View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Online Publikationen der Universität Stuttgart

Causal models for decision making via integrative inference

Von der Fakultät 5: Informatik, Elektrotechnik und Informationstechnik der Universität Stuttgart zur Erlangung der Würde eines Doktors der Naturwissenschaften (Dr. rer. nat.) genehmigte Abhandlung

Vorgelegt von
Philipp Geiger
aus Weingarten

Hauptberichter: Prof. Dr. Marc Toussaint
Mitberichter: Prof. Dr. Bernhard Schölkopf

Tag der mündlichen Prüfung: 20.06.2017

Institut für Parallele und Verteilte Systeme der Universität Stuttgart
2017

Acknowledgments

First I would very much like to thank my three supervisors Dominik Janzing, Bernhard Schölkopf and Marc Toussaint for accepting me as a PhD student and for supporting me. I learned a lot about causality, but also about mathematics, machine learning, physics and decision making from them. I am particularly grateful for the freedom that they gave me.

Also I would like to thank my current and former colleagues for the good times I had with them, for helping and supporting me, and for all the things I learned from them, which was not less than what I learned from my supervisors.

Generally I had a great time living in the beautiful town of Tübingen and conducting research at the Empirical Inference Department of the Max Planck Institute for Intelligent Systems.

Special thanks also go to my collaborators, in particular Kun Zhang.

Last but not least I would like to thank my family and especially my friends.

Contents

Index of notation

Notation	Meaning (page)
$[v]_k$	k-th component of the vector v
$X \sim Y$	X and Y have the same distribution
CBN	causal Bayesian network, 17
$\mathrm{D}(p\|q)$	Kullback-Leibler divergence between p and q, 16
DAG	directed acyclic graph, 16
FCM	functional causal model, 17
GCM	graphical causal model, 17
$\mathbf{I}$	identity matrix
$\mathrm{H}(X\|Y)$	conditional Shannon entropy, 16
i.i.d.	independent and identically distributed
$\mathrm{I}(X:Y\|Z)$	conditional mutual information of X and Y given Z, 16
PA_X^G	parents of X in the graph G (superscript G is dropped if the graph is clear), 16
PCM	probabilistic causal model, 17
SEM	structural equation model, 17
$X_{1:L}$	$X_1, \ldots, X_L$
VAR	vector autoregressive, 61

Abstract

Understanding causes and effects is important in many parts of life, especially when decisions have to be made. The systematic inference of causal models remains a challenge though. In this thesis, we study (1) "approximative" and "integrative" inference of causal models and (2) causal models as a basis for decision making in complex systems. By "integrative" here we mean including and combining settings and knowledge beyond the outcome of perfect randomization or pure observation for causal inference, while "approximative" means that the causal model is only constrained but not uniquely identified. As a basis for the study of topics (1) and (2), which are closely related, we first introduce causal models, discuss the meaning of causation and embed the notion of causation into a broader context of other fundamental concepts.

Then we begin our main investigation with a focus on topic (1): we consider the problem of causal inference from a non-experimental multivariate time series $(X_t)_{t\in\mathbb{Z}}$, that is, we integrate temporal knowledge. We take the following approach: We assume that $(X_t)_{t\in\mathbb{Z}}$ together with some potential hidden common cause – "confounder" – $(Z_t)_{t\in\mathbb{Z}}$ forms a first order vector autoregressive (VAR) process with structural transition matrix A. Then we examine under which conditions the most important parts of A are identifiable or approximately identifiable from only $(X_t)_{t\in\mathbb{Z}}$, in spite of the effects of $(Z_t)_{t\in\mathbb{Z}}$. Essentially, sufficient conditions are (a) non-Gaussian, independent noise or (b) no influence from $(X_t)_{t\in\mathbb{Z}}$ to $(Z_t)_{t\in\mathbb{Z}}$. We present two estimation algorithms that are tailored towards conditions (a) and (b), respectively, and evaluate them on synthetic and real-world data. We discuss how to check the model using $(X_t)_{t\in\mathbb{Z}}$.

Still focusing on topic (1) but already including elements of topic (2), we consider the problem of approximate inference of the causal effect of a variable X on a variable Y in i.i.d. settings "between" randomized experiments and observational studies. Our approach is to first derive approximations (upper/lower bounds) on the causal effect,

in dependence on bounds on (hidden) confounding. Then we discuss several scenarios where knowledge or beliefs can be integrated that in fact imply bounds on confounding. One example is about decision making in advertisement, where knowledge on partial compliance with guidelines can be integrated.

Then, concentrating on topic (2), we study decision making problems that arise in cloud computing, a computing paradigm and business model that involves complex technical and economical systems and interactions. More specifically, we consider the following two problems: debugging and control of computing systems with the help of sandbox experiments, and prediction of the cost of "spot" resources for decision making of cloud clients. We first establish two theoretical results on approximate counterfactuals and approximate integration of causal knowledge, which we then apply to the two problems in toy scenarios.

Kurzzusammenfassung

Ursachen und Wirkungen zu verstehen ist von großer Bedeutung in vielen Bereichen des Lebens, insbesondere, wenn Entscheidungen gefällt werden müssen. Die systematische Inferenz von kausalen Modellen bleibt jedoch eine Herausforderung. In dieser Dissertation werden (1) "approximative" und "integrative" Inferenz kausaler Modelle und (2) kausale Modelle als Grundlage für Entscheidungsfindung untersucht. Mit "integrativ" ist hier gemeint, dass Szenarien und Wissen jenseits von perfekt randomisierten Experimenten und reinen Observationsstudien mit einbezogen und kombinierten werden, während sich "approximativ" darauf bezieht, dass das wahre kausale Modell eingegrenzt, aber nicht eindeutig identifiziert wird. Als Grundlage für die darauffolgenden Abhandlungen zu den genannten Themen (1) und (2), die eng miteinander zusammenhängen, werden zunächst kausale Modelle eingeführt, die Bedeutung des Begriffs der Kausalität wird diskutiert, und der Begriff der Kausalität wird in einen breiteren Kontext von anderen grundlegenden Begriffen eingebettet.

Dann beginnt die Hauptuntersuchung mit einem Schwerpunkt auf Thema (1): es wird das Problem der kausalen Inferenz von einer nicht-experimentellen, multivariaten Zeitreihe $(X_t)_{t\in\mathbb{Z}}$ betrachtet, d.h. es wird zeitliches Wissen integriert. Dabei wird der folgende Ansatz verfolgt: Es wird angenommen, dass $(X_t)_{t\in\mathbb{Z}}$ zusammen mit einer potentiellen versteckten gemeinsamen Ursache – kurz "Confounder" – $(Z_t)_{t\in\mathbb{Z}}$ einen vektorautoregressiven Prozess erster Ordnung mit struktureller Übergangsmatrix A bildet. Dann wird untersucht, unter welchen Bedingungen die wichtigsten Teile von A identifizierbar oder approximativ identifizierbar sind auf Grundlage von $(X_t)_{t\in\mathbb{Z}}$, trotz der Einflüsse von $(Z_t)_{t\in\mathbb{Z}}$. Im Wesentlich sind die Folgenden Bedingungen hinreichend: (a) nicht-normalverteiltes, unabhängiges Rauschen oder (b) kein Einfluss von $(X_t)_{t\in\mathbb{Z}}$ nach $(Z_t)_{t\in\mathbb{Z}}$. Es werden zwei Schätzalgorithmen vorgestellt, die auf Bedingung (a) bzw. (b)

zugeschnitten sind, und auf künstlichen und Realdaten evaluiert. Des Weiteren wird diskutiert, wie $(X_t)_{t\in\mathbb{Z}}$ genutzt werden kann um die Modellannahmen zu überprüfen.

Als nächster Schritt im Rahmen von Thema (1), jedoch auch schon Elemente von Thema (2) beinhaltend, wird das Problem der approximativen Inferenz des Effekts einer Variable X auf eine variable Y in I.i.d.-Szenarien "zwischen" randomisierten Experimenten und Observationsstudien erforscht. Unser Ansatz besteht darin, zuerst Approximationen (untere/obere Schranken) bzgl. des kausalen Effekts, in Abhängigkeit von als gegeben angenommenen Schranken bzgl. verstecktem Confounding, herzuleiten. Daraufhin werden verschiedene Szenarien diskutiert, in denen Wissen oder Vermutungen integriert werden können, die Schranken in Bezug auf verstecktes Confounding implizieren. Ein Beispiel behandelt Entscheidungsfindung im Bereich der Werbung, wo Wissen bzgl. der partiellen Einhaltung von Vorschriften integriert werden kann.

Dann wird der Schwerpunkt auf Thema (2) gelegt, beginnend mit einer Untersuchung von Entscheidungsproblemen die im Bereich des Cloud-Computing auftreten, einem Computing-Paradigma und Geschäftsmodell, das komplexe technische und ökonomische Systeme und Interaktionen beinhaltet. Genauer geht es um die folgenden zwei Probleme: Debugging und Regelung von Computersystemen mithilfe von "Sandbox"-Experimenten einerseits, und Vorhersage der Kosten sogenannter "Spot"-Ressourcen für die Entscheidungsfindung von Cloud-Kunden andererseits. Wir beweisen zuerst zwei theoretische Resultate zu approximativen kontrafaktischen Wahrscheinlichkeiten und zur approximativen Integration von kausalem Wissen, die wir dann in Spielzeugszenarien auf die zwei genannten Probleme anwenden.

