainly causal, confounded, or which we were uncertain about.

- Causal: 13–16, 25–37, 43–46, 48, 54, 64, 69, 71–73, 76–80, 84, 86–87, 93, 96–98, 100
- Confounded: 65–67, 74–75, 99
- Uncertain: 1–12, 17–24, 38–42, 47, 49–53, 55–63, 68, 70, 81–83, 85, 88–92, 94–95

For example for pairs 5–11 it was unclear to us to what extent the age of an abalone should be considered as a causal factor to its length, height, weight, or other measurements, and to what extent all of these should simply be confounded by the underlying biological processes of development.

As another example, for pair 99 we believed that it is reasonable to suggest that the correlation between language test score of a child and socio-economic status of its family might more plausibly be explained by the unmeasured intelligence of parents and child — which are strongly correlated themselves due to high heritability of intelligence.

A.2 Proofs for Chapter 2

Lemma 2.1. *Let $X \sim P$ and f be some (measurable) function. Then X and $f(X)$ are statistically independent if and only if f is constant.*

Proof of 2.1. If X and $f(X)$ are independent, then for any two measurable sets A, B we have that

$$P(X \in A \cap f^{-1}(B)) = P(X \in A, f(X) \in B) = P(X \in A)P(f(X) \in B)\,.$$

In particular by setting $A = f^{-1}(B)$ we obtain

$$P(f(X) \in B) = P(f(X) \in B)^2\,,$$

which can only happen when $P(f(X) \in B) \in \{0, 1\}$. In particular, by setting $B = (-\infty, b]$, we see that

$$f(X) = b_0 = \arg\min_b \{b : P(f(X) \leq b) \neq 0\}\ ,$$

so that f is indeed constant on the range of X. □

Theorem 2.2 (Kolmogorov Does Not Incorrectly Detect Confounders)**.** *For any distribution $P(X)$, the following inequality holds*

$$\inf_{P(X,Z)\in P^s} K(P(X,Z)) \overset{+}{\leq} K(P(X))\,,$$

where the infimum is over the set P^s of all joint distributions $P(X, Z)$ with fixed marginal $P(X)$ and jointly independent Z. Conversely, if a joint distribution $P(X, Z) \in P^s$ exists such that the inequality

$$K(P(X,Z)) \overset{+}{<} K(P(X))\,,$$

holds, then the true generating mechanism of X includes latent variables influencing some subset $X_S \subseteq X$ of the observed variables.

Proof of Theorem 2.2. To prove the first statement, let Z be jointly independent and let there be no edges $X \to Z$. Pick P such that $P(Z = 0) = 1$. Then Z contains no information about X so that $K(P(X, Z)) \leq K(P(X))+K(P(Z)) = K(X) + O(1)$, with constant $K(P(Z)) = O(1)$ independent of $P(X)$.
For the second statement, consider the case where the true generating mechanism for X does not include any latent variables for any subset X_S. Then as noted in Equation (2.1) and the discussion preceding it, *all* information needed to compress $P(X)$ is already present in the graph G^*_X giving the optimal factorization of $P(X)$. Hence $K(P(X, Z)) \geq K(P(X)) + K(P(Z|X)) > K(P(X))$. □

Theorem 2.3 (Consistency of COCO)**.** *Let x^n, y^n be n be samples from the distribution M^* which is contained in $\mathcal{M}_{\mathrm{ca}} \cup \mathcal{M}_{\mathrm{co}}$. Then*

$$\lim_{n\to\infty} n^{-1} \left(L_{\mathrm{co}}(x^n, y^n) - L_{\mathrm{ca}}(x^n, y^n)\right) \begin{cases} \leq 0 & \text{if } M^* \in \mathcal{M}_{co} \\ \geq 0 & \text{if } M^* \in \mathcal{M}_{ca}\,, \end{cases}$$

with strict inequalities if M^* is contained in precisely *one of the two classes.*

Proof of 2.3. Let x^n, y^n be samples from $M^* \in \mathcal{M}_{\mathrm{ca}}$. Then by MDL consistency we know that in the limit $n \to \infty$, we have that $L_{\mathrm{ca}}(x^n, y^n) = L(x^n, y^n \mid M^*) + o(n)$ (Grünwald, 2007). But since this is the best *any* compression scheme can asymptotically perform in the limit $n \to \infty$, it follows that $\lim n^{-1} L_{\mathrm{co}}(x^n, y^n) - L_{\mathrm{ca}}(x^n, y^n) \geq 0$. Conversely, when $M^* \in \mathcal{M}_{\mathrm{co}}$, the same holds in the other direction. □

A.3 Proofs for Chapter 3

Proposition 3.1 (Confounders and Cliques)**.** *Let $P(X, Z)$ be the joint distribution of X, Z where Z is one-dimensional and let $\mathcal{I} = \{i : Z \to X_i\}$ be the set of indices of variables co-caused by Z. Then, any graph G_X capturing the correlations in $P(X)$ contains a clique over $X_{\mathcal{I}}$.*

Proof of Proposition 3.1. For any two $i, j \in S$ we know that, since they are direct descendants of Z, $X_i \not\perp\!\!\!\perp X_j \mid U$ for any $U \subset \{X_1, \ldots, X_m\} \setminus \{X_i, X_j\}$. Hence all edges $\{X_i, X_j\}$ are in G so that S is a clique in G. □

Theorem 3.2 (Identifiability of the SLC Model)**.** *Let Z be of dimension $l \leq m/4$, and let $P(X, Z)$ be described by the linear SCM of Equation* (3.1)

$$X = AX + BZ + \varepsilon\,.$$

Further, let Assumptions A–C hold. Then, both the number l of confounders and its parameters B are identifiable up to trivial indeterminacies (column permutations and rescaling). Furthermore, if either all noise variables ε_i are non-Gaussian or all ε_i have equal variances, then A is also identifiable.

Proof of Theorem 3.2. We prove this statement in two steps. First, we show that all b_{ij} are identifiable. Let $i \in \{1, \ldots, m\}$ and $j \in \{1, \ldots, l\}$. Then, by assumption (A2) there exists a distinct quadruple (X_i, X_u, X_v, X_w) of nodes that are conditionally independent given Z_j. In order to make every quadruple (X_i, X_u, X_v, X_w) be dependent conditional on Z_j, it would have to have either an edge between them or a common predecessor, which would require at least $|S_j| - 3$ incoming edges to S_j from sources that are not Z_j.

Therefore, for any two variables (X_λ, X_μ) in our quadruple we know that $\sigma_{\lambda\mu} = \mathrm{cov}(X_\lambda, X_\mu) = b_\lambda b_\mu$ and in particular

$$\sigma_{iu}\sigma_{vw} = b_i b_u b_v b_w = \sigma_{iv}\sigma_{uw} \,.$$

We can therefore write

$$b_i^2 = \sigma_{iu}\sigma_{iv}/\sigma_{uv} \,.$$

Furthermore, no quadruple $(X_i, X_{u'}, X_{v'}, X_{w'})$ this is not conditionally independent given Z can satisfy the constraint $\sigma_{iu}\sigma_{vw} = \sigma_{iv}\sigma_{uw}$ by assumption (A3). Hence, all b_{ij}^2 are identifiable, and since we assume that Z_j is symmetric around 0, so is the effect $b_{ij}Z_j$ of Z_j on X_i. Now, knowing these effects $b_{ij}Z_j$ on each X_i, we can determine the distribution $P(X \mid Z)$, which depends only on A and σ_ε^2. Now, if $X \mid Z$ is either a linear SCM with non-Gaussian noise variables, identifiability follows from LiNGAM (Shimizu et al., 2006). Meanwhile, if the noise variables are all Gaussian with equal noise variances, the result follows from Peters and Bühlmann (2014). □

Theorem 3.3 (Identifiability for Large Dense Graphs)**.** *Let Assumptions A and C hold and let the true causal graph G^* over X, Z be sampled from a directed Erdős-Rényi model $ER(m + l, p)$, with m observed and l latent nodes and edge probability $p < 1$. Then in the limit of infinitely many variables, the matrices A and B are identifiable with probability* 1*,*

$$\lim_{m\to\infty} P(A, B \textit{ identifiable}) = 1 \,,$$

where the limit is taken over DAGs with fixed topological order.

Proof of Theorem 3.3. To prove this result, we need to show that for any $p < 1$, for every variable X_i, we can find triples (X_u, X_v, X_w) such that (X_i, X_u, X_v, X_w) are conditionally independent given Z_j as in the previous proof by which to identify the value of b_{ij}.

We do this through a simple counting argument. Let $Y = X_i$ be fixed, and $U = X_u \in X$ be another variable. What is the probability that U is not admissible in a quadruple as above? There are two possibilities:

a) Either, U is ruled out by there existing an edge $X - U$ or edges $X \leftarrow R \rightarrow U$.
b) Or, U is ruled out by being mutually connected to too many other nodes V, which are not ruled out by step 1.

We begin with the first case. The probability of this occurring is $r \coloneqq p + p^2 - p^3 = p + (1 - p)p^2$ where the last term is subtracted because otherwise, we

would be counting the intersection twice. Note that for $p < 1$, we have $r < 1$. Next note that for any $r < 1$ and any $s < 1$ we have

$$r + (1 - r)s < 1 \,.$$

Hence we need to show that the probability s of the second case above is < 1. This is, however, trivial since U can only be ruled out in this way if there is at least one edge to another variable V, i.e., it cannot have zero edges, such that the probability $s < 1$ for any $p < 1$.
Therefore, the probability of a node U being ruled out is strictly less than 1 so that in the limit $m \to \infty$, we are guaranteed to find at least one valid quadruple. □

Theorem 3.4 (Consistency of BIC for Gaussian SLCs)**.** *Let $x = x^n$ be a sample from the SLC of Equation* (3.1) *with Gaussian distributions $P(Z), P(\varepsilon)$ and let Assumptions A–C hold. Let $\mathcal{M}$ be the corresponding model class and $\mathcal{M}_0$ the subset of $\mathcal{M}$ with $B = 0$ fixed. Further, consider the score*

$$L(x^n, M) = -\log P(x^n \mid A, B, \sigma_\varepsilon^2) + \lambda \left\|A\right\|_0 + \lambda \left\|B\right\|_0 \,, \tag{3.2}$$

and denote its minimizers by $\widehat{A}, \widehat{B}$. Then, for $\lambda = \log(n)/2$, our score L is the BIC score and is consistent for detecting confounders. That is,

$$\lim_{n\to\infty} P\left(\min_{M\in\mathcal{M}} L(x^n, M) < \min_{M\in\mathcal{M}_0} L(x^n, M)\right) = 1 \,.$$

Furthermore, $\widehat{A}$ and $\widehat{B}$ converge to the true A, B with probability 1*,*

$$\lim_{n\to\infty} P(\widehat{A} = A, \widehat{B} = B) = 1 \,.$$

Proof of Theorem 3.4. As we've seen in the proof of Thm. 3.2, for each X_i there exists a distinct quadruple of variables (X_i, X_u, X_v, X_w) that are conditionally independent given Z_j by Assumption B. Hence, all correlations between these four variables can be explained by the parameters in B. Furthermore, by Proposition 3.1, no pair of variables X_μ, X_λ can be d-separated in any DAG over X, so that by setting $b_{ij} = 0$ we would require *at least* four additional entries of A to be non-zero, instead of only one in B.
Hence, since in the limit we have $\widehat{b}_{ij}\widehat{b}_{vj} - \sigma_{iv} \to 0$, the matrix $\widehat{B}$ converges towards B (Haughton, 1988). Furthermore, given a good approximation of $P(X)$ and of B, due to the joint continuity of $L(x^n, M)$ in the matrices A, B, we obtain a good approximation of A (Chickering, 2002b; van de Geer and Bühlmann, 2013). □

Theorem 3.5 (MDL Consistency for SLCs)**.** *Let the assumptions of Theorem 3.4 hold. Then the minimizer* $\widehat{G}$*,*

$$\widehat{G} = \arg\min_{G} L(x^n; \mathcal{M}(G)) \,,$$

converges to the ground truth graph G^* *with probability one,*

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

Proof of Theorem 3.5. From standard MDL theory we know that the Bayesian MDL score $L(x^n, \mathcal{M}(G))$ is asymptotically equivalent to the BIC score $L(x^n, M)$ in the sense (Grünwald, 2007)

$$|L(x^n, \mathcal{M}(G)) - L(x^n, M^*(G))| = o(1) \,,$$

where $M^*(G)$ is the best model in $\mathcal{M}$. But since for any $G \neq G^*$ we have

$$L(x^n, M^*(G)) - L(x^n, M^*(G^*)) \propto \log(n) \,,$$

the same is true for $L(x^n; \mathcal{M}(G))$ and so we are guaranteed to pick the same minimizing $\widehat{G} = G^*$ in the limit. □

Proposition 3.6 (Consistency of COCO for Discovering Confounded Nodes)**.** *Let* x^n *be the an i.i.d. sample from* $P \in \mathcal{M}(G^*)$ *defined in Equation* (3.3)*, let Assumptions A–C hold and let* $\mathcal{I}_i^*$ *be the set of nodes affected by* Z_i*. Assume that* $\bigcap_{s\in\mathcal{I}_i^*} \mathrm{MB}_{G^*}(X_s) \setminus \{Z_i\} \subsetneq \mathcal{I}_i^*$*. Let* $\mathcal{A}$ *be consistent for recovering the Markov equivalence class of the graph* G_X *for distribution* $P(X)$*. Let* $\widehat{\mathcal{I}}_i$ *be the set of nodes confounded by* Z_i *discovered by* COCO*. Then*

$$\lim_{n\to\infty} P(\widehat{\mathcal{I}}_i = \mathcal{I}_i^*) = 1 \,.$$

Proof of Proposition 3.6. By Proposition 3.1 we know that in the limit $\mathcal{I}_j^*$ forms a clique in the marginal graph $\widehat{G}_X$ inferred over X by a consistent $\mathcal{A}$. This clique is maximal due to our assumption no node being in the Markov Blanket of all $s \in \mathcal{I}_j^*$. Further, since x^n is a sample from Equation (3.3) we know from the MDL principle for selecting nested model classes (Grünwald, 2007) that in the limit no other set can be compressed better by introducing a confounder than the set $\mathcal{I}_j^*$ itself. □

A.4 Proofs for Chapter 4

Theorem 4.1 (Identifiability in the Sparse PNL Model)**.** *Let the distribution $P(X, Z)$ be described by the nonlinear SCM of Equation* (4.4)*, i.e.,*

$$X = \tau\left((I - A)^{-1}(BZ + \varepsilon)\right),$$

for some Z of dimension $l \leq m/4$. Further, let Assumptions A–C hold. Then both the number l of confounders Z and the causal effects of the confounder, B, are identifiable up to trivial indeterminacies (column permutations and rescaling). Furthermore, if Assumption D also holds, then A is also identifiable.