Chapter 1.

Introduction

1.1. Motivation

Many questions that arise in life, especially in the course of decision making, are about causal relations. One may wonder about the cause of the abdominal pain one feels at some point in time, and how a certain drug or a different diet will affect it; a manufacturer may try to find out the factors that drive the demand for her product in general, or try to infer the influence of a specific factor, say advertisement, to inform her decision making; a policy maker may wonder about the influence of state debt on future economic growth, or about reasons for the rise of nationalistic movements.

Causal questions – some similar, some different from the examples just given – have certainly played a role for a long time in human history [Falcon, 2015, Hulswit, 2004]. It is a rather new development though, that such questions and ways to answer them are systematically studied from a mathematical and algorithmic perspective [Granger, 1969, Imbens and Rubin, 2015, Shadish et al., 2002, Pearl, 2000, Spirtes et al., 2000, Shimizu et al., 2006, Mooij et al., 2016, Peters et al., 2017]. One motivation for this development is that formalization can help to clarify concepts, arguments and communication. Another reason for this development are the economical and technological trends of automation and digitalization, which prompt various issues in terms of design and mathematical analysis of algorithms for causal inference.

This thesis lines up in the mentioned mathematical and algorithmic work on causation. It makes heavily use of causal models as introduced by Pearl [2000], Spirtes et al. [2000] to make further steps towards answering relevant causal questions, in particular those that arise in the course of decision making. The thesis is especially driven by the following issues:

- Randomized experiments are the gold standard for causal inference, but often they are expensive, unethical or impossible to perform. On the other hand, plenty of "cheap" observational (i.e., non-experimental) data is available. Approaches, often based on causal models, have been developed to more heavily integrate observational data into causal inference [Pearl, 2000, Spirtes et al., 2000, Shimizu et al., 2006, Mooij et al., 2016, Peters et al., 2017]. A limitation of these approaches is that they often either need strong assumptions, or they only draw weak conclusions.

  Is it possible to *integrate further forms* of "cheap" knowledge (beyond observations) as well as alternative forms of experimentation for causal inference? To what extent can temporal information [Granger, 1969, Schreiber, 2000b, Eichler, 2012], "imperfect" experimentation [Thistlewaite and Campbell, 1960, Shadish et al., 2002], or say system specifications (in case inference is w.r.t. engineered systems) help for causal inference, beyond established results? To what extent can causal models help for the formal side of such integration? In which cases can *approximate* but still meaningful results be established, which are often more realistic than unique identification of a causal model?

- Decisions concerning natural, social and technical systems of high complexity have to be made by humans, to stir them towards predefined goals. Furthermore, complicated "decisions" also have to be made by controllers and, more generally, intelligent machines. Ideally, decision making is performed on the basis of an understanding of the effects of executing a decision, in particular when decisions are about specific manipulations of the system (although, clearly, effects are not the only criterion to judge a decision). How can causal models help here? In particular, can they help when the available information is heterogeneous?

- Being a concept used so frequently in everyday life, it is surprising how much the meaning of causation is still subject to debate. And while causal models as intro-

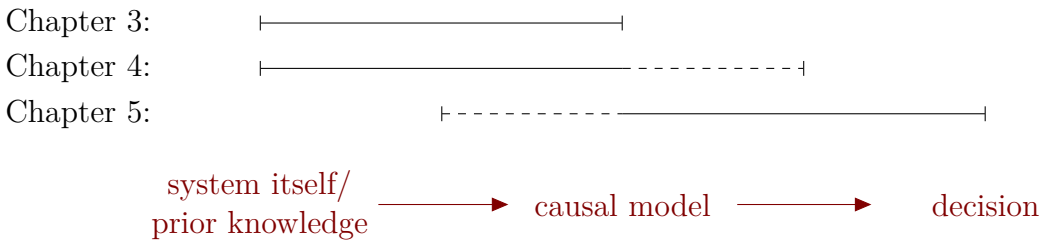


Figure 1.1.: Illustration of the main part of this thesis. We study fragments of the "inference path" that starts at a given system and information about it and goes via causal models towards the goal of an informed decision (concerning the system), depicted in red along the x-axis. The focus of the respective chapter is depicted by a solid black line, while topics that are briefly touched are depicted by a shorter dashed line.

duced by Pearl [2000], Spirtes et al. [2000] clarify some aspects of causal reasoning, they also "mask" some issues, for instance about the notion of an intervention. Can the meaning of causation be further clarified by better understanding its relation to other fundamental concepts, such as time? Can such a clarification help to advance causal inference methods?

1.2. Structure

The thesis is structured as follows:

- Chapter 2 contains prerequisites, and a summary of the subsequent chapters as well as the main contributions.
- Then, Chapters 3 to 5 contain the main part of this thesis: a study of approximative and integrative inference of causal models, and causal models as a basis for decision making in complex systems. We schematically illustrate the structure of these chapters in Figure 1.1.
- We conclude with Chapter 6, by weighing accomplishments and limitations of causal models in general, and this thesis in particular.

Chapter 2.

Preliminaries and overview

Here we first provide the conceptual background for this thesis, in Section 2.1, and then summarize content and contributions of the main part of this thesis in Section 2.2.

2.1. Preliminaries

We start, in Section 2.1.1, by introducing the rigorous mathematical causal modeling language the whole thesis is based on, followed by an informal discussion of what we mean by causation in Section 2.1.2. Afterwards, Sections 2.1.3 and 2.1.4 contain some background regarding the two main topics we will use causal models for: learning of causal models, and causal models as a basis for decision making. Last, in Section 2.1.5, we embed the concept of causation into a context of other important concepts. Generally, a significant part of the current chapter is devoted to painting a bigger picture, which may sometimes be vague, before we go into rigorous mathematical details in the main part of this paper, staring with Chapter 3.[1]

2.1.1. Mathematical concepts for causal modeling

We assume familiarity with basic probability theory as described, e.g., by Klenke [2013]. We usually presuppose some underlying probability space w.r.t. which random variables

[1] Thoughts in Sections 2.1.2 and 2.1.5 are – besides the mentioned references – based on personal communication with Bernhard Schölkopf and Dominik Janzing.

are defined, without necessarily mentioning it explicitly. We consider random variables with discrete as well as continuous domains. We usually denote the distribution of a random variable X by P_X or $P(X)$, and the conditional distribution of X given Y by $P_{X|Y}$ or $P(X|Y)$. By a (probability) density we either refer to a density w.r.t. the Lebesgue measure, in the continuous case, or w.r.t. the counting measure, in the discrete case, respectively. Usually, we write the density of a random variable X as $p_X(x)$, $p(x)$ or $p(X)$ and the conditional density of X given $Y = y$ is usually written as $p_{X|Y}(x|y)$, $p_{X|Y=y}(x)$, $p(x|y)$ or sometimes $p(X|y)$. If V is a tuple or set of random variables, then we may denote their joint distribution and density by P_V and p_V, respectively.

We also assume the reader to be familiar with basic information theory as described by Cover and Thomas [1991]. By $\mathrm{H}(\cdot)$ ($\mathrm{H}(\cdot|\cdot)$) we denote the (conditional) Shannon entropy, by $\mathrm{I}(\cdot : \cdot)$ ($\mathrm{I}(\cdot : \cdot|\cdot)$) the (conditional) mutual information, and by $\mathrm{D}(\cdot\|\cdot)$ the Kullback-Leibler (KL) divergence, usually based on logarithms with base 2. Keep in mind that, regarding the KL divergence of conditional densities $p(x|y), q(x|y)$, we use the following notation:

$$\mathrm{D}(p(X|Y)\|q(X|Y)) := \sum_{x,y} p(x,y) \log \frac{p(x|y)}{q(x|y)},$$
$$\mathrm{D}(p(X|y)\|q(X|y)) := \sum_{x} p(x|y) \log \frac{p(x|y)}{q(x|y)},$$

and similarly for continuous X, Y.

We assume familiarity with basic concepts from graph theory and probabilistic graphical models as described, e.g., by Lauritzen [1996], Spirtes et al. [2000], Pearl [2000]. In particular, we will make use of the concepts of a directed acyclic graph (DAG), Markovianity, faithfulness, (directed) paths, blocking (of paths), d-separation, and skeleton of a DAG. By PA_X^G we denote the set of parents of a node X in the DAG G (superscript G is dropped if the graph is clear).

Now we define causal models mathematically following Pearl [2000], Spirtes et al. [2000]. We give two closely connected definitions and discuss their relationship in Remark 2.1. Both definitions will be used in this thesis – it depends on the context which definition is more helpful. Let V be a set of variables, and let $\mathrm{dom}(X)$ denote the domain of a variable X.