Proof of Theorem 4.1. Note first of all that for any variables X_i, X_j we have that the mutual information $I(X_i; X_j) = I(\tau_i^{-1}(X_i); \tau_j^{-1}(X_j))$. Hence, since $X = \tau((I - A)^{-1}(BZ + \varepsilon))$, it suffices to study the fully linear Gaussian case. Note that for two Gaussian variables, we have

$$I(X_i; X_j) = -\frac{1}{2}\log(1 - \rho_{ij}^2),$$

where ρ_{ij} is the pair's correlation coefficient.
It, therefore, suffices to show that the parameters ρ_{ij} fully determine the matrices C and B when Assumptions A–D hold. To this end, note that the variances of each variable are observable and therefore ρ_{ij} and $\sigma_{ij} = \mathrm{cov}(X_i, X_j)$ contain equal amounts of information.
Hence, b_{ij}^2 and the effects $b_{ij}Z_j$ on X_i are identifiable as in Theorem 3.2. Given these effects, the matrix C (and therefore A) is identifiable by the general identifiability result of Corollary 31(ii) (Peters et al., 2014) when all τ_i are strictly nonlinear. When $\tau_i = \mathrm{id}$ and all noise variables share the same variance, then identifiability is again guaranteed by Peters and Bühlmann (2014). □

Corollary 4.2. *Let the Assumptions of Theorem 4.1 hold and let all τ_i be strictly increasing and standardized to satisfy $\tau_i(1) = 1$. Further, let the variances $\sigma_{Z_j}^2$ of each Z_j be known, and for each j let the sign of b_{ij} be known for at least one $X_i \in S_j$. Then B is identifiable up to permutations of its columns.*

Proof of Corollary 4.2. In the above, we could identify B only up to sign and scale due our results only distinguishing between b_{ij}^2, and the influence of $\sigma_{Z_j}^2$ and τ_i. However, by assuming precisely that $\tau_i(1) = 1$, we fix the scale of $\tau^{-1}(X)$, and by assuming that $\sigma_{Z_j^2}$ are known, we fix the scale of Z. Furthermore, by assuming that for each Z_j we know the sign of at least one $b_{ij} \neq 0$, we obtain the signs of all b_{ij}, so that B becomes identifiable up to column permutations. □

Theorem 4.3 (Consistency under Sparsity)**.** *Let x^n be a sample generated from the model in Equation* (4.4) *with $\tau = \mathrm{id}$ and let Assumptions A–D hold. Let L be the L^0-penalized ELBO score given by*

$$L(x^n; A, B) \coloneqq -L_{\mathrm{ELBO}} + \lambda_A \|A\|_0 + \lambda_B \|B\|_0 \ ,$$

and let $\widehat{A}, \widehat{B}$ be its minimizers subject to acyclicity, i.e.,

$$\widehat{A}, \widehat{B} = \operatorname*{arg\,min}_{A,B} L(x^n; A, B)$$
$$s.t. \ \ h(A) = 0 \, .$$

Then for sufficiently small $\sigma_\varepsilon(X), \sigma_z(X)$ the score L is consistent for recovering the matrices A, B when $\lambda_A = \lambda_B = \log(n)/2$:

$$\lim_{n\to\infty} P(\widehat{A} = A, \ \widehat{B} = B) = 1 \, .$$

Proof of Theorem 4.3. Let us write $\widehat{x}$ for our reconstruction of x. A typical sample loss of our L_{ELBO} score is then given by (Yu et al., 2019)

$$-L_{\mathrm{ELBO}} = \frac{1}{2} \sum (x_{ij} - \widehat{x}_{ij})^2 + \frac{1}{2} \sum H_{ij}^2 \ .$$

Furthermore in the case $\tau = \mathrm{id}$ the reconstruction is perfect, $\widehat{x} = x$ and $H = (I - A)x$ so that the L_{ELBO} can be written as

$$-L_{\mathrm{ELBO}} = \frac{1}{2} \|x - Ax - Bz\|_F^2 \ .$$

As before, the goal is therefore to model the empirical correlations between variables. Using Assumptions A–D, as in the proof of Theorem 4.1, for each i, there exists a distinct quadruple which is independent given Z_j, which permits us to explain the correlations between them by way of the parameters b_{ij} and since the variables cannot be rendered independent without reference to Z_j, any explanation of the same correlations involving only variables X_u would need six parameters a_{iu} instead of the four parameters b_{ij}, so that due to the use of $\|\cdot\|_0$ penalties the explanation using Z_j is preferred. We therefore have $\widehat{b}_{ij}\widehat{b}_{uj} - \sigma_{iu} \to 0$ so that $\widehat{B}$ converges to B.

Furthermore, given the matrix B describing how Z affects X, we can learn the matrix A from the consistency result of van de Geer and Bühlmann (2013). □

A.5 Proofs for Chapter 5

Theorem 5.1 (Identifiability under Noiseless Selection for Exponential Families)**.** *Let $\mathcal{M}$ be an exponential family with parameter space Θ and sufficient statistics $T(x)$ non-constant on every half-space. Further, let $\mathcal{A}$ be the set of (normalized) selection vectors a such that*

$$0 < P_\theta(a^\top X > 0) < 1\,, \tag{5.2}$$

for all θ. Then the parameters (θ, a) of $Q_{\theta,a}$ are identifiable. In particular, P_θ is fully determined by the distribution $Q_{\theta,a}$.

Proof of Theorem 5.1. Let $(a,\theta), (a',\theta')$ be two pairs of parameters as in the assumptions and denote $Q \coloneqq Q_{\theta,a}, Q' \coloneqq Q_{\theta',a'}$. Note that in particular we have $\operatorname{supp}(Q) = \operatorname{supp}(Q')$ so that

$$\operatorname{supp}(P_\theta) \cap \left\{a^\top X > 0\right\} = \operatorname{supp}(P_{\theta'}) \cap \left\{a'^\top X > 0\right\}\,,$$

so that from $P_\theta(a^\top X > 0) \in (0,1)$ it follows that $a = a'$.
Now given $a = a'$ it follows that for $u \in \left\{a^\top X > 0\right\}$ we have

$$\frac{P_\theta(u)}{P_\theta(a^\top X > 0)} = Q(u) = Q'(u) = \frac{P_{\theta'}(u)}{P_{\theta'}(a^\top X > 0)}\,,$$

so that

$$\log P_\theta(u) = \log P_{\theta'}(u) + \log\left(\frac{P_\theta(a^\top X > 0)}{P_{\theta'}(a^\top X > 0)}\right)\,.$$

Now if the second term on the right is zero, the claim $\theta = \theta'$ follows because P_θ form an exponential family. If it is not, then w.l.o.g. we can assume it is > 0 so that P_θ assigns strictly larger probabilities to every value of u than $P_{\theta'}$. But since $P_{\theta'}$ is a probability distribution that would imply that $\int dP_\theta(u) > \int dP_{\theta'}(u) = 1$, which is in contradiction to P_θ being a probability distribution.

□

Theorem 5.2 (Identifiability of Noisy Selection Effects in the Gaussian Family)**.** *Let $\mathcal{M}$ be the Gaussian exponential family with parameter space $\Theta = \{(\mu,\Sigma)\}$ and let $\varepsilon \sim N(0,1)$. Further, let the biased distribution be*

$$Q_{\mu,\Sigma,a,\zeta}(X) = P_{\mu,\Sigma}(X \mid a^\top X + \zeta\varepsilon > 0)\,.$$

Then the parameters $(\mu,\Sigma) \in \Theta, a \in \mathcal{A}, \zeta > 0$ are jointly identifiable.

Proof of Theorem 5.2. Let $\theta = (\mu, \Sigma, a, \zeta), \theta' = (\mu', \Sigma', a', \zeta')$ be two different vectors such that $Q = Q_\theta = Q_{\theta'} = Q'$. Let q, q' be the corresponding densities for Q respectively Q'. By plugging in the definition of the Gaussian density,

$$1 = \frac{q(u)}{q'(u)} = \frac{e^{-(u-\mu)^\top \Sigma^{-1}(u-\mu)} \Phi(a^\top u/\zeta)}{e^{-(u-\mu')^\top \Sigma'^{-1}(u-\mu')} \Phi(a'^\top u/\zeta')} \,.$$

By taking logarithms we obtain for all u

$$0 = \left(- \|u-\mu\|_\Sigma^2 + \|u-\mu'\|_{\Sigma'}^2 + \log\phi - \log\phi' \right) ,$$

where $\|\cdot\|_\Sigma$ is the Mahalanobis norm for Σ and $\phi = \Phi(a^\top u/\zeta)$. By taking derivatives we obtain

$$\begin{aligned} 0 &= A^\top (u-\mu) - A'^\top (u-\mu') \\ &\quad + \frac{\Phi'(a^\top u/\zeta)}{\phi} \frac{a}{\zeta} - \frac{\Phi'(a'^\top u/\zeta')}{\phi'} \frac{a'}{\zeta} \,. \end{aligned}$$

Note that if $A \neq A'$ we have $\Phi'(a^\top u/\zeta) \to 0$ but $\left\|(A-A')^\top u\right\| \to \infty$. Hence $A = A'$. By setting $u = 0$ we obtain that $\frac{a}{\zeta} - \frac{a'}{\zeta} = A^\top(\mu - \mu')$. However, by setting $u = \mu'$ we also obtain $\frac{\Phi'(a^\top \mu'/\zeta)}{\phi} \frac{a}{\zeta} - \frac{\Phi'(a'^\top \mu'/\zeta)}{\phi'} \frac{a'}{\zeta'} = A^\top(\mu' - \mu) = -(\frac{a}{\zeta} - \frac{a'}{\zeta'})$ which can only hold if $a = a'$ and $\zeta = \zeta'$. But then also $\mu = \mu'$ and we have proved what we wanted to prove. □

Theorem 5.3 (Identifiability of Selection under Invariance)**.** *Let $\mathcal{M}$ be a set of probability distributions and J be strongly distinguishable for each $P \in \mathcal{M}$. Assume that for all $P \in \mathcal{M}$ there is $j \neq \mathrm{id} \in J$ such that $P(X) = P(j(X))$. Let $P \in \mathcal{M}$ and $\mathcal{A}$ be the set of a for which there exists $j \in J$ such that*

$$\left\{a^\top X > 0\right\} \cap j^{-1}\left(\left\{a^\top X < 0\right\}\right) \neq \emptyset \,.$$

Then a is identifiable. Further, if all distributions $P_1, P_2 \in \mathcal{M}$ satisfy $P_1(\cdot \mid a^\top X > 0) = P_2(\cdot \mid a^\top X > 0)$ iff $P_1 = P_2$ then P is identifiable too.

Proof of Theorem 5.3. Let $j \in J$ be an invariance such that $E = \left\{a^\top X > 0\right\} \cap j^{-1}(\left\{a^\top X < 0\right\}) \neq \emptyset$. Then by definition $j(E) \cap \left\{a^\top X < 0\right\} \neq \emptyset$ and the points $y \in j(E) \cap \left\{a^\top X < 0\right\}$ are uniquely separated from points in $\left\{a^\top X > 0\right\}$ by the boundary a. □

Proposition 5.4. *Let x be a sample from the distribution $Q_{\theta,a}(X)$ with known underlying exponential family $\mathcal{M}$. Then, the saddle points $(\widehat{\theta}, \widehat{a})$ of Equation* (5.3) *satisfy the following conditions:*

a) $\widehat{a}$ *intersects the convex hull of* x.
b) Furthermore, if $\eta(\theta) = \theta$, *then* (θ^*, a^*) *is the unique global maximum of the large sample limit of the data log-likelihood* $\lim_{n\to\infty} \frac{1}{n} l_{\theta,a}(x)$.

Proof of Proposition 5.4. a) If $d = \operatorname{dist}(\widehat{a}, C) > 0$, where C is the convex hull of x, then by shifting $\widehat{a}$ towards $\widehat{a}'$ which is closer to C by any distance $d - \varepsilon$ for $\varepsilon > 0$, we would obtain a set $\left\{\widehat{a}'^{\top} X < 0\right\} \supsetneq \left\{\widehat{a}^{\top} X < 0\right\}$, which has strictly larger mass since P_θ is a continuous probability distribution.

b) In the limit $n \to \infty$, the only boundary a that does not discard sampled points $x \in \left\{a^{\top} X > 0\right\}$, but is such that $P\left(\left\{a^{\top} X < 0\right\}\right) > 0$ is $\widehat{a} = a^*$. But then, once a^* is known, the θ which maximizes $\log P_\theta(x^n \mid a^* x^n > 0)$ is θ^*.