Definition 2.1 (Functional causal model)**.** *A* functional causal model (FCM)*, or* structural equation model (SEM)*, M over V consists of the following components:*

- *a* background variable U_X *for each* $X \in V$ *(we may also denote it by* N_X *and refer to it as* noise *or* exogenous variable*),*
- *a distribution on* $\Pi_{X \in V} \operatorname{dom}(U_X)$ *that is a product distribution, denoted by* P_U *and referred to as* background distribution *(i.e., a joint distribution on the background variables that makes them independent),*
- *a* structural equation
$$X := f_X(PA_X, U_X)$$
for each $X \in V$ *and some set of variables* $PA_X \subset V$ *called the* parents of X*, where* f_X *is called the* structural (equation) function for X*.*

We call the elements of V *the* (endogeneous) variables of M*.*

Definition 2.2 (Graphical causal model)**.** *A* graphical causal model (GCM)*, or causal graphical model (CGM), or* causal Bayesian network (CBN)*, M over V consists of the following components:*

- *a DAG* G *with* V *as node set, called* causal diagram *or* causal DAG*,*
- *a conditional probability density* $p_{X|PA_X=pa_X}$ *(defined for all* $pa_X \in \operatorname{dom}(PA_X)$*) for each* $X \in V$*.*[2]

Again we call the elements of V *the* (endogeneous) variables of M*.*

Definition 2.3 (Probabilistic causal model)**.** *We refer to FCMs and GCMs jointly as* probabilistic causal models (PCMs) *or* causal models *for short.*[3]

Remark 2.1 (Objects implied by PCMs)**.** *Let* V *denote the set of endogenous variables.*

[2] Restricting to cases where densities are defined is broad enough for this thesis, although a more general definition may be possible.

[3] One may also read "PCM" as "Pearl/Pittsburgh causal model", as Judea Pearl is probably the main contributor to their theory, while the other important contributors, Peter Spirtes, Clark Glymour and Richard Scheines all are or were faculty at Carnegie-Mellon University in Pittsburgh.

- *Generally, we consider a structural equation as a "stronger", asymmetric form of equation. In particular, a structural equation $A := B$ implies the classical equation $A = B$.*

- *An FCM with background variables $U_X, X \in V$, naturally induces an underlying probability space with outcome space $\Pi_{X \in V} \operatorname{dom}(U_X)$ and distribution P_U (the background variables U_X can be seen as random variables on that probability space – they are simply projections – which renders the symbol P_U for their joint distribution consistent with our probability theoretic notation introduced above, when defining $U = (U_X)_{X \in V}$). And for each $X \in V$, the structural equations of the FCM "turn" X into a random variable on that underlying probability space, in case all structural functions f_Y, $Y \in V$ are measurable.*

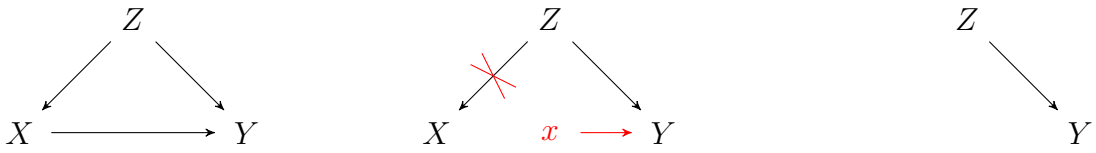
- *Similarly, GCMs naturally induce a joint density p_V over the variables in V, by multiplying the conditionals, and the variables $X \in V$ can then be seen as random variables. If p_V has support everywhere, then it, together with the causal DAG, already fully determines the GCM.*

- *For simplicity, we will usually treat the probability spaces and random variables induced by PCMs as part of the models themselves.*

- *The relation between FCMs and GCMs is as follows. Each FCM induces a unique GCM in a natural way: the parents in the structural equations define the parents in the causal diagram, and $p_{X|PA_X=pa_X}$ is defined as the density of $f_X(pa_X, U_X)$, for all variables X. It is easy to see though, that usually a given GCM is induced by many FCMs, so a GCM does not determine a (unique) FCM. Later, in Example 5.1, we give a specific example of a property of an FCM that is often not determined by a GCM.*

- *We will usually consider the causal diagram induced by an FCM as part of the FCM.*

- *If we want to make clear w.r.t. which PCM the distribution of a random variable X is meant, we may write $P^M(X)$ if we mean its distribution under M.*



(a) The joint density is $p_{Y|X,Z}(y|x,z)p_{X|Z}(x|z)p_Z(z)$.

(b) Remove $p_{X|Z}(x|z)$.

(c) The new joint density is $p_{Y|X,Z}(y|x,z)p_Z(z)$.

Figure 2.1.: Example for calculating the post-interventional density $p^M_{Y|\mathrm{do}\,X=x}$, given a causal model M (part (a)), based on deriving the post-interventional causal model $M_{\mathrm{do}\,X=x}$ (part (c)).

Now we introduce the formal concept of an intervention, which can be seen as an operator on causal models. This operator needs the causal structure. It is not determined from the joint probability distribution only.

Definition 2.4 (Post-interventional causal model and distribution)**.** *Given a causal model M and a tuple of variables X of M, the* post-interventional causal model $M_{\mathrm{do}\,X=x}$ *is defined as follows:*

- *if M is an FCM: drop the structural equations for all variables in X and replace, in all remaining structural equations, variables of X by the corresponding constant entries of x;*
- *if M is a GCM: drop the variables in X and all incoming arrows from the causal diagram, drop the conditional density $p_{X|PA_X=pa_X}$ from the model, and fix the value of variables in X to the corresponding entry of x in all remaining conditional densities.*

Based on this, we define the post-interventional distribution of Y after setting X to x in M*, denoted by $P^M_{Y|\mathrm{do}\,X=x}$ or $P^M(Y|\mathrm{do}\,X=x)$, by the distribution of Y in $M_{\mathrm{do}\,X=x}$ (we may drop the additional "in M", and the superscript M, if the underlying causal model is clear).*

If we explicitly want to refer to the variable Y as it is interpreted by $M_{\mathrm{do}\,X=x}$, we may write $Y_{M_{\mathrm{do}\,X=x}}$ or $Y_{\mathrm{do}\,X=x}$ instead of Y.

Let us give an example.

Example 2.1 (Calculating post-interventional model and distribution)**.** *Consider a GCM M with causal DAG as in Figure 2.1a (here X is an individual variable, compared to Definition 2.4 where we allowed it to be a tuple of variables). To calculate the post-interventional density $p^M_{Y|\mathrm{do}\,X=x}$, first perform the transformation of M depicted in Figure 2.1b, resulting in $M_{\mathrm{do}\,X=x}$ depicted in Figure 2.1c. Then, simply calculate $p^M_{Y|\mathrm{do}\,X=x}$ as*

$$p_Y^{M_{\mathrm{do}\,X=x}}(y) = \sum_z p_{Y|X,Z}(y|x,z)p_Z(z).$$

Post-interventional distributions will be essential to defining the semantics of causal models in Definition 2.6 below, by interpreting them as the predicted outcomes of randomized experiments. For a succinct terminology, we will consider the observational distribution P_Z, for any set of variables Z of a causal model M, as a special case of a post-interventional distribution.

A further remark regarding the relation between an FCM M and the post-interventional FCM $M_{\mathrm{do}\,X=x}$ is due here:

Remark 2.2. *Let M be an FCM and X, Y be individual variables or sets of variables in M. Following Pearl [2000, chapter 7.1.1 and chapter 7.2.2], we consider the random variables contained in M and the random variables in $M_{\mathrm{do}\,X=x}$, respectively, to be defined over the* same *underlying probability space (outcome space $\Pi_{X\in V}\,\mathrm{dom}(U_X)$ with distribution P_U). This allows expressions like $P(Y_{\mathrm{do}\,X=x}|X{=}x')$, where X is a random variable in M and $Y_{\mathrm{do}\,X=x}$ a random variable in $M_{\mathrm{do}\,X=x}$, to be meaningful. This quantity (the "counterfactual", see also Section 5.3.1) is sometimes written as $P(Y|\mathrm{do}\,X{=}x, X{=}x')$. Note that this quantity is not uniquely determined by a GCM only, it is necessary to know the "underlying" FCM (see also Example 5.1).*

Keep in mind the following statement, which links causal model and observational distribution, and justifies the causal Markov assumption which we will briefly discuss in Section 2.1.3.2.

Fact 2.1 (Causal model implies causal Markov assumption [Pearl, 2000])**.** *Let M be a causal model over variables $X_1, \ldots, X_n$ with causal DAG G. Then the joint distribution $X_1, \ldots, X_n$ induced by M is Markovian w.r.t. G.*

Also keep in mind the following terminology [Pearl, 2000, Spirtes et al., 2000, Peters et al., 2017].

Definition 2.5. *Let M be a causal model with variable set V and causal DAG G.*

- *We call a linear ordering over V a* causal ordering (relative to M)*, if it is a topological sorting of G, i.e., a linear ordering of the nodes of G such that there is no arrow form a "smaller" to a "larger" node.*
- *A variable $Z \in V$ is called a* confounder *or* common cause *of variables $X_1, \ldots, X_n \in V$ different from Z, if for all i there is a directed path from Z to X_i that is not blocked by any X_j with $j \neq i$. (For instance, in the causal DAG in Figure 2.1a on page 19, Z is a confounder of X and Y.)*
- *A set of variables $W \subset V$ is called* causally sufficient*, if all confounders $Z \in V$ of variables in W are already contained in W.*[4]
- *Relative to a given setting, we call the variables in V that are measured in that setting* observed variables / observables*, and those not measured* unobserved/hidden variables*. Usually we depict hidden variables by (dashed) gray circles, such as the Z in Figure 2.1a.*
- *If we are interested in inferring the causal effect of a variable $X \in V$ on a variable $Y \in V$, we may call X the* treatment variable*, Y the* outcome variable*, and $P(X|PA_X)$ the* assignment mechanism *[Imbens and Rubin, 2015].*
- *Let Q be any (joint) distribution over $\prod_{X \in V} \mathrm{dom}(X)$. We say that Q satisfies* causal minimality *w.r.t. the DAG G, if it is Markovian w.r.t. G but not w.r.t. any proper subgraph of G.*

Note that some further, more specific definitions will be given in the respective chapters where they are used.