□

A.6 Proofs for Chapter 6

Lemma 6.1 (Significance and Power)**.** *Let* X, Y *be unconfounded and* $X \to Y$. *Let* Π_X, Π_Y *be the corresponding partitions. Then*

$$\lim_{n_s \to \infty} P(t > q_{1-\alpha}) \to \alpha\,,$$

where $q_{1-\alpha}$ *is the* $1-\alpha$*-quantile of the standard normal distribution. Conversely, if* X, Y *are confounded, then for* $\alpha > 0$ *in the limit we obtain power*

$$\beta = \lim_{n_s \to \infty} P(t > q_{1-\alpha}) \to 1\,.$$

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_1, X_2, their partitions satisfy

$$EI(\Pi_1, \Pi_2) \geq \frac{n_s}{2} H(p) \gg EI(\Pi_1', \Pi_2')\,,$$

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 C. Note that the relation $\frac{n_s}{2} H(p) \gg EI(\Pi_1', \Pi_2')$ follows from the fact that $\lim_{n_s\to\infty} \frac{1}{n_s} I(\Pi_1', \Pi_2') = 0$ $\mathcal{P}$-almost surely so that $EI(\Pi_1', \Pi_2')$ cannot be extensive in n_s. Since $I(\Pi_1, \Pi_2)$ also concentrates around its mean, the result follows. □

Theorem 6.2. *Let $X_{\mathcal{I}}$ be a set of variables such that all $X_i, X_j \in X_{\mathcal{I}}$ are pairwise confounded. Then $X_{\mathcal{I}}$ is jointly confounded if and only if for each triple $X_i, X_j, X_k \in X_{\mathcal{I}}$ we have*

$$\lim_{n_s \to \infty} P(T(\Pi_i, \Pi_j, \Pi_k) < I(\Pi_i, \Pi_j) + I(\Pi_j., \Pi_k)) = \begin{cases} 1, & X_i, X_j, X_k \text{ jointly confounded} \\ 0, & \text{otherwise}. \end{cases}$$

Proof. As we have seen, 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 correlations between the partitions are shared, which can only happen due to joint confounding of more than two variables at a time. Now, let us assume that some set $\mathcal{I}$ of $s \geq 4$ pair-wise confounded nodes satisfying this inequality, does not share the same confounder between all nodes. Without loss of generality let us call these variables $X_1, \ldots, X_s$. Then the way for *every* triplet to have a shared confounder, and which requires the least number of edges into the set, 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 F. □

Proposition 6.3 (Consistency for Pairs of Variables)**.** *If a variable pair X, Y is confounded by a variable Z, then there exists some constant $\rho > 0$ such that*

$$P(I(\Pi_X^*, \Pi_Y^*) < I(\Pi_X, \Pi_Y)) = 1 - O(e^{-\rho n_s}).$$

Proof. Following Perry et al. (2022) we show more precisely that

$$\mathcal{P}(I(\Pi_X^*, \Pi_Y^*) < I(\Pi_X, \Pi_Y)) = 1 - O\left((p + (1-p)(1-p+p^2))^{\lfloor n_s/2 \rfloor}\right),$$

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 environments c_{2i}, c_{2i+1} are independent of each other, the probability of this happening in *all* environments is therefore $(p + (1-p)(1-p+p^2))^{\lfloor n_s/2 \rfloor}$ and since $p + (1-p)(1-p+p^2) \leq 1$ as convex combination of 1 and $(1-p+p^2)$, the result follows. □

Proposition 6.4 (Consistency for Recovering Parents)**.** *Let X_i be a target variable and let G and G' be two graphs in the MEC of the marginal distribution $P^s(X)$. Assume that only one of the two graphs correctly recovers the parents of X_i, $\mathrm{Pa}_i = \mathrm{Pa}_i^*$ and $\mathrm{Pa}_i' \neq \mathrm{Pa}_i^*$, and further assume that the number of latent confounders affecting X_i plus spurious siblings is bounded by $\frac{\log(0.5)}{\log(1-p)}$. Then*

$$\begin{aligned} P(I(\Pi_i, \{\Pi_j : j \in \mathrm{Pa}_i\}) < I(\Pi_i', \{\Pi_j' : j \in \mathrm{Pa}_i'\})) \\ = 1 - O(e^{-\rho n_s}) \,. \end{aligned}$$

Proof. More precisely, we will show that

$$\begin{aligned} &\mathcal{P}(I(\Pi_i, \{\Pi_j : j \in \mathrm{Pa}_i\}) < I(\Pi_i', \{\Pi_j' : j \in \mathrm{Pa}_i'\})) \\ &= 1 - O\left(\left((1-(1-p)^r) + (1-p)^r(1-p+p^2)\right)^{n_s/2}\right) \,, \end{aligned}$$

where r is the number of latent parents of X_i plus the number of other variables with which it is pair-wise confounded. In essence, these variables are precisely those which could make us not detect changes between two environments, just as in the previous proof changes in the mechanism of Z between environments could prevent us from detecting changes in the mechanisms of X or Y.

To this end, note that if $\mathrm{Pa}_i' \neq \mathrm{Pa}_i^*$ then there exists either a variable in Pa_i^* that is missing in Pa_i' or a child of X_i in Pa_i'. In either case, additional joint shifts are introduced between X_i and these variables and therefore the mutual information increased. 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_i is less than 0.5. □

Theorem 6.5 (Consistency)**.** *Let G^* be the true graph over V and let G_x^* be the induced graph on X, and assume that for all X_i the number of latent parents plus spurious siblings is at most $\frac{\log(0.5)}{\log(1-p)}$. Then with high probability, G_x^* and its partitions $\Pi_1^*, \ldots, \Pi_k^*$ are the unique minimum of total correlation,*

$$P\left(\underset{G, \Pi_1, \ldots, \Pi_m}{\arg\min}\ T(\Pi_1, \ldots, \Pi_m) = (G_x^*, \Pi_1^*, \ldots, \Pi_m^*)\right) = 1 - O(e^{-\rho n_s}) \,.$$

Proof. Let m be the number of observed variables and r be an upper bound on all the $r = \max\{r_i\}$ from the above Proposition. Then we specifically show

that

$$\mathcal{P}\left(\arg\min_{G} T(\Pi_1,\ldots,\Pi_m) = \{G_x^*\}\right)$$
$$= 1 - O\left(\frac{m^2(m-1)}{2}\left((1-(1-p)^r) + (1-p)^r(1-p+p^2)\right)^{n_s/2}\right).$$

To this end let us assume that the true causal ordering over X is given by $X_1 \leq \cdots \leq X_m$. Then note that by construction we have $T(\Pi_1,\ldots,\Pi_m) = \sum_i I(\Pi_i, \{\Pi_j : j \in \mathrm{Pa}_i\})$ so that the inside of our statement here is simply the sum of all terms in Proposition 6.4. As such, the total correlation is the unique minimum if the above proposition holds for all i and when compared against *any* other graph G, resulting in the union bound above. □

References

M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. TensorFlow: A System for Large-Scale Machine Learning. In *OSDI16*, pages 265–283, 2016.

C. Adami. The use of information theory in evolutionary biology. *Ann. N.Y. Acad. Sci.*, 1256(1):49–65, 2012.

J. Adams, N. Hansen, and K. Zhang. Identification of Partially Observed Linear Causal Models: Graphical Conditions for the Non-Gaussian and Heterogeneous Cases. *NeurIPS*, 34:22822–22833, 2021.

S. Affeldt, L. Verny, and H. Isambert. 3off2: A network reconstruction algorithm based on 2-point and 3-point information statistics. In *BMC Bioinformatics*, volume 17, pages 149–165. BioMed Central, 2016.

R. Agrawal, C. Squires, N. Prasad, and C. Uhler. The DeCAMFounder: nonlinear causal discovery in the presence of hidden variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(5):1639–1658, 2023.

D. Ahn, J. Yang, M. Cha, H. Yang, J. Kim, S. Park, S. Han, E. Lee, S. Lee, and S. Park. A human-machine collaborative approach measures economic development using satellite imagery. *Nat. Commun.*, 14(6811):1–10, 2023.

K. Ahuja, J. S. Hartford, and Y. Bengio. Weakly Supervised Representation Learning with Sparse Perturbations. *NeurIPS*, 35:15516–15528, 2022.

K. Ahuja, D. Mahajan, Y. Wang, and Y. Bengio. Interventional Causal Representation Learning. In *ICML*, pages 372–407. PMLR, 2023.

E. R. Álvarez Buylla, J. Dávila-Velderrain, and J. C. Martínez-García. Systems Biology Approaches to Development beyond Bioinformatics: Nonlinear Mechanistic Models Using Plant Systems. *BioScience*, 66(5):371–383, 2016.

K. Andriopoulos and J. Pouwelse. Augmenting LLMs with Knowledge: A survey on hallucination prevention. *arXiv:2309.16459*, 2023.

A. Anglemyer, H. T. Horvath, and L. Bero. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. *Cochrane Database of Systematic Reviews*, 2014(4), 2014.

J. D. Angrist. Conditional independence in sample selection models. *Economics Letters*, 54(2):103–112, 1997.

J. D. Angrist and J.-S. Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2009.

J. D. Angrist, G. W. Imbens, and D. B. Rubin. Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996.

O. A. Arah. Bias Analysis for Uncontrolled Confounding in the Health Sciences. *Annual Review of Public Health*, 38:23–38, 2017.

M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant Risk Minimization. *arXiv:1907.02893*, 2019.

M. Ashman, C. Ma, A. Hilmkil, J. Jennings, and C. Zhang. Causal Reasoning in the Presence of Latent Confounders via Neural ADMG Learning. In *ICLR*, 2023.

J. Austrian, F. Mendoza, A. Szerencsy, L. Fenelon, L. I. Horwitz, S. Jones, M. Kuznetsova, and D. M. Mann. Applying A/B Testing to Clinical Decision Support: Rapid Randomized Controlled Trials. *J. Med. Internet Res.*, 23 (4), 2021.

A. J. Averitt, C. Weng, P. Ryan, and A. Perotte. Translating evidence into practice: eligibility criteria fail to eliminate clinically significant differences between real-world and study populations. *NPJ Digital Medicine*, 3(1):67, 2020.

J. C. Baez. The Fundamental Theorem of Natural Selection. *Entropy*, 23(11): 1436, 2021.

Y. Bahri, J. Kadmon, J. Pennington, S. S. Schoenholz, J. Sohl-Dickstein, and S. Ganguli. Statistical Mechanics of Deep Learning. *Annu. Rev. Condens. Matter Phys.*, (Volume 11, 2020):501–528, 2020.

E. J. Barber and P. T. Barber. *When They Severed Earth from Sky: How the Human Mind Shapes Myth.* Princeton University Press, 2006.

E. Bareinboim and J. Pearl. Controlling Selection Bias in Causal Inference. In *AISTATS*, pages 100–108. PMLR, 2012.

E. Bareinboim and J. Pearl. Causal inference and the data-fusion problem. *PNAS*, 113(27):7345–7352, 2016.

E. Bareinboim, J. Tian, and J. Pearl. Recovering from Selection Bias in Causal and Statistical Inference. In *AAAI*, volume 28, 2014.

R. Barrangou and J. A. Doudna. Applications of CRISPR technologies in research and beyond. *Nat. Biotechnol.*, 34:933–941, 2016.

D. Barth, N. W. Papageorge, K. Thom, and M. Velásquez-Giraldo. Genetic Endowments, Income Dynamics, and Wealth Accumulation Over the Lifecycle. Technical report, National Bureau of Economic Research, 2022.

A. G. Baydin, B. A. Pearlmutter, A. A. Radul, and J. M. Siskind. Automatic Differentiation in Machine Learning: a Survey. *JMLR*, 18:1–43, 2018.

A. Beller and T. Gerstenberg. Causation, meaning, and communication. 2024.

K. Bello, B. Aragam, and P. Ravikumar. DAGMA: Learning DAGs via M-matrices and a Log-Determinant Acyclicity Characterization. *NeurIPS*, 35: 8226–8239, 2022.

A. Bellot and M. van der Schaar. Linear Deconfounded Score Method: Scoring DAGs With Dense Unobserved Confounding. *IEEE Trans. Neural Networks Learn. Syst.*, 35(4):4948–4962, 2024.

A. Benavoli, G. Corani, F. Mangili, M. Zaffalon, and F. Ruggeri. A Bayesian Wilcoxon signed-rank test based on the Dirichlet process. In *ICML*, pages 1026–1034. PMLR, 2014.

G. Benton, M. Finzi, P. Izmailov, and A. G. Wilson. Learning Invariances in Neural Networks from Training Data. *NeurIPS*, 33:17605–17616, 2020.

J. Berkson. Limitations of the Application of Fourfold Table Analysis to Hospital Data. *Biometrics Bulletin*, 2(3):47–53, 1946.

B. Berlin and P. Kay. *Basic Color Terms: Their Universality and Evolution.* University of California Press, Ewing, NJ, USA, 1969.

D. P. Bertsekas. *Nonlinear Programming.* Athena Scientific, 1997.

R. Bhattacharya, T. Nagarajan, D. Malinsky, and I. Shpitser. Differentiable Causal Discovery Under Unmeasured Confounding. In *AISTATS*, pages 2314–2322. PMLR, 2021.

K. Bi, L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619:533–538, 2023.

S. Bickel, M. Brückner, and T. Scheffer. Discriminative Learning Under Covariate Shift. *JMLR*, 10(9), 2009.

C. M. Bishop and N. M. Nasrabadi. *Pattern Recognition and Machine Learning*, volume 4. Springer, 2006.

P. Blöbaum, P. Götz, K. Budhathoki, A. A. Mastakouri, and D. Janzing. DoWhy-GCM: An extension of DoWhy for causal inference in graphical causal models. *arXiv:2206.06821*, 2022.

M. Blondel and V. Roulet. The Elements of Differentiable Programming. *arXiv:2403.14606*, 2024.

E. B. Bonawitz, D. Ferranti, R. Saxe, A. Gopnik, A. N. Meltzoff, J. Woodward, and L. E. Schulz. Just do it? Investigating the gap between prediction and action in toddlers' causal inferences. *Cognition*, 115(1):104–117, 2010.

S. Bond-Taylor, A. Leach, Y. Long, and C. G. Willcocks. Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(11):7327–7347, 2022.

J. Boyle, D. Yeter, M. Aschner, and D. C. Wheeler. Estimated IQ points and lifetime earnings lost to early childhood blood lead levels in the United States. *Sci. Total Environ.*, 778:146307, 2021.

N. Bröckelmann, S. Balduzzi, L. Harms, J. Beyerbach, M. Petropoulou, C. Kubiak, M. Wolkewitz, J. J. Meerpohl, and L. Schwingshackl. Evaluating agreement between bodies of evidence from randomized controlled trials and cohort studies in medical research: a meta-epidemiological study. *BMC Medicine*, 20(1):174, 2022.

S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, et al. Sparks of Artificial General Intelligence: Early experiments with GPT-4. *arXiv:2303.12712*, 2023.

S. Buchholz, G. Rajendran, E. Rosenfeld, B. Aragam, B. Schölkopf, and P. Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing. *NeurIPS*, 36, 2024.

D. Buchsbaum, S. Bridgers, D. S. Weisberg, and A. Gopnik. The power of possibility: causal learning, counterfactual reasoning, and pretend play. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1599): 2202, 2012.