[4] See Peters et al. [2017, Definition 9.1] for a refined definition of causal sufficiency called "interventional sufficiency".

2.1.2. Meaning of "causation"

While we already gave some examples in Section 1.1, here we try to explicitly define what we mean by the causal effect of a variable X on a variable Y. In particular, we restrict to a meaning of causal effect (or "influence") which can be formalized by a conditional density of Y given $X = x$ and denote it by $p^c_{Y|x}$ for the moment. We present two definitions: one that reduces causation (essentially) to interventions, which is somewhat unsatisfactory from the empirical point of view; and one which reduces causation to randomized experiments.

This section is necessarily more vague, and may contain statements more subject to debate, than the previous one. Roughly speaking, our goal here is to express the meaning of causation in terms that are clearer and "more empirical" than causation itself. In this sense, we will try to work out precisely which definition is relative to which other concepts, in particular to avoid circularity.

2.1.2.1. Relative to interventions and PCMs

One possibility is to define the causal effect of a variable X on a variable Y as the density of Y after *intervening* on X, setting it to a fixed value x. (We will give further details on what is meant by an intervention in Section 2.1.2.3.)

And there is a clear way to model interventions with PCMs: Given say a GCM M, it seems natural to translate the intervention on X into a transformation of M where we remove the conditional $p_{X|PA_X=pa_X}$ from the joint density, and the corresponding arrow from the causal DAG, and fix the value of X to x in all remaining conditional densities. The resulting density of Y exactly coincides with the post-interventional distribution $p^M_{Y|\mathrm{do}\,X=x}$ we introduced in Definition 2.4. This allows us to use PCMs to formally reason in various ways. In particular, given a PCM M, *relative to* M the meaning of the causal effect, for which we introduced the term $p^c_{Y|x}$ above, is simply given by $P^M_{Y|\mathrm{do}\,X=x}$.[5]

[5] If the predicted outcome of an intervention is wrong then either the specific model is poor or it was not an intervention – or PCMs do not capture well the notion of an intervention in general.

2.1.2.2. Relative to randomized experiments

While illustrating how the language of PCMs works, clearly we have to go beyond the above "definition" of causation: Usually, the causal model M does not just fall into one's lap, instead a definition has to be more based on the empirical world. Unfortunately in empirical settings, (point-)interventions are hard to define and perform. Therefore, we base our definition on randomized experiments, as is common practice [Imbens and Rubin, 2015].

Definition 2.6 (Causation and correct causal model)**.** *We define actual causal effect and correct causal model as follows.*

(a) The (actual/true) causal effect *of a variable X on a variable Y, $p^c_{Y|x}$, is defined as the conditional density of Y given X obtained from a randomized experiment, where X is randomized (and similarly when X, Y are* sets *of variables).*

(b) Given a set V of variables, we say that a causal model M over these variables is a correct/true causal model*, if*

- *the joint density p^M_V coincides with the observational density of the variables in V,*
- *the causal diagram of M is causally minimal[6] w.r.t. the observational density of the variables in V,*
- *for any two sets of variables $X, Y \subset V$, the post-interventional density $p^M_{Y|\mathrm{do}\, X=x}$ coincides with the actual causal effect $p^c_{Y|x}$ of X on Y as we defined it in part (a).*

2.1.2.3. Remarks

Clearly, the above Definition 2.6 is not perfect either and swipes under the rug various issues. Therefore, let us make some remarks.

[6] This is required because otherwise the causal diagram may contain "too many" arrows which do not correspond to actual effects.

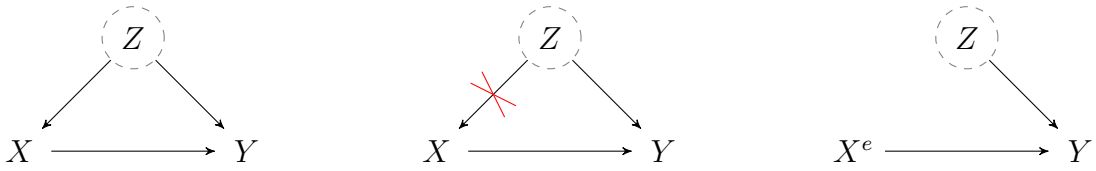
- **Why randomization and not intervention.** Consider some system with variables X, Y and assume we want to infer the effect of X on Y. If we could "perfectly" intervene on X, then no randomization would be necessary. One would simply intervene several times, setting X to different values (possibly rerun the system for the same value several times if it is stochastic). By a "perfect" intervention here we mean an intervention on the system similar to how we defined an intervention in a PCM: one destroys the mechanism that governs X, sets X to a specific value, but keeps the remaining mechanisms the system consists of invariant (on the population level, if the system is stochastic).

  However, in practice it can never be ruled out that the decisions on when and how to "intervene" are governed by some factor which also affects the subsequent Y, i.e., which does *not* leave the rest of the system invariant. For instance, it may happen that the experimenter would systematically increase the value of X over time, while the evolution of time also changes the system in a systematic way.

  That is, it is hard to make sure that what the experimenter does is a perfect intervention. This is why randomization is important, as it ensures that the value X is set to is independent of the (variation of the) rest of the system. It can still be argued that the idea underlying randomized experiments is based on some form of an intervention – a "soft" intervention – and we will comment on this below.

- **What is a valid randomized experiment?** We leave open what precisely constitutes a valid randomized experiment. Important concepts in this regard are "external validity" (whether it is appropriate to generalize from the experimental population[7]) and "internal validity" (whether the experimental setup ensures that the causal effect is correctly estimated w.r.t. the experimental population, in particular, that there is no hidden confounder) [Shadish et al., 2002, Imbens and Rubin, 2015]. These concepts are widely applied in the context of causality and

[7] We sometimes speak of a "population" as the object study, and sometimes of a "system". The former seems more suitable in case we are given different samples, say humans, from a population which can be defined based on some unifying trait. The latter seems more suitable for cases where in fact only "one (stochastic) individual" is given, say the global economy, but we observe this individual in various states – e.g., a finite trajectory of the system over time. While sometimes it may be important to distinguish between both, here we use the terms more or less interchangeably. Most of the time we reason on the level of probability distributions anyway, which can be done in both cases.



(a) Causal DAG for undisturbed, original population. Gray means hidden.

(b) Soft intervention – removing the original assignment mechanism ...

(c) ... and replacing it by randomization.

Figure 2.2.: PCMs and the notion of a soft intervention can help to argue why randomized experiments are a valid method of causal inference. The difference to Figure 2.1 is that there we used a "point intervention", which is how causation is defined in PCMs, setting X to a fixed value x and deleting the mechanism for X entirely, while here we replace the mechanism by a new, randomized one, whose output we denote by X^e.

experimentation, and are usually not based on PCMs. However, validity can also be discussed using PCMs, and we will do so in the next point.

- **Definition and analysis of validity based on PCMs, "interventions" and "soft interventions".** Consider some population whose correct causal structure is given by M with causal DAG as depicted in Figure 2.2a (which does not limit generality, as the hidden Z could be anything). Let us, for the moment, define a valid randomized experiment to be one that corresponds to the *soft intervention*[8] [Eberhardt and Scheines, 2007] of replacing the conditional $p_{X|Z}(x|z)$ in M by the new[9], unconditional density $p_{X^e}(x)$ (with support everywhere) of the variable X^e that replaces X. The intuition is that randomizing X means replacing its generating mechanism by randomization, which by definition makes the new X^e independent of the past of the universe and thus of all possible confounders Z. See Figure 2.2b for a graphical illustration. (Clearly, this definition of validity is rather far from empirical, as we based it on M, which we want to infer in the first place; but nonetheless, it clarifies the idea of validity.) Let M^e denote the resulting causal structure, which models the experimental setting, whose DAG is depicted in Figure 2.2c, and whose joint density we denote by $p^e_{X^e,Y,Z}(x,y,z)$. Now

[8]The author of this thesis was not able to find a precise definition of "soft" interventions in the literature. It needs to be mentioned though, that also interventions where the conditional $p_{X|Z}(x|z)$ is replaced by another *conditional*, which still depends on Z, are considered as "soft" interventions [Eberhardt and Scheines, 2007].

[9]It generally woks out better to introduce new variables instead of redefining old ones in such cases.

the crucial point is that

$$p^e_{Y|X^e}(y|x) = \sum_z p^e_{Y|X^e,Z}(y|x,z)p_Z(z) \tag{2.1}$$

$$= \sum_z p_{Y|X,Z}(y|x,z)p_Z(z) \tag{2.2}$$

$$= p_{Y|\mathrm{do}\, X=x}(y|x), \tag{2.3}$$

where Equality 2.2 is based on how we obtained M^e from M: the conditional for Y was not changed, i.e., $p^e_{Y|X^e,Z} = p_{Y|X,Z}$; and Equality 2.3 is simply Definition 2.4. This means that a valid randomized experiment identifies the true causal effect, when accepting the definition of validity based on soft interventions as well as that of causal effect based on interventions in Section 2.1.2.1. This sort of reasoning has been analyzed recently [Pearl and Bareinboim, 2011a, Bareinboim and Pearl, 2014], on a more general level, under the title "transportability".