K. Budhathoki and J. Vreeken. MDL for Causal Inference on Discrete Data. In *ICDM*, pages 751–756. IEEE, 2017.

R.-S. Bun, J. Scheer, S. Guillo, F. Tubach, and A. Dechartres. Meta-analyses frequently pooled different study types together: a meta-epidemiological study. *J. Clin. Epidemiol.*, 118:18–28, 2020.

A. Butler, P. Hoffman, P. Smibert, E. Papalexi, and R. Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.*, 36:411–420, 2018.

P. Bühlmann, J. Peters, J. Ernest, et al. CAM: Causal Additive Models, High-Dimensional Order Search and Penalized Regression. *The Annals of Statistics*, 42:2526–2556, 2014.

R. Cai, F. Xie, C. Glymour, Z. Hao, and K. Zhang. Triad Constraints for Learning Causal Structure of Latent Variables. *NeurIPS*, 32, 2019.

R. Cai, Z. Huang, W. Chen, Z. Hao, and K. Zhang. Causal Discovery with Latent Confounders Based on Higher-Order Cumulants. In *ICML*, pages 3380–3407. PMLR, 2023.

C. F. Camerer, A. Dreber, F. Holzmeister, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, G. Nave, B. A. Nosek, T. Pfeiffer, A. Altmejd, N. Buttrick, T. Chan, Y. Chen, E. Forsell, A. Gampa, E. Heikensten, L. Hummer, T. Imai, S. Isaksson, D. Manfredi, J. Rose, E.-J. Wagenmakers, and H. Wu. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nat. Hum. Behav.*, 2:637–644, 2018.

D. Card and A. B. Krueger. Minimum Wages and Employment: A Case Study of the Fast Food Industry in New Jersey and Pennsylvania. *NBER*, 1993.

S. Caton and C. Haas. Fairness in Machine Learning: A Survey. *ACM Computing Surveys*, 56(7):1–38, 2024.

D. Ćevid, P. Bühlmann, and N. Meinshausen. Spectral Deconfounding via Perturbed Sparse Linear Models. *JMLR*, 21(232):1–41, 2020.

V. Chandrasekaran, P. A. Parrilo, and A. S. Willsky. Latent variable graphical model selection via convex optimization. In *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1610–1613. IEEE, 2010.

K. K. Chang, M. Cramer, S. Soni, and D. Bamman. Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7312–7327. Association for Computational Linguistics, 2023.

P. Chao, P. Blöbaum, and S. P. Kasiviswanathan. Interventional and Counterfactual Inference with Diffusion Models. *arXiv:2302.00860*, 2023.

J. Chen, Z. Gao, X. Wu, and J. Luo. Meta-causal learning for single domain generalization. In *CVPR*, pages 7683–7692. IEEE, 2023.

S. Chen, E. Dobriban, and J. H. Lee. A Group-Theoretic Framework for Data Augmentation. *The JMLR*, 21(1):9885–9955, 2020.

W. Chen, M. Drton, and Y. S. Wang. On causal discovery with an equal-variance assumption. *Biometrika*, 106(4):973–980, 2019.

Z. Chen, F. Xie, J. Qiao, Z. Hao, K. Zhang, and R. Cai. Identification of Linear Latent Variable Model with Arbitrary Distribution. *AAAI*, 36(6):6350–6357, 2022.

D. Cheng and T. Cai. Adaptive Combination of Randomized and Observational Data. *arXiv:2111.15012*, 2021.

Y. Cherapanamjeri, C. Daskalakis, A. Ilyas, and M. Zampetakis. What Makes a Good Fisherman? Linear Regression under Self-Selection Bias. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 1699–1712, 2023.

M. Chevalley, Y. Roohani, A. Mehrjou, J. Leskovec, and P. Schwab. CausalBench: A Large-scale Benchmark for Network Inference from Single-cell Perturbation Data. *arXiv:2210.17283*, 2022.

D. M. Chickering. Learning Equivalence Classes of Bayesian-Network Structures. *The JMLR*, 2:445–498, 2002a.

D. M. Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002b.

Y. Chikuse. *Statistics on Special Manifolds*, volume 174. Springer Science & Business Media, 2003.

S. Choi, H. Kang, N. Kim, and J. Kim. How Does Artificial Intelligence Improve Human Decision-Making? Evidence from the AI-Powered Go Program. *arXiv:2310.08704*, 2023.

P. Cisek and J. F. Kalaska. Neural mechanisms for interacting with a world full of action choices. *Annual review of neuroscience*, 33(1):269–298, 2010.

O. M. Cliff, A. G. Bryant, J. T. Lizier, N. Tsuchiya, and B. D. Fulcher. Unifying pairwise interactions in complex dynamics. *Nat. Comput. Sci.*, pages 1–11, 2023.

B. Colnet, I. Mayer, G. Chen, A. Dieng, R. Li, G. Varoquaux, J.-P. Vert, J. Josse, and S. Yang. Causal Inference Methods for Combining Randomized Trials and Observational Studies: A Review. *Statistical Science*, 39:165 – 191, 2024.

D. Colombo, M. H. Maathuis, M. Kalisch, and T. S. Richardson. Learning High-Dimensional Directed Acyclic Graphs with Latent and Selection Variables. *The Annals of Statistics*, 40(1):294–321, 2012.

P. Comon. Independent component analysis, A new concept? *Signal Process.*, 36:287–314, 1994.

R. Compton, L. Zhang, A. Puli, and R. Ranganath. When More is Less: Incorporating Additional Datasets Can Hurt Performance By Introducing Spurious Correlations. In *Machine Learning for Healthcare Conference*, pages 110–127. PMLR, 2023.

D. M. Condon. *An Organizational Framework for the Psychological Individual Diffences: Integrating the Affective, Cognitive, and Conative Domains*. PhD thesis, Northwestern University, 2014.

S. B. Cooper. *Computability theory*. Taylor & Francis, 2017.

B. Corominas-Murtra, R. V. Goé, and C. Rodríguez-Caso. On the origins of hierarchy in complex networks. *PNAS*, 110(33):13316–13321, 2013.

S. Cortes-Gomez, M. Dulce, C. Patino, and B. Wilder. Statistical Inference Under Constrained Selection Bias. *arXiv:2306.03302*, 2023.

N. R. Council. *Issues in the Integration of Research and Operational Satellite Systems for Climate Research: Part I. Science and Design*. 2000.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 1999.

H. Cramér. Über eine Eigenschaft der normalen Verteilungsfunktion. *Math. Z.*, 41(1):405–414, 1936.

A. Crespel, J. Lindström, K. R. Elmer, and S. S. Killen. Evolutionary relationships between metabolism and behaviour require genetic correlations. *Phil. Trans. R. Soc. B*, 379(1896):20220481, 2024.

A. Cutler and D. M. Condon. Deep lexical hypothesis: Identifying personality structure in natural language. *Journal of Personality and Social Psychology*, 2022.

J. O. Dada and P. Mendes. Multi-scale modelling and simulation in systems biology. *Integr. Biol.*, 3(2):86–96, 2011.

I. J. Dahabreh, L. C. Petito, S. E. Robertson, M. A. Hernán, and J. A. Steingrimsson. Toward Causally Interpretable Meta-analysis: Transporting Inferences from Multiple Randomized Trials to a New Target Population. *Epidemiology*, 31(3):334, 2020.

A. D'Amour. On Multi-Cause Approaches to Causal Inference with Unobserved Counfounding: Two Cautionary Failure Cases and A Promising Alternative. In *AISTATS*, pages 3478–3486. PMLR, 2019.

P. Daniusis, D. Janzing, J. Mooij, J. Zscheischler, B. Steudel, K. Zhang, and B. Schölkopf. Inferring deterministic causal relations. In *UAI*, pages 143–150. PMLR, 2012.

T. Dao, A. Gu, A. Ratner, V. Smith, C. De Sa, and C. Ré. A Kernel Theory of Modern Data Augmentation. In *ICML*, pages 1528–1537. PMLR, 2019.

G. Darmois. Analyse générale des liaisons stochastiques: etude particulière de l'analyse factorielle linéaire. *Revue de l'Institut international de statistique*, pages 2–8, 1953.

M. de Carvalho, G. L. Page, and B. J. Barney. On the Geometry of Bayesian Inference. *Bayesian Analysis*, 14(4):1013–1036, 2019.

J. de la Fuente, G. Davies, A. D. Grotzinger, E. M. Tucker-Drob, and I. J. Deary. A general dimension of genetic sharing across diverse cognitive traits inferred from molecular data. *Nat. Hum. Behav.*, 5:49–58, 2021.

B. de Raad and B. Mlačić. The Lexical Foundation of the Big Five Factor Model. *OUP Academic*, 2015.

A. Deaton and N. Cartwright. Understanding and misunderstanding randomized controlled trials. *Social science & medicine*, 210:2–21, 2018.

A. Decelle. An Introduction to Machine Learning: a perspective from Statistical Physics. *Physica A*, 631:128154, 2023.

O. Del Fabbro and P. Christen. Philosophy-Guided Modelling and Implementation of Adaptation and Control in Complex Systems. In *2022 International Joint Conference on Neural Networks*, pages 18–23. IEEE, 2022.

T. Deleu, A. óis, C. Emezue, M. Rankawat, S. Lacoste-Julien, S. Bauer, and Y. Bengio. Bayesian structure learning with generative flow networks. In *UAI*, pages 518–528. PMLR, 2022.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society: series B*, 39(1):1–22, 1977.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, pages 20–25. IEEE, 2009.

K. Desai, B. Nachman, and J. Thaler. Symmetry discovery with deep learning. *Phys. Rev. D*, 105(9), 2022.

Y. Deshpande and A. Montanari. Information-theoretically optimal sparse PCA. In *IEEE International Symposium on Information Theory*, pages 2197–2201. IEEE, 2014.

A. Dhir and M. van der Wilk. Causal Discovery using Bayesian Model Selection. *arXiv:2306.02931*, 2023.

D. Dillion, N. Tandon, Y. Gu, and K. Gray. Can AI language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600, 2023.

C. Ding, M. Gong, K. Zhang, and D. Tao. Likelihood-Free Overcomplete ICA and Applications In Causal Discovery. *NeurIPS*, 32, 2019.

A. A. Dominguez, W. A. Lim, and L. S. Qi. Beyond editing: repurposing CRISPR–Cas9 for precision genome regulation and interrogation. *Nature Reviews Molecular Cell Biology*, 17(1):5–15, 2016.

W. E. Donath and A. J. Hoffman. Algorithms for partitioning of graphs and computer logic based on eigenvectors of connection matrices. *IBM Technical Disclosure Bulletin*, 15(3):938–944, 1972.

R. Durrett. *Probability: Theory and Examples*. Cambridge university press, 2019.

N. Elhage, T. Hume, C. Olsson, N. Schiefer, T. Henighan, S. Kravec, Z. Hatfield-Dodds, R. Lasenby, D. Drain, C. Chen, R. Grosse, S. McCandlish, J. Kaplan, D. Amodei, M. Wattenberg, and C. Olah. Toy Models of Superposition. *Transformer Circuits Thread*, 2022. URL `https://transformer-circuits.pub/2022/toy_model/index.html`.

G. Elidan, N. Lotner, N. Friedman, and D. Koller. Discovering Hidden Variables: A Structure-Based Approach. *NeurIPS*, 13, 2000.

M. R. Elliot. Combining Data from Probability and Non- Probability Samples Using Pseudo-Weights. *Survey Practice*, 2(6), 2009.

J. Eriksson and V. Koivunen. Identifiability, separability, and uniqueness of linear ICA models. *IEEE Signal Processing Letters*, 11(7):601–604, 2004.

R. J. Evans. Graphs for margins of Bayesian networks. *Scandinavian Journal of Statistics*, 43(3):625–648, 2016.

R. J. Evans and T. S. Richardson. Smooth, identifiable supermodels of discrete DAG models with latent variables. *Bernoulli*, 25(2):848–876, 2019.

J. A. Ewing. On the production of transient electric currents in iron and steel conductors by twisting them when magnetised or by magnetising them when twisted. *Proc. R. Soc. Lond.*, 33(216-219):21–23, 1882.

K.-T. Fang, S. Kotz, and K. W. Ng. *Symmetric Multivariate and Related Distributions* . Chapman and Hall/CRC, 1990.

E. Fedorenko, S. T. Piantadosi, and E. A. Gibson. Language is primarily a tool for communication rather than thought. *Nature*, 630(8017):575–586, 2024.

U. Feige, S. Goldwasser, L. Lovász, S. Safra, and M. Szegedy. Approximating clique is almost NP-complete. In *Proceedings 32nd Annual Symposium of Foundations of Computer Science*, pages 2–12. IEEE Computer Society, 1991.

M. A. Figueiredo and C. Oliveira. Distinguishing Cause from Effect on Categorical Data: The Uniform Channel Model. In *Conference on Causal Learning and Reasoning*, pages 122–141. PMLR, 2023.

R. A. Fisher. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7(2):179–188, 1936.

P. Forré and J. M. Mooij. Causal Calculus in the Presence of Cycles, Latent Confounders and Selection Bias. In *UAI*, pages 71–80. PMLR, 2020.

V. Fortuin. Priors in Bayesian Deep Learning: A Review. *International Statistical Review*, 2022.

H. Frankfort and H. A. Groenewegen-Frankfort. *Before Philosophy: The Intellectual Adventure of Ancient Man.* Penguin Books, 1946.

K. Fukushima. Cognitron: A self-organizing multilayered neural network. *Biol. Cybern.*, 20(3):121–136, 1975.

A. Gabel, V. Klein, R. Valperga, J. S. Lamb, K. Webster, R. Quax, and E. Gavves. Learning lie group symmetry transformations with neural networks. In *Topological, Algebraic and Geometric Learning Workshops 2023*, pages 50–59. PMLR, 2023.

F. J. Gallego, N. Kussul, S. Skakun, O. Kravchenko, A. Shelestov, and O. Kussul. Efficiency assessment of using satellite data for crop area estimation in Ukraine. *Int. J. Appl. Earth Obs. Geoinf.*, 29:22–30, 2014.

F. Galton. *The Measurement of Character.* Prentice-Hall, 1884.

J. L. Gamella, A. Taeb, C. Heinze-Deml, and P. Bühlmann. Characterization and Greedy Learning of Gaussian Structural Causal Models under Unknown Interventions. *arXiv:2211.14897*, 2022.

T. Gao, K. Fadnis, and M. Campbell. Local-to-Global Bayesian Network Structure Learning. In *ICML*, pages 1193–1202. PMLR, 2017.

P. Gärdenfors. *Conceptual Spaces: The Geometry of Thought.* MIT press, 2004.

P. Gärdenfors. *The Geometry of Meaning: Semantics Based on Conceptual Spaces.* MIT press, 2014.

S. A. Gelman and C. H. Legare. Concepts and Folk Theories. *Annual Review of Anthropology*, 40:379, 2011.

A. S. Gerber, G. A. Huber, D. Doherty, and C. M. Dowling. The Big Five Personality Traits in the Political Arena. *Annual Review of Political Science*, 14:265–287, 2011.

A. Ghassami, A. Yang, I. Shpitser, and E. T. Tchetgen. Causal Inference with Hidden Mediators. *arXiv:2111.02927*, 2021.

M. Girardeau and R. Mazo. Variational Methods in Statistical Mechanics. *Advances in Chemical Physics*, pages 187–255, 1973.

C. Glymour, K. Zhang, and P. Spirtes. Review of Causal Discovery Methods Based on Graphical Models. *Frontiers in Genetics*, 10:524, 2019.

M. M. Glymour and S. Greenland. Causal Diagrams. *Modern Epidemiology*, 3:183–209, 2008.

L. H. Goetz and N. J. Schork. Personalized medicine: motivation, challenges, and progress. *Fertil. Steril.*, 109(6):952–963, 2018.

N. D. Goodman, T. D. Ullman, and J. B. Tenenbaum. Learning a theory of causality. *Psychological Review*, 118(1):110, 2011.

M. Gordon, D. Viganola, M. Bishop, Y. Chen, A. Dreber, B. Goldfedder, F. Holzmeister, M. Johannesson, Y. Liu, C. Twardy, J. Wang, and T. Pfeiffer. Are replication rates the same across academic fields? Community forecasts from the DARPA SCORE programme. *R. Soc. Open Sci.*, 7(7):200566, 2020.

K. B. Gorman, T. D. Williams, and W. R. Fraser. Ecological Sexual Dimorphism and Environmental Variability within a Community of Antarctic Penguins (Genus Pygoscelis). *PloS One*, 9(3):e90081, 2014.

B. Gourévitch, B.-J. Le, and G. Faucon. Linear and nonlinear causality between signals: methods, examples and neurophysiological applications. *Biol. Cybern.*, 95(4):349–369, 2006.

L. Gresele, J. Von Kügelgen, V. Stimper, B. Schölkopf, and M. Besserve. Independent Mechanism Analysis, a New Concept? *NeurIPS*, 34:28233–28248, 2021.

A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, and B. Schölkopf. Covariate Shift by Kernel Mean Matching. *Dataset Shift in Machine Learning*, 3(4):5, 2009.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A Kernel Two-Sample Test. *JMLR*, 13(25):723–773, 2012.

S. D. Grosse and Y. Zhou. Monetary Valuation of Children's Cognitive Outcomes in Economic Evaluations from a Societal Perspective: A Review. *Children*, 8(5), 2021.

J. Grossman and F. J. Mackenzie. The Randomized Controlled Trial: gold standard, or merely standard? *Perspectives in Biology and Medicine*, 48(4): 516–534, 2005.

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

A. Grzegorczyk. On the definitions of computable real continuous functions. *Fund. Math.*, 44:61–71, 1957.

S. Guo, V. Tóth, B. Schölkopf, and F. Huszár. Causal de Finetti: On the identification of invariant causal structure in exchangeable data. *NeurIPS*, 36, 2024.

W. Gurnee and M. Tegmark. Language Models Represent Space and Time. *arXiv:2310.02207*, 2023.

I. Guyon, C. Aliferis, G. Cooper, A. Elisseeff, J.-P. Pellet, P. Spirtes, and A. Statnikov. Datasets of the Causation and Prediction Challenge. Technical report, 2008.

D. Ha and J. Schmidhuber. World Models. *arXiv:1803.10122*, 2018.

A. Hajat, R. F. MacLehose, A. Rosofsky, K. D. Walker, and J. E. Clougherty. Confounding by Socioeconomic Status in Epidemiological Studies of Air Pollution and Health: Challenges and Opportunities. *Environ. Health Perspect.*, 2021.

D. M. Haughton. On the choice of a model to fit data from an exponential family. *The annals of statistics*, pages 342–355, 1988.

K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *CVPR*, pages 1026–1034. IEEE, 2015a.

Y. He, J. Jia, and B. Yu. Counting and Exploring Sizes of Markov Equivalence Classes of Directed Acyclic Graphs. *The JMLR*, 16(1):2589–2609, 2015b.

L. G. Hemkens, D. G. Contopoulos-Ioannidis, and J. P. A. Ioannidis. Routinely collected data and comparative effectiveness evidence: promises and limitations. *CMAJ*, 188(8):E158–E164, 2016.

A. Herbert, G. Griffith, G. Hemani, and L. Zuccolo. The spectre of Berkson's paradox: Collider bias in Covid-19 research. *Significance*, 17(4):6–7, 2020.

G. M. Hodgson. *Economics and Evolution: Bringing Life Back into Economics.* University of Michigan Press, 1996.

A. Hooda, M. Christodorescu, M. Allamanis, A. Wilson, K. Fawaz, and S. Jha. Do Large Code Models Understand Programming Concepts? A Black-box Approach. *arXiv:2402.05980*, 2024.

P. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. *NeurIPS*, 21, 2008a.

P. O. Hoyer, S. Shimizu, A. J. Kerminen, and M. Palviainen. Estimation of causal effects using linear non-Gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49:362–378, 2008b.

Y. Hu, F. J. van Lenthe, and J. P. Mackenbach. Income inequality, life expectancy and cause-specific mortality in 43 European countries, 1987 extendash2008: a fixed effects study. *Eur. J. Epidemiol.*, 30(8):615, 2015.

Y. Hu, Y. Wu, L. Zhang, and X. Wu. A Generative Adversarial Framework for Bounding Confounded Causal Effects. *AAAI*, 35(13):12104–12112, 2021.

B. Huang, K. Zhang, J. Zhang, J. Ramsey, R. Sanchez-Romero, C. Glymour, and B. Schölkopf. Causal Discovery from Heterogeneous/Nonstationary Data. *JMLR*, 21(1), 2020.

B. Huang, C. J. H. Low, F. Xie, C. Glymour, and K. Zhang. Latent Hierarchical Causal Structure Discovery with Rank Constraints. *NeurIPS*, 35:5549–5561, 2022.

A. Hyvärinen and P. Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439, 1999.

G. W. Imbens and T. Lancaster. Combining Micro and Macro Data in Microeconometric Models. *Rev. Econom. Stud.*, 61(4):655–680, 1994.

G. W. Imbens and T. Lemieux. Regression discontinuity designs: A guide to practice. *J. Econometrics*, 142(2):615–635, 2008.

A. Immer, T. F. van der Ouderaa, V. Fortuin, G. Rätsch, and M. van der Wilk. Invariance Learning in Deep Neural Networks with Differentiable Laplace Approximations. *NeurIPS*, 35:12449–12463.

D. Janzing and B. Schölkopf. Causal inference using the algorithmic Markov condition. *IEEE Trans. Inf. Theory*, 56:5168–5194, 2010.

D. Janzing and B. Schölkopf. Detecting Confounding in Multivariate Linear Models via Spectral Analysis. *Journal of Causal Inference*, 6(1), 2018.

D. Janzing and B. Schölkopf. Detecting Non-Causal Artifacts in Multivariate Linear Regression Models. In *ICML*, pages 2245–2253. JMLR, 2018.

D. Janzing, J. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniušis, B. Steudel, and B. Schölkopf. Information-geometric approach to inferring causal directions. *Artif. Intell.*, 182:1–31, 2012.

D. Janzing, B. Steudel, N. Shajarisales, and B. Schölkopf. Justifying Information-Geometric Causal Inference. In *Measures of Complexity*, pages 253–265. Springer, 2015.

P. Jaworski, F. Durante, W. K. Hardle, and T. Rychlik. *Copula Theory and Its Applications*, volume 198. Springer, 2010.

J. Jhun. Economics, Equilibrium Methods, and Multi-Scale Modeling. *Erkenntnis*, 86(2):457–472, 2021.

H. Jiang, L. Ge, Y. Gao, J. Wang, and R. Song. Large Language Model for Causal Decision Making. *arXiv:2312.17122*, 2023a.

Z. Jiang, M. Yang, M. Tsirlin, R. Tang, Y. Dai, and J. Lin. "Low-Resource" Text Classification: A Parameter-Free Classification Method with Compressors. *ACL Anthology*, pages 6810–6828, 2023b.

Z. Jin, J. Liu, Z. Lyu, S. Poff, M. Sachan, R. Mihalcea, M. Diab, and B. Schölkopf. Can Large Language Models Infer Causation from Correlation? *arXiv:2306.05836*, 2023.

T. Jiralerspong, X. Chen, Y. More, V. Shah, and Y. Bengio. Efficient Causal Graph Discovery Using Large Language Models. *arXiv:2402.01207*, 2024.

W. Johnson and T. J. Bouchard. The structure of human intelligence: It is verbal, perceptual, and image rotation (VPR), not fluid and crystallized. *Intelligence*, 33(4):393–416, 2005.

D. Kahneman and A. Tversky. rospect Theory: An Analysis of Decision Under Risk. In *Handbook of the fundamentals of financial decision making: Part I*, pages 99–127. World Scientific, 2013.

O. Kallenberg. *Probabilistic Symmetries and Invariance Principles*. Springer, 2005.

N. Kallus, X. Mao, and M. Udell. Causal Inference with Noisy and Missing Covariates via Matrix Factorization. *NeurIPS*, pages 6920–6931, 2018.

D. Kaltenpoth and J. Vreeken. We Are Not Your Real Parents: Telling Causal from Confounded using MDL. In *SDM*, pages 199–207. SIAM, 2019.

D. Kaltenpoth and J. Vreeken. Causal Discovery with Hidden Confounders Using the Algorithmic Markov Condition. In *UAI*, pages 1016–1026. PMLR, 2023a.

D. Kaltenpoth and J. Vreeken. Identifying Selection Bias from Observational Data. *AAAI*, 37(7):8177–8185, 2023b.

D. Kaltenpoth and J. Vreeken. Nonlinear Causal Discovery with Latent Confounders. In *ICML*, pages 15639–15654. PMLR, 2023c.

M. Kang, B. G. Ragan, and J.-H. Park. Issues in Outcomes Research: An Overview of Randomization Techniques for Clinical Trials. *Journal of Athletic Training*, 43(2):215–221, 2008.

R. Karlsson and J. Krijthe. Detecting hidden confounding in observational data using multiple environments. In *NeurIPS*, volume 36, pages 44280–44309, 2023.

A. Karvonen. Emergent World Models and Latent Variable Estimation in Chess-Playing Language Models. *arXiv:2403.15498*, 2024.

S. A. Kauffman. *At Home in the Universe: The Search for Laws of Self-organization and Complexity*. Oxford University Press, USA, 1995.

I. Kaźmierczak, A. Zajenkowska, R. Rogoza, P. K. Jonason, and D. Ścigała. Self-selection biases in psychological studies: Personality and affective disorders are prevalent among participants. *PLoS One*, 18(3), 2023.

G. Keropyan, D. Strieder, and M. Drton. Rank-Based Causal Discovery for Post-Nonlinear Models. In *AISTATS*, pages 7849–7870. PMLR, 2023.

M. Kertel and N. Klein. Boosting Causal Additive Models. *arXiv:2401.06523*, 2024.

I. Khemakhem, D. Kingma, R. Monti, and A. Hyvärinen. Variational Autoencoders and Nonlinear ICA: A Unifying Framework. In *AISTATS*, pages 2207–2217. PMLR, 2020.

I. Khemakhem, R. Monti, R. Leech, and A. HyvarinenHyvärinen. Causal Autoregressive Flows. In *AISTATS*, pages 3520–3528. PMLR, 2021.

E. Kıcıman, R. Ness, A. Sharma, and C. Tan. Causal Reasoning and Large Language Models: Opening a New Frontier for Causality. *arXiv:2305.00050*, 2023.

Y. Kim, G. R. B. Saunders, A. Giannelis, E. A. Willoughby, C. G. DeYoung, and J. J. Lee. Genetic and neural bases of the neuroticism general factor. *Biol. Psychol.*, 184:108692, 2023.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014.

D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. In *ICLR*, 2014.

D. P. Kingma and M. Welling. *An Introduction to Variational Autoencoders*. Now Foundations and Trends, 2019.

L. Kirfel, T. Icard, and T. Gerstenberg. Inference from explanation. *Journal of Experimental Psychology: General*, 151(7):1481, 2022.

K.-R. Kladny, J. von Kügelgen, B. Schölkopf, and M. Muehlebach. Causal effect estimation from observational and interventional data through matrix weighted linear estimators. In *UAI*, pages 1087–1097. PMLR, 2023.

M. Kocaoglu, A. G. Dimakis, S. Vishwanath, and B. Hassibi. Entropic Causal Inference. In *AAAI*, pages 1156–1162, 2017.

M. Kocaoglu, S. Shakkottai, A. G. Dimakis, C. Caramanis, and S. Vishwanath. Entropic Latent Variable Discovery. *arXiv:1807.10399*, 2018.

M. Kocaoglu, A. Jaber, K. Shanmugam, and E. Bareinboim. Characterization and Learning of Causal Graphs with Latent Variables from Soft Interventions. In *NeurIPS*, pages 14346–14356. 2019.

B. Koch, E. Denton, A. Hanna, and J. G. Foster. Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research. In *NeurIPS Datasets and Benchmarks*. 2021.

R. Kohavi, D. Tang, Y. Xu, L. G. Hemkens, and J. P. A. Ioannidis. Online randomized controlled experiments at scale: lessons and extensions to medicine. *Trials*, 21, 2020.

W. Köhler. Gestalt Psychology. *Psychologische Forschung*, 31(1):XVIII–XXX, 1967.

M. Koivisto and S. Grassini. Best humans still outperform artificial intelligence in a creative divergent thinking task. *Sci. Rep.*, 13(13601):1–10, 2023.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT press, 2009.