It remains subject to debate to what extent the notion of a soft intervention can be made empirically meaningful. How would one test whether the experimenter does not accidentally also manipulate the mechanism that generates Y – $p_{Y|X,Z}$ in the above calculation? (See also the brief discussion of modularity in Section 2.1.5.)

- **An "ideal-empirical" definition.** Instead of defining causation based on interventions or randomized experiments, one could give the following *"ideal-empirical"* definition[10] (which is similar to, but tries to go deeper than what we referred to as "perfect" intervention above), trying to capture what empirical causal inference *aims* at: Assume we are concerned with the effect of a variable X, defined for some time point t, on a subsequent variable Y.

  The whole universe would have to "run" several times, each run starting at t, with completely identical initial condition except that the value of X varies. The causal effect would then be given by how the Y differs between the different runs.[11] This definition more or less coincides with the one based on "potential outcomes" [Imbens and Rubin, 2015], but also takes into account thoughts from Granger

[10] This definition is based on personal communication with Bernhard Schölkopf.

[11] Probably a similar definition could be given in terms of a distribution of initial conditions instead of a single one, which would be closely related to our remark on the concept of randomization below.

[1969] (see also Section 3.4.2). Note that we invoked the whole universe to avoid a circular definition: for instance, if we would just require that some sort of "isolated" system around X and Y would have to run several times, then in turn we would have to define "isolated" which may require some form of causal notion. Also note that the definition can be seen as some sort of "twin study" on the level of the universe.

It is important to note that in this definition, whether the variation in X between the several runs is due to some imaginary investigator *setting* X, or if the several runs are just *observed*, does not matter, as the rest of the initial condition is assumed to be invariant between the several runs, which excludes the possibility of a hidden confounder for X and Y.

One reasonable path to causal inference seems to be to start with such an ideal definition and then successively write down the assumptions that are necessary to infer causation in *practice*. Such a path was – to some extent – taken by Imbens and Rubin [2015].[12]

Clearly, the above ideal definition is not free of issues, some of which are: to compare the different outcomes of Y, one would have to stand "outside" the universe, which somewhat contradicts the definition of a universe; it may not be possible to vary one variable in the initial condition of the universe while keeping the others invariant, or it may even be impossible to conceptually distinguish between a variable and "rest of the universe"[13]; the way the definition relies on the notion of (global) time may be problematic under physical theories such as general relativity theory; and in general, ideal definitions are probably more a matter of taste than empirical definitions.

- **A remark on randomization.** Let us make a remark regarding the concept of *randomization*, which is central to causal inference, as its meaning seems clearer

[12] In contrast, in the work on PCMs by Pearl [2000], Spirtes et al. [2000] and others, such considerations are often swiped under the rug; "interventions" are treated as some kind of notion living in both worlds – the model world and the empirical world. But on an ideal level, a definition like the ideal one we gave above may be more helpful, and on a practical level, it is often still unclear what an intervention is supposed to be, as we discussed. But it may well be that when the notion of an intervention is better understood it can perform the balancing act between model and empirical world, such as the notion of force did in Newtonian mechanics.

[13] This is based on personal communication with Dominik Janzing.

than the meaning of causation itself (though not empirically testable). As we argued above, based on PCMs, if X is randomized, then its correlation with the subsequent Y coincides with its causal effect on Y. Without using PCMs, the argument can also be stated as follows, which is related to our "ideal-empirical" definition above: randomization, i.e., independence between X and all other variables measured at the same time or before X, ensures that all other variables measured at the same time or before X have the same distribution for both $X = 0$ and $X = 1$ (in case X is binary); so all differences in the distribution of Y between $X = 0$ and $X = 1$ have to be due to the variation in X. Either way, w.r.t. the *population in a randomized experiment*, randomization reduces the difficult notion of causation to the much simpler notion of correlation. Clearly, this ignores the problem that often we want to infer a causal effect w.r.t. some *original population* instead of the experimental population, which brings us back to the problem of validity of a randomized experiment, which we commented on above. In spite of its importance, we only briefly discuss randomization in this thesis (see also Section 2.1.5).

- **Statistical issues.** In Definition 2.6, we ignored statistical finite-sample issues, or, more broadly speaking, the problem of induction from finite observations. Instead, we pretended that experiments and observations would directly give us population-level distributions. In a more precise definition, one would rather have to speak of *falsification* (w.r.t. some fixed significance level) and *estimation* of the causal effect of X on Y.

- **Other meanings and definitions.** All in all, the concepts which we reduce causation to – interventions and randomized experiments – are themselves not free of controversy.[14] But arguably, these concepts seem significantly less opaque than causation itself.

  It is worth emphasizing that Definition 2.6 does not provide a meaning for all *usages* of "causation". For instance, recall the abdominal pain example from Section 1.1 which concerned the cause of one individual event instead of a persistent variable.

[14] Another criticism of these sort of definitions, especially randomized experiments, would be that they confuse the *meaning* of causation with how to *empirically test* causal statements. But the difference between both is hard to discern.

And there are yet other usages of "causation". For instance, Hume and Hendel [1955] writes: "we may define a cause to be an object, followed by another, and where all the objects similar to the first are followed by objects similar to the second". If we take this literally, not variables and not even events form the causal relata, but rather *objects*. ("Causal relata" means the subject and object of a causal statement.) Yet another usage can be observed in Aristoteles' work [Falcon, 2015, Hulswit, 2004]: He considered the "material" cause as one type of cause (out of four). But the "material" cause more or less means the physical substance a body is made of. This indicates that back then, the meaning of causation was somewhat broader than and different from what it is today.

Also keep in mind that even when restricting to variables as causal relata, there are yet other definitions of causation. Sometimes, causation is defined in terms of the "underlying mechanisms" [Pearl, 2000] that may be known based on say physical or chemical theory. Another definition was suggested by Granger [1969], and we will come back to it in Section 3.4.2. Historically, causal statements were also seen as answers to "why?"-questions and we will get back to that in Section 2.1.5.

2.1.3. Learning causal models

One of the two main topics of this thesis, mainly spanning Chapters 3 and 4, is causal learning in the framework of PCMs. In this section, we introduce some background as well as terminology, and briefly discuss the parts of this thesis that fall under this topic.

2.1.3.1. Definition and classification of causal learning

By *causal learning* (alternatively: *causal inference* or *causal induction*) we mean the material and mental process that aims at concluding correct causal models based on prior knowledge as well as interaction between investigator and investigated system (in

the sense of measurements and manipulations of the system).[15] The process may also be (partially) automated.

Remark 2.3 (Provisional features for classifying causal inference methods)**.** *One focus of this thesis is on the* variety *of methods for causal inference. A systematic classification of them is difficult due to their heterogeneity, but useful to compare and understand them. In this sense, within the scope of this thesis, we propose to provisionally classify*[16][17] *causal inference based on the following* two features of causal inference methods*, which are closely intertwined:*

(1) the form of constraints on the causal model the method yields (for instance, it may output constraints on $P(Y|\mathrm{do}\, X = x)$, for some X and Y),

(2) precisely which characteristics of the settings the method is applicable to, including the form of potential prior knowledge, lead to the constraints (for instance, temporal knowledge implies a constraint on the causal ordering).

2.1.3.2. Classical methods

We briefly (and without raising any claim to completeness) review some established classes of causal inference methods with a focus on the two features we proposed above.

- **Experimental causal inference:** It is immediate from Definition 2.6, part (a), that the causal effect of X on Y can be inferred through randomized experiments

[15] Alternatively, one could also define *causal inference* as any inference that aims at drawing causal conclusions (statements about cause-effect relationships), and as *causal learning* or *causal induction* the type of inference where the causal conclusion does not follow with "logical" necessity from the premises.

[16] This is just a provisional classification for this thesis. More systematic classifications have been proposed. For instance Imbens and Rubin [2015] classify w.r.t. the assignment mechanism, i.e., the mechanism that generated the "treatment" variable, whose effect on the "outcome variable" we aim to infer.

[17] Based on our definition of causal learning, another important class of methods are those that *sequentially* decide about the next experimentation step, which cannot be described so well by our two features. Such methods are also closely related to reinforcement learning, which we will briefly touch in Section 2.1.4. Note that, figuratively speaking, sequential methods would add an arrow from "decision" to "system itself / prior knowledge" in the diagram in Figure 1.1 on page 14.

– no reasoning left to be done.[18] But if one starts from the intervention-based definition in Section 2.1.2.1 instead, one can argue why and how randomized experiments form a valid causal inference method, as we did in Section 2.1.2.2 based on the notion of soft interventions.

In terms of the two features of causal inference methods we proposed in Remark 2.3, the argument can be (re-)phrased as follows. Randomization implies the absence of confounding. Assuming that the experiment is a soft intervention implies that only the assignment mechanism for X changes between original population and experimental population. Together, this allows to conclude that the experimental conditional distribution of Y given X uniquely identifies the post-interventional distribution $P_{Y|\mathrm{do}\,X=x}$.

- **Observational causal learning:** By observational causal learning we mean causal learning from a presumably independent and identically distributed (i.i.d.) sample of measurements from some multivariate distribution, *without any further causally relevant information*. In particular, the sample usually does not come from an experiment, and even if it came from an experiment, this would not be known.[19] Stated differently, by observational causal learning we mean any causal learning method which "at most" uses some multivariate distribution as input – nothing more, but potentially less (say only a finite sample from the distribution). Although not being in the scope of this thesis, we discuss observational causal learning in some detail as it is probably the class of causal inference methods that has been studied most intensively within the framework of PCMs, and also because it inspired some ideas in this thesis. (It would be equally justified to consider causal inference from time series as falling under – a broader notion of – observational causal inference, and then a significant part of this thesis would in fact fall under this topic.)

  Keep in mind that, as usual [Spirtes et al., 2000, Pearl, 2000], we say that, given variables $X_1, \ldots, X_n$, their joint distribution obeys the *causal Markov assumption*

[18] Clearly, meaning and inference of a statement (or how to argue for a statement / its "truth") are closely intertwined. The goal of inference usually is to conclude a *correct* statement, but the correctness of a statement can hardly be judged without having its meaning.

[19] Clearly, also causal inference from experiments is based on *observations* – but besides these observations, the setting includes (randomized) manipulations (which produce the observations).

if statistical conditional independence between them implies d-separation in the correct causal DAG over them. And if the converse holds true, we say that the joint distribution obeys the *(causal) faithfulness* assumption. The causal Markov assumption can be seen as a generalization of "Reichenbach's principle" which we will briefly discuss in Section 2.1.5. Causal Markov and faithfulness (and causal sufficiency) assumption together allow the reasoning underlying one of the most popular methods for causal inference from purely observational data, the "PC algorithm" [Spirtes et al., 2000]. This algorithm is solely based on conditional independences, and usually is not able to identify the causal diagram uniquely, instead it just outputs the so-called "Markov equivalence class" [Spirtes et al., 2000].

Other methods go beyond conditional independences, taking into account more properties of the observed distribution [Peters et al., 2017, Mooij et al., 2016]. Examples include methods based on the additive noise model [Peters et al., 2014, Shimizu et al., 2006] or the information-geometric approach to causal inference (IGCI) [Janzing et al., 2012], where the latter only applies to the case of two variables so far. These methods *uniquely* identify the causal DAG, given their underlying assumptions are correct.

Generally, since causal knowledge is significantly richer than statistical knowledge, causal inference from purely observational data seems limited in its possibilities. The more it tries to identify the causal model, the stronger assumptions are necessary, assumptions which may only hold in special cases (and the domain these special cases belong to may be unknown).[20] Nonetheless, there are empirical hints that causal learning from pure observations works to some extent also in broader domains [Mooij et al., 2014]. And on a more theoretical level, while pure observations cannot uniquely identify the underlying causal structure in general, there seems to be no principle reason that excludes that in rather broad domains pure observations (1) can help to approximate the causal structure and (2) can out-

[20] Roughly speaking, general assumptions plus specific knowledge (in the form of just measurements or beyond) yield causal conclusions, and fixing the available knowledge, more assumptions imply stronger identifiability results, less assumptions weaker identifiability results. One motivation underlying this thesis is to find settings, were forms of knowledge – "inputs" to causal inference – are available beyond pure observations, such that less assumptions may still constrain the set of candidate causal models strong enough (though not lead to unique identification).

perform random guessing *on average* (that is, do not work in every case, but still "correlate" with causation when considering many instances).

The study of observational causal inference is strongly motivated by the cheapness of observational data (Section 1.1), and its value of "prioritizing the experimental search space": in case it is not clear what randomized experiment to perform next, nothing is lost by letting this decision be informed by observational causal inference methods.

Maybe one of the main challenges of observational causal inference is to explicitly identify the domains in which the respective observational causal learning methods work.

- **Back-door criterion.** An important results for causal inference based on PCMs is the so-called back-door criterion [Pearl, 2000, Spirtes et al., 2000] (and the front-door criterion, which is closely related). Given a set of variables V, a subset of observed variables $W \subset V$ whose joint distribution $P(W)$ we are given, and knowledge of the causal DAG G underlying V, the back-door criterion tells when and how a causal effect between variables in W can be (uniquely) identified from the given. (The trivial case is $W = V$ which means that we are already given the complete GCM and thus obviously can calculate all post-interventional distributions – unless the joint distribution $P(W)$ does not have support everywhere.)

  The basic idea underlying the back-door criterion is to look at the definition of the post-interventional distribution (Definition 2.4) and see which parts of the joint distribution it depends on and which not.

  Let us give an example.
  Example 2.2 (Back-door criterion)**.** *Consider the causal DAG in Figure 2.3. We*

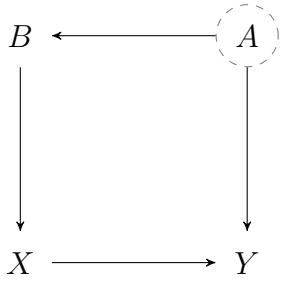


Figure 2.3.: Simple example of the back-door criterion: the effect of X on Y can be identified in spite of A being hidden.

have, based on Definition 2.4,

$$p(y|\mathrm{do}\, x) = \sum_{a,b} p(y|x,a,b)p(a,b) \tag{2.4}$$

$$= \sum_{a,b} p(y|x,a,b)p(a|b)p(b) \tag{2.5}$$

$$= \sum_{a,b} p(y|x,a,b)p(a|x,b)p(b) \tag{2.6}$$

$$= \sum_{a,b} p(y,a|x,b)p(b) \tag{2.7}$$

$$= \sum_{b} p(y|x,b)p(b), \tag{2.8}$$

where Equation 2.6 is due to d-separation in the causal DAG. So we identified $p(y|\mathrm{do}\, x)$ *from* $p(x,y,b)$ *only (which we assumed to be observed), as Equation 2.8 demonstrates. Observe that* B *blocks the "back-door" path from the hidden confounder* A *to* X*, which gives the back-door criterion its name. But the back-door criterion also applies to more complex causal DAGs.*

Let us briefly summarize one of the main ideas underlying inference based on PCMs: Past observables are related to future observables by assuming that some mechanisms (i.e., structural equations) of the system that generated the past observations reoccur in the system underlying the future observations, while allowing some other mechanisms to vary between the two systems. Depending on the precise variations and invariances, this allows to constrain or even uniquely identify the distribution of the future observables. The most important example is the future system being a system where some observables are set to constant values, in which case certain mechanisms/causal influences are

revealed.

2.1.3.3. Class of methods this thesis focuses on

In this thesis we try to advance causal inference methods that fall into the following two classes, which are closely intertwined:

- Regarding feature (2) of causal inference methods we defined in Remark 2.3, we investigate *integrative* methods. We mean integration in several ways, which *overlap*.

  On the *level of direct interaction* with the system under investigation, we mean that we include settings beyond perfectly randomized experiments or purely observational (i.i.d.) data. Examples include "partial" randomization (Section 4.5.2) or time series measurements (Chapter 3). (As mentioned in Section 1, the motivation for such integration is that randomized experiments or observational studies alone are sometimes an unsatisfactory basis for causal inference, as they may be too "expensive" or contain too little causal information, respectively.)

  On the *level of knowledge given a priori* [21], we aim to integrate all causally relevant and potentially heterogeneous information about the system that is available. An example of such prior knowledge are system specifications and program code in the case of computer systems (Sections 5.4.2.1 and 5.6.2, where we only briefly touch this idea though). Ideally, one would also include knowledge in the form of descriptions in (simple) natural language.

  On the *level of inference*, we mean the *synthesis* of the available partial information into a – not necessarily unique – global causal model. Examples include integration of sandbox experiments (Sections 5.6.1 and 5.4.3) and "plugging" together marginal and/or conditional distributions, by (partially) knowing the causal DAG (Sections 5.3.2 and 5.5.2).

[21] "A priori knowledge" which serves as inputs to causal inference is not to be confused with the *assumptions* that underly many causal learning methods and that are usually not based on specific knowledge of the systems they are applied to. Often, assumptions are lawful and so are "true" only if they are tautologies, in the narrow sense, while knowledge is true by definition, in the narrow sense, although we may sometimes mix knowledge with belief. Rigorously distinguishing between knowledge and assumptions (and beliefs) is not always possible though.

Note that inferring a global causal model on the basis of heterogeneous data sets, for instance, data sets from overlapping but different sets of variables, has been investigated by Tsamardinos et al. [2012] under the name "integrative causal analysis". And while we use "integrative" as a very preliminary working title in this thesis not to be meant as a proposal of a lasting definition, it can be seen as a generalization of the definition by Tsamardinos et al. [2012], including more kinds of "inputs".

- Regarding feature (1) of causal inference methods we defined in Remark 2.3, we investigate *approximative* methods: methods that often do not lead to a *unique* identification of the correct causal model, but still constrain the set of explanatory causal models.[22]

  An important result in this direction is the PC algorithm we mentioned in Section 2.1.3.2 above, which works on the basis of purely observational data and outputs the Markov equivalence class of the correct causal DAG (if the underlying assumptions hold true). But one can think of a whole variety of ways in which causal models can be constrained – for instance in terms of causal ordering. PCMs provide an expressive language for formalizing the various constraints. Within this thesis, examples include identification of structural coefficients up to a finite number of possibilities (Section 3.6.2), approximations of the causal effect based on bounds on confounding (Section 4.4), approximation of the structural counterfactual, a property of an FCM (Section 5.3.1) and approximate integration of conditionals (Section 5.3.2).

  Clearly, it depends on the specific scenario whether approximate causal inference does provide helpful insights, or if the approximations are to coarse to be meaningful. Nonetheless, it seems that aiming for approximate identification of causal models is often *more realistic* than aiming for unique identification.