M. Kosinski. Evaluating Large Language Models in Theory of Mind Tasks. *arXiv:2302.02083*, 2023.

S. Kotz and S. Nadarajah. *Multivariate t-Distributions and Their Applications*. Cambridge University Press, 2004.

B. Kovács and A. J. Sharkey. The Paradox of Publicity: How Awards Can Negatively Affect the Evaluation of Quality. *Administrative Science Quarterly*, 59(1):1–33, 2014.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *NeurIPS*, 25, 2012.

A. Kucukelbir, D. Tran, R. Ranganath, A. Gelman, and D. M. Blei. Automatic Differentiation Variational Inference. *JMLR*, 18:430–474, 2017.

E. Kummerfeld and J. Ramsey. Causal Clustering for 1-Factor Measurement Models. In *KDD*, pages 1655–1664. ACM, 2016.

R. Kundu, X. Shi, J. Morrison, J. Barrett, and B. Mukherjee. A framework for understanding selection bias in real-world healthcare data. *J. R. Stat. Soc. Ser. A Stat. Soc.*, page qe039, 2024.

M. Kuroki and Z. Cai. On recovering a population covariance matrix in the presence of selection bias. *Biometrika*, 93(3):601–611, 2006.

E. Kwessi. Information Theory in a Darwinian Evolution Population Dynamics Model. *arXiv:2403.05044*, 2024.

S. Lachapelle, P. Brouillard, T. Deleu, and S. Lacoste-Julien. Gradient-Based Neural DAG Learning. In *ICLR*, 2019.

G. Lakoff and M. Johnson. *Metaphors We Live By*. University of Chicago press, 1980.

G. Lakoff and R. Núñez. *Where Mathematics Comes From*. New York: Basic Books, 2000.

L. Landeck, C. Kneip, J. Reischl, and K. Asadullah. Biomarkers and personalized medicine: current status and further perspectives with special focus on dermatology. *Experimental Dermatology*, 25(5):333–339, 2016.

C. G. Langton. Computation at the edge of chaos: Phase transitions and emergent computation. *Physica D: Nonlinear Phenomena*, 42(1-3):12–37, 1990.

C. Lauditi, E. Troiani, and M. Mézard. Sparse Representations, Inference and Learning. *arXiv:2306.16097*, 2023.

N. Lawrence. Probabilistic Non-linear Principal Component Analysis with Gaussian Process Latent Variable Models. *JMLR*, 6:1783–1816, 2005.

R. D. Leclerc. Survival of the sparsest: robust gene networks are parsimonious. *Mol. Syst. Biol.*, 4(1):213, 2008.

J. Lee, A. Wibisono, and E. Zampetakis. Learning Exponential Families from Truncated Samples. *NeurIPS*, 36, 2024.

K. Lee and M. C. Ashton. Psychometric Properties of the HEXACO Personality Inventory. *Multivariate Behavioral Research*, pages 329–358, 2010.

P. Lehnert, M. Niederberger, U. Backes-Gellner, and E. Bettinger. Proxying economic activity with daytime satellite imagery: Filling data gaps across time and space. *PNAS Nexus*, 2(4), 2023.

J. Lemeire and D. Janzing. Replacing Causal Faithfulness with Algorithmic Independence of Conditionals. *Minds &. Machines*, 23(2):227–249, 2013.

C. Li and J. Flanigan. Task Contamination: Language Models May Not Be Few-Shot Anymore. *AAAI*, 38(16):18471–18480, 2024.

M. Li and P. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer Science & Business Media, 2009.

W. Li, Y. Li, S. Zhu, Y. Shao, J. Hao, and Y. Pang. GFlowCausal: Generative Flow Networks for Causal Discovery. *arXiv:2210.08185*, 2022.

W. Liang, A. Kekić, J. von Kügelgen, S. Buchholz, M. Besserve, L. Gresele, and B. Schölkopf. Causal Component Analysis. *NeurIPS*, 36:32481–32520, 2023a.

W. Liang, Y. Zhang, H. Cao, B. Wang, D. Ding, X. Yang, K. Vodrahalli, S. He, D. Smith, Y. Yin, D. McFarland, and J. Zou. Can large language models provide useful feedback on research papers? A large-scale empirical analysis. *arXiv:2310.01783*, 2023b.

A. Liddle. *An Introduction to Modern Cosmology*. John Wiley & Sons, 2015.

H. Lin and H. Peng. Smoothed rank correlation of the linear transformation regression model. *Comput. Statist. Data Anal.*, 57(1):615–630, 2013.

P. Lippe, S. Magliacane, S. Löwe, Y. M. Asano, T. Cohen, and E. Gavves. BISCUIT: Causal Representation Learning from Binary Interactions. In *UAI*, pages 1263–1273. PMLR, 2023.

M. Liu, X. Sun, Y. Qiao, and Y. Wang. Causal Discovery with Unobserved Variables: A Proxy Variable Approach. *arXiv:2305.05281*, 2023.

F. Locatello, S. Bauer, M. Lucic, G. Rätsch, S. Gelly, B. Schölkopf, and O. Bachem. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. In *ICML*, pages 4114–4124. PMLR, 2019.

J. C. Loehlin. *Latent Variable Models: An Introduction to Factor, Path, and Structural Analysis*. Psychology Press, 1998.

T. M. Ludden. Nonlinear Pharmacokinetics: Clinical Implications. *Clin. Pharmacokinet.*, 20(6):429–446, 1991.

C. Lyle, M. van der Wilk, M. Kwiatkowska, Y. Gal, and B. Bloem-Reddy. On the Benefits of Invariance in Neural Networks. *arXiv:2005.00178*, 2020.

A. L. Maas, A. Y. Hannun, A. Y. Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *ICML*, volume 30, page 3, 2013.

A. Mallick, K. Hsieh, B. Arzani, and G. Joshi. Matchmaker: Data Drift Mitigation in Machine Learning for Large-Scale Systems. *Proceedings of Machine Learning and Systems*, 4:77–94, 2022.

S. Mameche, D. Kaltenpoth, and J. Vreeken. Discovering Invariant and Changing Mechanisms from Data. In *KDD*, pages 1242–1252. ACM, 2022.

S. Mameche, D. Kaltenpoth, and J. Vreeken. Learning Causal Mechanisms under Independent Changes. volume 36, pages 75595–75622, 2023.

S. Mameche, J. Vreeken, and D. Kaltenpoth. Identifying Confounding from Causal Mechanism Shifts. In *AISTATS*. PMLR, 2024.

R. Marino. Where do hard problems really exist? *arXiv:2309.16253*, 2023.

G. Markozannes, G. Vourli, and E. Ntzani. A survey of methodologies on causal inference methods in meta-analyses of randomized controlled trials. *Syst. Rev.*, 10(1):1–9, 2021.

S. Marks and M. Tegmark. The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets. *arXiv:2310.06824*, 2023.

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

A. Marx and J. Vreeken. Causal Inference on Multivariate and Mixed-Type Data. *ECMLPKDD*, pages 655–671, 2019.

A. Marx and J. Vreeken. Formally Justifying MDL-Based Inference of Cause and Effect. In *AAAI Workshop ITCI*, 2022.

A. Maul. Rethinking Traditional Methods of Survey Validation. *Measurement: Interdisciplinary Research and Perspectives*, 15(2):51–69, 2017.

R. T. McCoy, S. Yao, D. Friedman, M. Hardy, and T. L. Griffiths. Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve. *arXiv:2309.13638*, 2023.

R. R. McCrae, A. B. Zonderman, P. T. Costa Jr, M. H. Bond, and S. V. Paunonen. Evaluating replicability of factors in the Revised NEO Personality Inventory: Confirmatory factor analysis versus Procrustes rotation. *Journal of Personality and Social Psychology*, 70(3):552, 1996.

C. Meek. *Complete orientation rules for patterns.* Carnegie Mellon University, 1995.

J. Mefford and J. S. Witte. The Covariate's Dilemma. *PLoS Genetics*, 8(11), 2012.

L. Meincke, E. R. Mollick, and C. Terwiesch. Prompting Diverse Ideas: Increasing AI Idea Variance. *arXiv:2402.01727*, 2024.

A. N. Meltzoff. Infants' causal learning: Intervention, observation, imitation. *OUP Academic*, 2007.

A. Mey and R. M. Castro. Invariant Causal Prediction with Locally Linear Models. *arXiv:2401.05218*, 2024.

M. Mezard and A. Montanari. *Information, Physics, and Computation.* Oxford University Press, 2009.

O. A. Mian, A. Marx, and J. Vreeken. Discovering fully oriented causal networks. In *AAAI*, volume 35, pages 8975–8982, 2021.

P. Mirowski. An Evolutionary Theory of Economics Change: A Review Article. *Journal of Economic Issues*, 17(3):757–768, 1983.

R. P. Monti, I. Khemakhem, and A. Hyvarinen. Autoregressive flow-based causal discovery and inference. *arXiv:2007.09390*, 2020.

J. Mooij, S. Magliacane, and T. Claassen. Joint Causal Inference from Multiple Contexts. *JMLR*, 21:99:1–99:108, 2020.

J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf. Distinguishing Cause from Effect Using Observational Data: Methods and Benchmarks. *JMLR*, 17(1):1103–1204, 2016.

K. Muandet, K. Fukumizu, B. Sriperumbudur, B. Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

H. Naci and J. P. A. Ioannidis. Comparative effectiveness of exercise and drug interventions on mortality outcomes: metaepidemiological study. *BMJ*, 347: f5577, 2013.

S. Nazlioglu. World oil and agricultural commodity prices: Evidence from nonlinear causality. *Energy Policy*, 39(5):2935–2943, 2011.

Y. Nishiyama, K. Hitomi, Y. Kawasaki, and K. Jeong. A consistent nonparametric test for nonlinear causality—Specification in time series regression. *J. Econometrics*, 165(1):112–127, 2011.

J. M. Ogarrio, P. Spirtes, and J. Ramsey. A Hybrid Causal Search Algorithm for Latent Variable Models. In *Conference on Probabilistic Graphical Models*, pages 368–379, 2016.

T. X. Olausson, J. P. Inala, C. Wang, J. Gao, and A. Solar-Lezama. Is Self-Repair a Silver Bullet for Code Generation? *arXiv:2306.09896*, 2023.

Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251), 2015.

OpenAI, :, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Baltescu, H. Bao, M. Bavarian, J. Belgum, I. Bello, J. Berdine, G. Bernadett-Shapiro, C. Berner, L. Bogdonoff, O. Boiko, M. Boyd, A.-L. Brakman, G. Brockman, T. Brooks, M. Brundage, K. Button, T. Cai, R. Campbell, A. Cann, B. Carey, C. Carlson, R. Carmichael, B. Chan, C. Chang, F. Chantzis, D. Chen, S. Chen, R. Chen, J. Chen, M. Chen, B. Chess, C. Cho, C. Chu, H. W. Chung, D. Cummings, J. Currier, Y. Dai, C. Decareaux, T. Degry, N. Deutsch, D. Deville, A. Dhar, D. Dohan, S. Dowling, S. Dunning, A. Ecoffet, A. Eleti, T. Eloundou, D. Farhi, L. Fedus, N. Felix, S. P. Fishman, J. Forte, I. Fulford, L. Gao, E. Georges, C. Gibson, V. Goel, T. Gogineni, G. Goh, R. Gontijo-Lopes, J. Gordon, M. Grafstein, S. Gray, R. Greene, J. Gross, S. S. Gu, Y. Guo, C. Hallacy, J. Han, J. Harris, Y. He, M. Heaton, J. Heidecke, C. Hesse, A. Hickey, W. Hickey, P. Hoeschele, B. Houghton, K. Hsu, S. Hu, X. Hu, J. Huizinga, S. Jain, S. Jain, J. Jang, A. Jiang, R. Jiang, H. Jin,

D. Jin, S. Jomoto, B. Jonn, H. Jun, T. Kaftan, Ł. Kaiser, A. Kamali, I. Kanitscheider, N. S. Keskar, T. Khan, L. Kilpatrick, J. W. Kim, C. Kim, Y. Kim, J. H. Kirchner, J. Kiros, M. Knight, D. Kokotajlo, Ł. Kondraciuk, A. Kondrich, A. Konstantinidis, K. Kosic, G. Krueger, V. Kuo, M. Lampe, I. Lan, T. Lee, J. Leike, J. Leung, D. Levy, C. M. Li, R. Lim, M. Lin, S. Lin, M. Litwin, T. Lopez, R. Lowe, P. Lue, A. Makanju, K. Malfacini, S. Manning, T. Markov, Y. Markovski, B. Martin, K. Mayer, A. Mayne, B. McGrew, S. M. McKinney, C. McLeavey, P. McMillan, J. McNeil, D. Medina, A. Mehta, J. Menick, L. Metz, A. Mishchenko, P. Mishkin, V. Monaco, E. Morikawa, D. Mossing, T. Mu, M. Murati, O. Murk, D. Mély, A. Nair, R. Nakano, R. Nayak, A. Neelakantan, R. Ngo, H. Noh, L. Ouyang, C. O'Keefe, J. Pachocki, A. Paino, J. Palermo, A. Pantuliano, G. Parascandolo, J. Parish, E. Parparita, A. Passos, M. Pavlov, A. Peng, A. Perelman, F. d. A. B. Peres, M. Petrov, H. P. d. O. Pinto, Michael, Pokorny, M. Pokrass, V. H. Pong, T. Powell, A. Power, B. Power, E. Proehl, R. Puri, A. Radford, J. Rae, A. Ramesh, C. Raymond, F. Real, K. Rimbach, C. Ross, B. Rotsted, H. Roussez, N. Ryder, M. Saltarelli, T. Sanders, S. Santurkar, G. Sastry, H. Schmidt, D. Schnurr, J. Schulman, D. Selsam, K. Sheppard, T. Sherbakov, J. Shieh, S. Shoker, P. Shyam, S. Sidor, E. Sigler, M. Simens, J. Sitkin, K. Slama, I. Sohl, B. Sokolowsky, Y. Song, N. Staudacher, F. P. Such, N. Summers, I. Sutskever, J. Tang, N. Tezak, M. B. Thompson, P. Tillet, A. Tootoonchian, E. Tseng, P. Tuggle, N. Turley, J. Tworek, J. F. C. Uribe, A. Vallone, A. Vijayvergiya, C. Voss, C. Wainwright, J. J. Wang, A. Wang, B. Wang, J. Ward, J. Wei, C. J. Weinmann, A. Welihinda, P. Welinder, J. Weng, L. Weng, M. Wiethoff, D. Willner, C. Winter, S. Wolrich, H. Wong, L. Workman, S. Wu, J. Wu, M. Wu, K. Xiao, T. Xu, S. Yoo, K. Yu, Q. Yuan, W. Zaremba, R. Zellers, C. Zhang, M. Zhang, S. Zhao, T. Zheng, J. Zhuang, W. Zhuk, and B. Zoph. GPT-4 Technical Report. *arXiv:2303.08774*, 2023.