A reoccurring issue in our investigation of causal learning will be *hidden confounding*: If ones assumes, besides having distributions, to know the causal ordering of the observed variables (to some extent), then hidden confounding remains as the primary challenge for

[22] We ask the reader to kindly excuse some imprecision in terminology here. It is clear that inductive inference always contains some uncertainty. What we mean here by "approximate methods" are methods that fail to uniquely identify the true causal model even on the *population-level*.

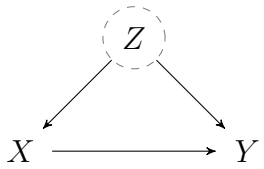


Figure 2.4.: Example of hidden confounding.

causal inference. This is because if we know the causal ordering, under the assumption of causal Markovianity and faithfulness, the observational distribution already determines the causal model – unless there are hidden confounders [Spirtes et al., 2000]. In this sense, hidden confounding will play a central role in Chapters 3 and 4.

Let us give an example to illustrate the problem of hidden confounding.

Example 2.3 (The problem of hidden confounding)**.** *We show how hidden confounding can distort (naive) observational causal inference. The example is stylized, but similar studies were in fact performed [Lawlor et al., 2004].*

Consider the variables X, Y, Z, where X denotes the dose of hormone replacement therapy applied, $Y \in \mathbb{R}$ denotes the severity of subsequent coronary heart disease, and Z denotes the wealth. Assume the true causal model is given by the DAG in Figure 2.4 and the structural equations

$$\begin{aligned} X &:= Z + N_X, \\ Y &:= 0.5X - Z + N_Y, \end{aligned}$$

where we leave the noise distributions unspecified for the moment. A a purely observational study, that only considers X, Y, yields

$$\mathbb{E}(Y|X = 1) - \mathbb{E}(Y|X = 0) = -0.5,$$

based on

$$Y = 0.5X - (X - N_X) + N_Y = -0.5X + N_X + N_Y.$$

This may mislead to the conclusion, that hormone replacement therapy X causes a less

severe coronary heart disease Y, while in fact

$$\mathbb{E}(Y|\mathrm{do}\, X = 1) - \mathbb{E}(Y|\mathrm{do}\, X = 0) = 0.5,$$

based on

$$\mathbb{E}(Y|\mathrm{do}\, X = x) = 0.5x + \mathbb{E}(-Z + N_Y),$$

calculated from Definition 2.4. That is, hormone replacement therapy X actually causes (a more severe) coronary heart disease Y. The reason for the observational study to be mislead is the strong negative correlation between X and Y that is induced by the unobserved wealth Z: a high Z induces higher intakes X and (say due to more money spend on health in general) lower Y.

2.1.4. Causal models for decision making

On the one hand, having a good causal understanding of the world can be seen as an end in itself. On the other hand, a main focus of this thesis is on causal models as a *means* for informed *decision making* towards given goals. We already gave examples in Section 1.1: a personal decision regarding which drug to take should be informed by knowledge on the effect of the available drugs; decisions on fiscal policy should be informed by models about the effects of (high) state debt; political or civil action against nationalist and populist movements should be based on an understanding of the causes for the emergence of such movements. Generally, it may be that directly the effect of a decision is of interest, or it may be that one is interested in causal effects that are only indirectly linked to a decision.

In this section, we give some background for decision making using PCMs and briefly discuss the parts of this thesis that fall under this topic.

Decision making has been investigated intensely in the field of "decision theory" [Steele and Stefánsson, 2016], which, roughly speaking, studies the reasoning that leads to an "instrumentally rational" agent's choice. By "instrumentally rational" we mean agents that (1) have (own) goals and (2) chose suitable means to achieve them.[23]

[23] Often, such behavior is simply called "rational". But it can be argued, that a general concept of rationality also comprised the choice of goals, a choice which does not itself have a clear goal, but

One of the most popular formal approaches is expected utility theory [Von Neumann and Morgenstern, 2007, Jeffrey, 1990], according to which instrumentally rational agents, in uncertain situations, make the choice that maximizes their expected utility. In other words, instrumentally rational agents are modeled as solving an *optimization* problem – under uncertainty – with the objective function formalizing their goal. This hypothesis can be justified based on a certain axiomatization of instrumental rationality, as Von Neumann and Morgenstern [2007] showed.

Closely related to this, the problem of designing intelligent machines (or "artificial intelligence") is often cast as designing automatic agents that sequentially *"decide"*, under uncertainty, such as to optimize some predefined utility (or "reward") function. This paradigm is often referred to as "reinforcement learning (RL)" [Sutton and Barto, 1998], and is closely related to control theory [Aström and Murray, 2010]. A core challenge in this paradigm is the trade-off between exploration to improve the agent's model and exploitation of (i.e., maximization of the utility under) the current model.

Causation is often not *explicitly* thematized in studies of decision making – which does not mean that it is not *implicitly* handled correctly. However, a subfield of decision theory, referred to as *causal decision theory* [Steele and Stefánsson, 2016, Weirich, 2016, Lewis, 1981, Woodward, 2005], emphasizes the importance of being explicit about causation in decision making.[24] Otherwise, one may make errors, such as confusing, on the one hand, a *correlation* of outcomes/utilities and actions with, on the other hand, outcomes/utilities that are *caused* by actions. To see why this can be an error, keep in mind that often one thinks of an agent as independent of the system under control and actions as exogenous to the system; but this implies that actions can be seen as *interventions*.[25] (Note that in this simple argument we only refer to actions that are not influenced by the system. But the argument can be extended to other actions as well, using soft interventions as introduced in Section 2.1.2.3.) Recently, there has also been some work that investigates the role of *PCMs* for (automated) decision making problems: Bottou et al. [2013], Bareinboim et al. [2015] establish connections between a

can still be based on arguments. A further discussion of this issue is far beyond the scope of this thesis and we refer the reader to [Kolodny and Brunero, 2016].

[24] Clearly, decisions should not only be judged in terms of their effects, but sometimes also in terms of ethical considerations, which may be based more on the motivation than the result Kant et al. [2002], although both are difficult to discern.

[25] This is based on personal communication with Michel Besserve.

standard model of RL – "bandits" – and causal models.

Note that causal models can help for decision making in two ways: On the one hand, if the decision corresponds to *one variable* D in a causal model M, then M allows to predict the outcome of a specific decision (formalized by an intervention on D) as well as the outcome of a policy for decision making (formalized by a soft intervention on D, see Section 2.1.2.3). On the other hand, if the decision is about a more general manipulation of a system modeled by a causal model M, then it can also be seen as an element of the set of all possible interventions on M, including interventions on *several variables*.

We do not consider this thesis as a general contribution to causal decision theory. However, we will show how causal models can help for decision making towards predefined goals in some *specific* cases involving complex systems. The basic idea is to first *formalize* questions that arise in decision making as *queries about parts of causal models* and then to study how these parts can be inferred (integratively and approximately).

- In Section 4.5.2, we consider decision making in a toy scenario in advertisement. Specifically, we inform the decision of whether to send out an advertisement letter or not by approximately inferring the strength of the causal effect of sending out such a letter.

- In Section 5.4, we consider decisions about how to *manipulate* a (cloud) computing system to debug its performance. We show how the outcomes of such manipulations can be formalized as counterfactual probabilities. We discuss how these probabilities can approximately be inferred from a given GCM (since FCMs, which would uniquely determine counterfactual probabilities, often cannot be inferred), and, to some extent, also how the GCM can be inferred. In Section 5.4 we also discuss automated allocation, i.e., *control*, of cloud computing resources based on causal models. We also show how causal models can help to integrate *sandbox experiments* for debugging and control decisions.

- Furthermore, in Section 5.5 we cast the outcomes of clients bidding for cloud computing resources as post-(soft)interventional distributions. We discuss how such distributions can be inferred approximately and integratively from knowledge distributed among the clients, potentially helping them in their decision making. However, we leave strategic considerations based on game theory to future work.

Note that while we always informally specify goals, we do not always explicitly formalize them by a utility function. In this sense, we also do not always run through the complete optimization procedure, but instead we rather show how reasoning based on PCMs can be *one step* in rational decision making.

It is worth emphasizing again that we do not claim that PCMs are *necessary* for the decision problems we apply them to. For instance, if a decent class or distribution of global models for past and future observations, conditioned on decisions, is *given*, then explicit causal reasoning is not needed, and the inference from observations to decisions can be done in an "end-to-end" manner. A prominent example of such a model class is given by "Markov decision processes" [Sutton and Barto, 1998, Barber, 2012]. However, to come up with such (classes or distributions of) global models (or approximations, see 5.3.2), or to "critically" reason about identifiability of parts of the model from the available data (which we will briefly do in Section 5.4.3), PCMs *can* be helpful. Furthermore, causal models may be of help whenever learning happens but the precise goal or the possible actions are not specified yet.

2.1.5. Contextualization

Before going into rather technical and mathematical details, of which a majority of this thesis consists, we want to take a step back and take a broader view on causation by putting this concept into context with other important concepts, and also by making some historical remarks.