P. A. Ortega, M. Kunesch, G. Delétang, T. Genewein, J. Grau-Moya, J. Veness, J. Buchli, J. Degrave, B. Piot, J. Perolat, T. Everitt, C. Tallec, E. Parisotto, T. Erez, Y. Chen, S. Reed, M. Hutter, N. de Freitas, and S. Legg. Shaking the foundations: delusions in sequence models for interaction and control. *arXiv:2110.10819*, 2021.

E. Oster. Unobservable Selection and Coefficient Stability: Theory and Validation. *NBER*, 2013.

P. K. Parida, T. Marwala, and S. Chakraverty. A Multivariate Additive Noise Model for Complete Causal Discovery. *Neural Networks*, 103:44–54, 2018.

J. Park, U. Shalit, B. Schölkopf, and K. Muandet. Conditional Distributional Treatment Effect with Kernel Conditional Mean Embeddings and U-Statistic Regression. In *ICML*, volume 139, pages 8401–8412. PMLR, 2021.

P. S. Park, S. Goldstein, A. O'Gara, M. Chen, and D. Hendrycks. AI deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5), 2024.

P. Parsana, C. Ruberman, A. E. Jaffe, M. C. Schatz, A. Battle, and J. T. Leek. Addressing confounding artifacts in reconstruction of gene co-expression networks. *Genome Biol.*, 20(1):1–6, 2019.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *NeurIPS*, 32, 2019.

R. Patel and E. Pavlick. Mapping Language Models to Grounded Conceptual Spaces. In *ICLR*, 2022.

A. Paullada, I. D. Raji, E. M. Bender, E. Denton, and A. Hanna. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336, 2021.

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

J. Pearl and T. S. Verma. A theory of inferred causation. In *Studies in Logic and the Foundations of Mathematics*, volume 134, pages 789–811. Elsevier, 1995.

B.-E. Perrin, L. Ralaivola, A. Mazurie, S. Bottani, J. Mallet, and F. d' Alché–Buc. Gene Networks Inference Using Dynamic Bayesian Networks. *Bioinformatics*, 19:138–148, 2003.

R. Perry, J. V. Kügelgen, and B. Schölkopf. Causal Discovery in Heterogeneous Environments Under the Sparse Mechanism Shift Hypothesis. In *NeurIPS*, 2022.

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

J. Peters and P. Bühlmann. Structural Intervention Distance for Evaluating Causal Graphs. *Neural Comput.*, 27(3):771–799, 2015.

J. Peters, J. M. Mooij, D. Janzing, B. Schölkopf, et al. Causal Discovery with Continuous Additive Noise Models . *JMLR*, 15:2009–2053, 2014.

J. Peters, P. Bühlmann, and N. Meinshausen. Causal Inference by using Invariant Prediction: Identification and Confidence Intervals. 78(5):947–1012, 2016.

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

V. Pham and S. Cunningham. ChatGPT Can Predict the Future when it Tells Stories Set in the Future About the Past. *arXiv:2404.07396*, 2024.

Plato. *Phaedrus*. Cambridge University Press, 1952.

R. J. Prill, D. Marbach, J. Saez-Rodriguez, P. K. Sorger, L. G. Alexopoulos, X. Xue, N. D. Clarke, G. Altan-Bonnet, and G. Stolovitzky. Towards a rigorous assessment of systems biology models: the DREAM3 challenges. *PloS One*, 5(2), 2010.

X. Pu, M. Gao, and X. Wan. Summarization is (Almost) Dead. *arXiv:2309.09558*, 2023.

P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *ACL Anthology*, pages 2383–2392, 2016.

R. Ramb, M. Eichler, A. Ing, M. Thiel, C. Weiller, C. Grebogi, C. Schwarzbauer, J. Timmer, and B. Schelter. The impact of latent confounders in directed network analysis in neuroscience. *Philos. Trans. Royal Soc. A*, 371(1997):20110612, 2013.

J. Ramsey and B. Andrews. Py-Tetrad and RPy-Tetrad: A New Python Interface. In *Causal Analysis Workshop Series*, pages 40–51. PMLR, 2023.

J. Ramsey, M. Glymour, R. Sanchez-Romero, and C. Glymour. A million variables and more: The Fast Greedy Equivalence Search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *Int J of Data Sci Anal.*, 3:121–129, 2017.

M. J. D. Ramstead, D. A. R. Sakthivadivel, C. Heins, M. Koudahl, B. Millidge, L. Da Costa, B. Klein, and K. J. Friston. On Bayesian mechanics: a physics of and by beliefs. *Interface Focus*, 13(3):20220029, Apr. 2023. ISSN 2042-8901. doi: 10.1098/rsfs.2022.0029.

R. Ranganath and A. Perotte. Multiple Causal Inference with Latent Confounding. *arXiv:1805.08273*, 2018.

R. Ranganath, L. Tang, L. Charlin, and D. Blei. Deep Exponential Families. In *AISTATS*, pages 762–771. PMLR, 2015.

R. Rashid, J. Chowdhury, and G. Terejanu. From Causal Pairs to Causal Graphs. In *ICMLA*, pages 802–807. IEEE, 2022.

G. Raskutti and C. Uhler. Learning directed acyclic graph models based on sparsest permutations. *Stat*, 7(1), 2018.

S. Ravanbakhsh, J. Schneider, and B. Poczos. Equivariance Through Parameter-Sharing. In *ICML*, pages 2892–2901. PMLR, 2017.

V. J. Reddi, C. Cheng, D. Kanter, P. Mattson, G. Schmuelling, C.-J. Wu, B. Anderson, M. Breughe, M. Charlebois, W. Chou, R. Chukka, C. Coleman, S. Davis, P. Deng, G. Diamos, J. Duke, D. Fick, J. S. Gardner, I. Hubara, S. Idgunji, T. B. Jablin, J. Jiao, T. S. John, P. Kanwar, D. Lee, J. Liao, A. Lokhmotov, F. Massa, P. Meng, P. Micikevicius, C. Osborne, G. Pekhimenko, A. T. R. Rajan, D. Sequeira, A. Sirasao, F. Sun, H. Tang, M. Thomson, F. Wei, E. Wu, L. Xu, K. Yamada, B. Yu, G. Yuan, A. Zhong, P. Zhang, and Y. Zhou. MLPerf Inference Benchmark. In *International Symposium on Computer Architecture*, pages 2020–03. IEEE, 2020.

A. G. Reddy, S. Dash, A. Sharma, and V. N. Balasubramanian. Counterfactual Generation Under Confounding. In *NeurIPS Workshop on Causality for Real-world Impact*, 2022.

H. Rein. A proposal for community driven and decentralized astronomical databases and the Open Exoplanet Catalogue. *arXiv:1211.7121*, 2012.

A. Reisach, C. Seiler, and S. Weichwald. Beware of the Simulated DAG! Causal Discovery Benchmarks May Be Easy to Game. *NeurIPS*, 34:27772–27784, 2021.

L. Rendsburg, L. C. Vankadara, D. Ghoshdastidar, and U. von Luxburg. A Consistent Estimator for Confounding Strength. *arXiv:2211.01903*, 2022.

D. Rezende and S. Mohamed. Variational Inference with Normalizing Flows. In *ICML*, pages 1530–1538. PMLR, 2015.

T. S. Richardson, R. J. Evans, J. M. Robins, and I. Shpitser. Nested Markov properties for acyclic directed mixed graphs, 2012.

E. A. Rios, W.-H. Cheng, and B.-C. Lai. DAF:re: A Challenging, Crowd-Sourced, Large-Scale, Long-Tailed Dataset For Anime Character Recognition. *arXiv:2101.08674*, 2021.

J. Rissanen. Modeling by shortest data description. *Automatica*, 14(1):465–471, 1978.

M. Ronen, R. Rosenberg, B. I. Shraiman, and U. Alon. Assigning numbers to the arrows: Parameterizing a gene regulation network by using accurate expression kinetics. *PNAS*, 99, 2002.

M. R. Rosenzweig and K. I. Wolpin. Natural "Natural Experiments" in Economics. *J. Econ. Lit.*, 38(4):827–874, 2000.

K. Roth, M. Ibrahim, Z. Akata, P. Vincent, and D. Bouchacourt. Disentanglement of Correlated Factors via Hausdorff Factorized Support. *arXiv:2210.07347*, 2022.

D. Rothenhäusler, P. Bühlmann, and N. Meinshausen. Causal Dantzig: Fast inference in linear structural equation models with hidden variables under additive interventions. *Ann. Stat.*, 47(3):1688–1722, 2019.

D. B. Rubin. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, 66:688–701, 1974.

W. Rudin. *Principles of Mathematical Analysis*. McGraw Hill Higher Education, 1953.

J. Runge. Modern causal inference approaches to investigate biodiversity-ecosystem functioning relationships. *Nat. Commun.*, 14(1917):1–3, 2023.

M. Rutter. Proceeding From Observed Correlation to Causal Inference: The Use of Natural Experiments. *Perspectives on Psychological Science*, 2(4): 377–395, 2007.

K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan. Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data. *Science*, 308(5721):523–529, 2005.

K. Sakurada and T. Ishikawa. Synthesis of causal and surrogate models by non-equilibrium thermodynamics in biological systems. *Sci. Rep.*, 14(1001): 1–14, 2024.

S. Salehkaleybar, A. Ghassami, N. Kiyavash, and K. Zhang. Learning Linear Non-Gaussian Causal Models in the Presence of Latent Variables. *JMLR*, 21(39):1–24, 2020.

J. Salvatier, T. V. Wiecki, and C. Fonnesbeck. Probabilistic programming in Python using PyMC3. *PeerJ Comp Sci*, 2, 2016.

N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. Paritosh, and L. M. Aroyo. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *Conference on Human Factors in Computing Systems*, pages 1–15. ACM, 2021.

E. Sanderson, M. M. Glymour, M. V. Holmes, H. Kang, J. Morrison, M. R. Munafò, T. Palmer, C. M. Schooling, C. Wallace, Q. Zhao, and G. Davey Smith. Mendelian randomization. *Nat. Rev. Methods Primers*, 2(6):1–21, 2022.

P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich. SuperGlue: Learning Feature Matching With Graph Neural Networks. In *CVPR*, pages 4938–4947. IEEE, 2020.

M. Scanagatta, C. P. de Campos, G. Corani, and M. Zaffalon. Learning Bayesian Networks with Thousands of Variables. In *NeurIPS*, pages 1864–1872, 2015.

R. Scheines. An introduction to causal inference. 1997.

T. Schoeler, D. Speed, E. Porcu, N. Pirastu, J.-B. Pingault, and Z. Kutalik. Participation bias in the UK Biobank distorts genetic associations and downstream analyses. *Nat. Hum. Behav.*, 7:1216–1227, 2023.

P. Schoenegger, I. Tuminauskaite, P. S. Park, and P. E. Tetlock. Wisdom of the Silicon Crowd: LLM Ensemble Prediction Capabilities Rival Human Crowd Accuracy. *arXiv:2402.19379*, 2024.

B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. Toward Causal Representation Learning. *Proc. IEEE*, 109(5): 612–634, 2021.

G. Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, pages 461–464, 1978.

M. H. Secrest, R. W. Platt, C. R. Dormuth, D. Chateau, L. Targownik, R. Nie, C. M. Doyle, S. Dell'Aniello, and K. B. Filion. Extreme restriction design as a method for reducing confounding by indication in pharmacoepidemiologic research. *Pharmacoepidemiology and Drug Safety*, 29(S1):26–34, 2020.

J. Seng, M. Zečević, D. S. Dhami, and K. Kersting. Tearing Apart NOTEARS: Controlling the Graph Prediction via Variance Manipulation. *arXiv:2206.07195*, 2022.

U. Shalit and G. Chechik. Coordinate-descent for learning orthogonal matrices through Givens rotations. In *ICML*, pages 548–556. PMLR, 2014.

C. R. Shalizi and J. P. Crutchfield. Computational Mechanics: Pattern and Prediction, Structure and Simplicity. *J. Stat. Phys.*, 104(3):817–879, 2001.

S. Shapiro. Looking to the 21st century: Have we learned from our mistakes, or are we doomed to compound them? *Pharmacoepidemiology and Drug Safety*, 13(4):257–265, 2004.

A. Sharma and E. Kiciman. DoWhy: An End-to-End Library for Causal Inference. *arXiv:2011.04216*, 2020.

X. Shen, P. Bühlmann, and A. Taeb. Causality-oriented Robustness: Exploiting General Additive Interventions. *arXiv:2307.10299*, 2023.

S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen. A Linear Non-Gaussian Acyclic Model for Causal Discovery. *JMLR*, 7:2003–2030, 2006.

S. Shimizu, T. Inazumi, Y. Sogawa, A. Hyvärinen, Y. Kawahara, T. Washio, P. O. Hoyer, and K. Bollen. DirectLiNGAM: A Direct Method for Learning a Linear Non-Gaussian Structural Equation Model. *JMLR*, 12:1225–1248, 2011.

T. Shimizu. Diffusion Model in Causal Inference with Unmeasured Confounders. *arXiv:2308.03669*, 2023.

I. Shpitser, R. J. Evans, T. S. Richardson, and J. M. Robins. Introduction to Nested Markov Models. *Behaviormetrika*, 41(1):3–39, 2014.

I. Shpitser, R. J. Evans, and T. S. Richardson. Acyclic Linear SEMs Obey the Nested Markov Property. In *UAI*, volume 2018. PMLR, 2018.

V. Shwartz and Y. Choi. Do Neural Language Models Overcome Reporting Bias? In *International Conference on Computational Linguistics*, pages 6863–6870, 2020.

R. Silva, R. Scheines, C. Glymour, P. Spirtes, and D. M. Chickering. Learning the Structure of Linear Latent Variable Models. *JMLR*, 7(2), 2006.

V. P. Skitovitch. On a property of the normal distribution. *DAN SSSR*, 89: 217–219, 1953.

M. Sklar. Fonctions de répartition à N dimensions et leurs marges. *Publ. inst. statist. univ. Paris*, 8:229–231, 1959.

J. Sliwa, S. Ghosh, V. Stimper, L. Gresele, and B. Schölkopf. Probing the Robustness of Independent Mechanism Analysis for Representation Learning. In *UAI CRL*, 2022.

A. Soleymani, A. Raj, S. Bauer, B. Schölkopf, and M. Besserve. Causal Feature Selection via Orthogonal Search. *TMLR*, 2020.

R. J. Solomonoff. A formal theory of inductive inference. Part I. *Information and Control*, 7(1):1–22, 1964a.

R. J. Solomonoff. A formal theory of inductive inference. Part II. *Information and Control*, 7(2):224–254, 1964b.

L. Solus, Y. Wang, and C. Uhler. Consistency guarantees for greedy permutation-based causal inference algorithms. *Biometrika*, 108(4):795–814, 2021.

P. Spirtes, C. Meek, and T. Richardson. Causal inference in the presence of latent variables and selection bias. In *UAI*, pages 499–506. PMLR, 1995.

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

C. Squires, A. Seigal, S. Bhate, and C. Uhler. Linear Causal Disentanglement via Interventions. *arXiv:2211.16467*, 2022a.

C. Squires, D. Shen, A. Agarwal, D. Shah, and C. Uhler. Causal Imputation via Synthetic Interventions. In *Conference on Causal Learning and Reasoning*, pages 688–711. PMLR, 2022b.

A. Statnikov, S. Ma, M. Henaff, N. Lytkin, E. Efstathiadis, E. R. Peskin, and C. F. Aliferis. Ultra-Scalable and Efficient Methods for Hybrid Observational and Experimental Local Causal Pathway Discovery. *JMLR*, 16:3219–3267, 2015.

M. Strathern. 'Improving ratings': audit in the British University system. *European Review*, 5(3):305–321, 1997.

S. H. Strogatz. *Nonlinear Dynamics and Chaos with Student Solutions Manual*. CRC press, 2018.

T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, I. William M. Mauck, Y. Hao, M. Stoeckius, P. Smibert, and R. Satija. Comprehensive Integration of Single-Cell Data. *Cell*, 177(7):1888–1902, 2019.

N. Sturma, C. Squires, M. Drton, and C. Uhler. Unpaired Multi-Domain Causal Representation Learning. *NeurIPS*, 36:34465–34492, 2023.

M. Sugiyama, M. Krauledat, and K.-R. Müller. Covariate Shift Adaptation by Importance Weighted Cross Validation. *JMLR*, 8(5), 2007.

A. Taleb and C. Jutten. Source separation in post-nonlinear mixtures. *Transactions on Signal Processing*, 47(10):2807–2820, 1999.

T. Tanaka. A Theory of Mean Field Approximation. *NeurIPS*, 11, 1998.

G. Tennenholtz, A. Hallak, G. Dalal, S. Mannor, G. Chechik, and U. Shalit. On Covariate Shift of Latent Confounders in Imitation and Reinforcement Learning. In *ICLR*, 2021.

Y. Tenzer and G. Elidan. Generalized Ideal Parent (GIP): Discovering Non-Gaussian Hidden Variables. In *AISTATS*, pages 222–230. PMLR, 2016.

L. L. Thurstone. The Vectors of Mind. *Psychological Review*, 41(1):1, 1934.

M. E. Tipping and C. M. Bishop. Probabilistic Principal Component Analysis. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 61:611–622, 1999.

J. Ton, D. Sejdinovic, and K. Fukumizu. Meta Learning for Causal Direction. *AAAI*, 35(11):9897–9905, 2021.

N. Trask, C. Martinez, K. Lee, and B. Boyce. Unsupervised physics-informed disentanglement of multimodal data for high-throughput scientific discovery. *arXiv:2202.03242*, 2022.

F. Träuble, E. Creager, N. Kilbertus, F. Locatello, A. Dittadi, A. Goyal, B. Schölkopf, and S. Bauer. On Disentangled Representations Learned from Correlated Data. In *ICML*, pages 10401–10412. PMLR, 2021.

R. Tu, C. Zhang, P. Ackermann, K. Mohan, H. Kjellström, and K. Zhang. Causal Discovery in the Presence of Missing Data. In *AISTATS*, pages 1762–1770. PMLR, 2019.

J. R. Turner and R. M. Baker. Complexity Theory: An Overview with Potential Applications for the Social Sciences. *Systems*, 7(1):4, 2019.

R. E. Ulanowicz. *A Third Window. Natural Life beyond Newton and Darwin.* Templeton Foundation Press, 2009.

S. van de Geer and P. Bühlmann. ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs. *The Annals of Statistics*, 41(2):536–567, 2013.

T. F. A. van der Ouderaa and M. van der Wilk. Learning invariant weights in neural networks. In *UAI*, pages 1992–2001. PMLR, 2022.

M. van der Wilk, M. Bauer, S. John, and J. Hensman. Learning Invariances using the Marginal Likelihood. *NeurIPS*, 31, 2018.

R. Varley. ExoData: A Python package to handle large exoplanet catalogue data. *Computer Physics Communications*, 207:298–309, 2016.

S. Vattikuti, J. J. Lee, C. C. Chang, S. D. H. Hsu, and C. C. Chow. Applying compressed sensing to genome-wide association studies. *GigaScience*, 3(1): 2047–217, 2014.

P. Versteeg, C. Zhang, and J. M. Mooij. Local Constraint-Based Causal Discovery under Selection Bias. *arXiv:2203.01848*, 2022.

N. X. Vinh, J. Epps, and J. Bailey. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *ICML*, pages 1073–1080. PMLR, 2009.

N. X. Vinh, J. Epps, and J. Bailey. Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. *JMLR*, 11(95):2837–2854, 2010.

H. Vollmer. *Introduction to Circuit Complexity: A Uniform Approach*. Springer Science & Business Media, 1999.

J. von Kügelgen, M. Besserve, L. Wendong, L. Gresele, A. Kekić, E. Bareinboim, D. M. Blei, and B. Schölkopf. Nonparametric Identifiability of Causal Representations from Unknown Interventions. *NeurIPS*, 36:48603–48638, 2023.

M. J. Wainwright, M. I. Jordan, et al. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355. Association for Computational Linguistics, 2018.

H. Wang and J. K. Kim. Information projection approach to propensity score estimation for handling selection bias under missing at random. *arXiv:2104.13469*, 2021.

X. Wang, H. Chen, S. Tang, Z. Wu, and W. Zhu. Disentangled Representation Learning. *arXiv:2211.11695*, 2022.

Y. Wang and D. M. Blei. The Blessings of Multiple Causes. *J. Am. Stat. Assoc.*, 2019.

Y. Wang and M. I. Jordan. Desiderata for Representation Learning: A Causal Perspective. *arXiv:2109.03795*, 2021.

Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10(1):57, 2009.

L. Wasserman. *All of Statistics: A Concise Course in Statistical Inference*, volume 26. Springer, 2004.

S. Watanabe. Information Theoretical Analysis of Multivariate Correlation. *IBM Journal of Research and Development*, 4:66–82, 1960.

D. Waxman, K. Butler, and P. M. Djurić. Dagma-DCE: Interpretable, Non-Parametric Differentiable Causal Discovery. *IEEE Open J. Signal Process.*, 5:393–401, 2024.

D. Wei, T. Gao, and Y. Yu. DAGs with No Fears: A Closer Look at Continuous Optimization for Learning Bayesian Networks. *NeurIPS*, 33:3895–3906, 2020.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. V. Le, and D. Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *NeurIPS*, 35:24824–24837, 2022.

D. S. Weisberg and A. Gopnik. Pretense, Counterfactuals, and Bayesian Causal Models: Why What Is Not Real Really Matters. *Cogn. Sci.*, 37(7):1368–1381, 2013.

Z. Wen and W. Yin. A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, 142(1):397–434, 2013.

B. M. Wiernik and J. A. Dahlke. Obtaining Unbiased Results in Meta-Analysis: The Importance of Correcting for Statistical Artifacts. *Advances in Methods and Practices in Psychological Science*, 3(1):94–123, 2020.

E. O. Wilson. *Consilience: The Unity of Knowledge*, volume 31. Vintage, 1999.

D. E. Worrall, S. J. Garbin, D. Turmukhambetov, and G. J. Brostow. Harmonic networks: Deep translation and rotation equivariance. In *CVPR*, pages 5028–5037. IEEE, 2017.

S. Wright. Correlation and Causation. *Journal of Agricultural Research*, 20(7): 557, 1921.

G. Wunsch. Confounding and control. *Demographic research*, 16:97–120, 2007.

A. C. Wysocki, K. M. Lawson, and M. Rhemtulla. Statistical Control Requires Causal Justification. *Advances in Methods and Practices in Psychological Science*, 5(2), 2022.

F. Xie, R. Cai, B. Huang, C. Glymour, Z. Hao, and K. Zhang. Generalized Independent Noise Condition for Estimating Latent Variable Causal Graphs. *NeurIPS*, 33:14891–14902, 2020.

F. Xie, B. Huang, Z. Chen, Y. He, Z. Geng, and K. Zhang. Identification of Linear Non-Gaussian Latent Hierarchical Structure. In *ICML*, pages 24370–24387. PMLR, 2022.

J. Yang, R. Walters, N. Dehmamy, and R. Yu. Generative Adversarial Symmetry Discovery. In *ICML*, pages 39488–39508. PMLR, 2023.

Y. Yang, A. Ghassami, M. Nafea, N. Kiyavash, K. Zhang, and I. Shpitser. Causal Discovery in Linear Latent Variable Models Subject to Measurement Error. *NeurIPS*, 35:874–886, 2022.

T. Yarkoni. The generalizability crisis. *Behav. Brain Sci.*, 45, 2022.

M. Yasunaga, X. Chen, Y. Li, P. Pasupat, J. Leskovec, P. Liang, E. H. Chi, and D. Zhou. Large Language Models as Analogical Reasoners. In *ICLR*, 2023.

I. Yildirim and L. A. Paul. From task structures to world models: what do LLMs know? *Trends in Cognitive Sciences*, 28(5):404–415, 2024.

T. Yu, P. Li, B. Chen, A. Yuan, and J. Qin. Maximum pairwise-rank-likelihood-based inference for the semiparametric transformation model. *J. Econometrics*, 235(2):454–469, 2023.

Y. Yu, J. Chen, T. Gao, and M. Yu. DAG-GNN: DAG Structure Learning with Graph Neural Networks. In *ICML*, pages 7154–7163. PMLR, 2019.

R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning Fair Representations. In *ICML*, pages 325–333. PMLR, 2013.

A. Zhang, F. Liu, W. Ma, Z. Cai, X. Wang, and T.-s. Chua. Boosting Differentiable Causal Discovery via Adaptive Sample Reweighting. In *ICLR*, 2023a.

J. Zhang. On the Completeness of Orientation Rules for Causal Discovery in the Presence of Latent Confounders and Selection Bias. *Artificial Intelligence*, 172:1873–1896, 2008.

J. Zhang, K. Greenewald, C. Squires, A. Srivastava, K. Shanmugam, and C. Uhler. Identifiability Guarantees for Causal Disentanglement from Soft Interventions. *NeurIPS*, 36:50254–50292, 2023b.

K. Zhang and A. Hyvärinen. On the identifiability of the post-nonlinear causal model. In *UAI*, pages 647–655. AUAI Press, 2009.

K. Zhang and A. Hyvärinen. Distinguishing causes from effects using nonlinear acyclic causal models. In *Causality: Objectives and Assessment*, pages 157–164. PMLR, 2010.

K. Zhang, J. Peters, D. Janzing, and B. Schölkopf. Kernel-Based Conditional Independence Test and Application in Causal Discovery. In *UAI*, page 804–813. AUAI Press, 2011.

K. Zhang, J. Zhang, B. Huang, B. Schölkopf, and C. Glymour. On the Identifiability and Estimation of Functional Causal Models in the Presence of Outcome-Dependent Selection. In *UAI*. AUAI Press, 2016.

K. Zhang, B. Huang, J. Zhang, C. Glymour, and B. Schölkopf. Causal Discovery from Nonstationary/Heterogeneous Data: Skeleton Estimation and Orientation Determination. *International Joint Conference on Artificial Intelligence*, pages 1347–1353, 2017.

X.-H. Zhang, L. Y. Tee, X.-G. Wang, Q.-S. Huang, and S.-H. Yang. Off-target Effects in CRISPR/Cas9-mediated Genome Engineering. *Molecular Therapy - Nucleic Acids*, 4, 2015.

Y. Zhang, Y. Zhang, Y. Gan, L. Yao, and C. Wang. Causal Graph Discovery with Retrieval-Augmented Generation based Large Language Models. *arXiv:2402.15301*, 2024.

X. Zheng, B. Aragam, P. K. Ravikumar, and E. P. Xing. DAGs with NO TEARS: Continuous Optimization for Structure Learning. *NeurIPS*, 31, 2018.

X. Zheng, C. Dan, B. Aragam, P. Ravikumar, and E. P. Xing. Learning Sparse Nonparametric DAGs. In *AISTATS*, pages 3414–3425. PMLR, 2020.

Y. Zheng, I. Ng, and K. Zhang. On the Identifiability of Nonlinear ICA: Sparsity and Beyond. *NeurIPS*, 35:16411–16422, 2022.

Y. Zheng, B. Huang, W. Chen, J. Ramsey, M. Gong, R. Cai, S. Shimizu, P. Spirtes, and K. Zhang. Causal-learn: Causal Discovery in Python. *JMLR*, 25(60):1–8, 2024.

S. Zhu, I. Ng, and Z. Chen. Causal Discovery with Reinforcement Learning. In *ICLR*, 2019.

J. Ziv and A. Lempel. Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory*, 24(5):530–536, 1978.

Index