On the one hand, this contextualization is of value when addressing the problem of causal inference: *to infer causation, it is important to understand how causation "a priori" relates to other concepts, because any such relation may allow to reason from non-causal statements, which are potentially easier to infer, to causal statements.* On the other hand, a (historical) contextualization is one of the most powerful standard methods for systematic reflection upon a topic. (Another powerful method is critical analysis – weighing the pros and cons, the potentials and limitations – and we will perform such analysis to a limited extent in Chapter 6. A third "method" may be seen in asking "why?"-questions, looking for the driving forces – searching for the underlying causes.)

We organize the contextualization by several key words. Some relations have already been touched in previous sections, in particular the ones to manipulation and correlation. Here we add some more "basic" and historical remarks.

- **Manipulation.** Our definitions in Section 2.1.2 already indicate the key role of *manipulations* for the concept of causation, as these definitions rely on interventions and experimentation. Generally, manipulation relates to causation in two ways: on the one hand, manipulation of an object helps to infer a causal model of it; on the other hand, a PCM predicts the outcomes of a special kind of manipulations of an object – interventions.[26] The latter direction makes causal models particularly interesting for decision making, which is often about potential manipulations of some system. We do not want to claim though, that causal models are necessary whenever outcomes of manipulations need to be predicted: there are other manipulations than interventions, and generally, there are cases where one can speak about manipulation without any need to resort to causation.

  Historically, the key role of manipulation for causal inference was recognized at least since Francis Bacon [Spirtes et al., 2000, Section 1.1]. Other important proponents of the necessity of manipulation are John Stuart Mill [Macleod, 2016], Donald Rubin and Paul Holland [Holland, 1986]. The growing emphasis on manipulation needs to be seen in the context of the development of the modern scientific method. It heavily relies on experimental interaction of the investigator with the object of interest – repeated, deliberate manipulation and observation – in contrast to sole deduction of (empirical) insight from first principles.[27] As Kant [1998, BXIII] describes it, in modern science, it is about performing experiments to *force* nature to answer the questions that we pose in *our* language. He gives the example of Galileo letting a ball whose mass he *picked himself* roll down an inclined plane to infer the physical laws governing it.

  In this thesis, the relation between manipulation and causation will in particular

[26] Clearly, the two ways in which "manipulation" relates to "causation" – inference and prediction – can be seen as two sides of the same coin.

[27] The growing connection between theory and manipulation of the empirical world may be driven by the economic development from slaveholder and feudal societies towards market economy [Marx, 2014, 1867, Smith and Recktenwald, 1986]. In a slaveholder society, there were probably less people with the intellectual education, a connection to practical problems, and incentives as well as a possibility to apply the former to solve the latter.

play a role for Chapter 5, where causal models are proposed to inform decisions about "which knob to turn" to debug a computer system.

- **Counterfactuals.** When considering causation w.r.t. individual units instead of populations, then usually one runs into the notion of counterfactual statements and questions [Pearl, 2000, Imbens and Rubin, 2015], which are closely related to the "potential outcomes" we touched in Section 2.1.2.3. For instance, one can read the statement "Taking the aspirin made my headache go away." by "If I had not taken the aspirin (that is, if I had intervened on my decision to take the aspirin), then my headache would not have gone away."

  As most people would agree, the latter statement cannot be verified or falsified in reality (without making further assumptions [Pearl, 2000]). This is referred to as the "fundamental problem of causation" [Holland, 1986].

  In this work, most of the time we avoid this problem by focusing on causal effects on the population-level instead of the unit-level, where one can (under appropriate assumptions) in fact apply treatment as well as control to (samples of) one and the same population, in case one samples and randomizes properly. In Chapter 5 though, we will interpret debugging questions in computer systems as unit-level counterfactuals, and answer them (approximately) with the help of PCMs, assuming, for instance, that the value of the background variables remains invariant [Pearl, 2000] (see also Remark 5.1).

- **Correlation.** Causation is closely related to correlation or, more precisely, (statistical) dependence in various ways. Probably the most well-known assumption in this direction is the so-called *common-cause principle* by Reichenbach [1956]. It states that whenever two variables X and Y are dependent, either they have a common cause, or one causes the other (the latter can be seen as a "degenerate" special case of the former). Going beyond the "qualitative" level of dependence towards a "quantitative" level, one may say that the causal structure underlying a set of variables is closely linked to their observational joint distribution. We already discussed observational causal inference methods, which are completely based on this link, in Section 2.1.3.2. In particular, the causal Markov assumption we introduced there is a generalization of Reichbach's common-cause principle. But also the definition of PCMs in general heavily relies on this link, as a PCM

is a multivariate distribution plus additional information. And within this thesis, in particular in Chapters 3 and 4, correlations – and observational distributions in general – will be one of the most important inputs to causal inference, although they will not be the *sole* input.

Interestingly, from a superficial view, the notions of causation and correlation were less well-separated earlier in history. For instance, Hume and Hendel [1955, chapter VII] states that "we may define a cause to be an object, followed by another, and where all the objects similar to the first are followed by objects similar to the second", which may rather be read as some form of correlation (or implication) than as intervention-based causation (when neglecting the use of "object" instead of "event", which we already commented on in Section 2.1.2.3). In this sense, the concept of correlation may be seen as a first, easier to define and infer, approximation to the concept of causation.

So far we were merely talking about the relationship between causation and correlation on a "population level". It is important to mention though, that statistical and causal inference face the same fundamental problem – the problem of *induction*, that is, drawing conclusions from past to future, from the special case to the general law, from finite sample to entire population, from the observed to the hidden– beyond tautologies. This problem was discussed by many philosophers, in particular Hume and Hendel [1955], who considered this problem especially w.r.t. causal inference. One of the most interesting "solutions" was given by Kant and Guyer [1998], who argued that whatever can be perceived and understood about the future would be perceived and understood by a *(human) subject*. Based on this, the problem of induction can – to some extent – be reduced to the problem of understanding the subject and the "necessary" conditions that enable and structure perception and insight. This is closely related to Kants view on modern science we mentioned above.

- **Time.** It is generally assumed that future does not influence past. Since time is usually easy to measure, this assumption can be very valuable. In Chapter 3, where we study causal inference from time series, we will heavily rely on this assumption. In that chapter we will show that measuring variables over time can also allow to better remove hidden confounding compared to just having i.i.d. measurements.

Note that it is not obvious how the assumption could be falsified: how could one prove an effect from present to past, if one existed? Let W_s, X_t, Y_u be variables at times $s < t < u$, respectively. In a classical randomized experiment, to infer the influence of X_t on Y_u, one randomizes the treatment variable X_t in order to make it independent of the past of the universe and thus independent of all potential factors that also influence the outcome variable Y_u. But, when randomizing X_t in an experiment to determine its effect on W_s, this randomization would by definition always disprove an effect of X_t on the preceding W_u, since otherwise the randomization would not have been performed properly. So a classical randomized experiment does not seem appropriate to disprove an effect of X_t on W_s. Instead, one would have to come up with other forms of experiments. Alternatively, it may be that there are arguments against effects from future to past which are based on physics or may even hold true *a priori*.

- **Explanation.** The concept of causation is closely related to the concept of explanation. For instance, according to Falcon [2015], Hulswit [2004], Aristoteles considered causal statements as explanatory answers to questions regarding *why* and *what* an object is.

  In this thesis we use the concept of causation in a rather narrow and technical sense, based on our definitions in terms of PCMs, interventions and randomized experiments (Section 2.1.2). When moving within this technical sphere, it is not always obvious, how results contribute to a genuine *deeper understanding* of the world. Nonetheless, the distant goal of this thesis is to develop methods that help to find answers to the important "why?"-questions, and that contribute to explanations and understanding of the world.

  As a side not, generally, it seems an open question how much formal models and quantitative results contribute to understanding the world. Without a doubt, they form *one step*: for instance, one can have some understanding of the world, then try to translate it into a quantitative hypothesis and then falsify this hypothesis. But by no means one should confuse the quantitative hypothesis with the understanding itself. On the other hand, it may also depend on the individual how quantitative "understanding" itself is.

- **Modularity.** An important concept which PCMs are based on is *modularity* [Woodward, 2005]. In a PCM, the causal conditionals can be changed individually, and assuming modularity means assuming that (1) interventions on the mechanisms represented by the causal conditionals, including the soft interventions we discussed in Section 2.1.2.3, are well defined in reality and (2) the PCM correctly predicts the outcome of such interventions. That is, modularity means that the remaining structural equations (in case of an FCM) "not intervened on" form a valid explanation of the new system, which makes it a fundamental assumption underlying PCMs. This necessity of modularity for PCM-based causal inference was a key motivation for Chapter 5: there we perform causal inference in *computer* systems, which are often modular *by design*. (A critical stance on modularity and interventions was taken e.g. by Wiener [1956], who pointed out that we can never make entirely sure that we performed a proper intervention which did not accidentally also affect other mechanisms.)

  As a side note, causal models are generally inspired by computer science and in particular (imperative) programming. There, values can be assigned to variables (which is often denoted by a "=" with asymmetrical interpretation similar to the ":=" in structural equations), or, on a physical level, loaded into memory. This entails the metaphor of considering "influence" as "assignment".

- **What else?** Causation is also linked to the *spatial* arrangement of variables and events, as emphasized, e.g., by Hume and Hendel [1955]. It seems that this relation is rarely harnessed for causal inference so far. More recent investigations also look at how causation is related to fundamental concepts from *physics*, such as entropy [Janzing et al., 2016].

  Certain usages of "causation" can be seen as an *"anthropomorphic"* world-view. Take, for instance, Aristoteles who considered one type[28] of cause to be the "efficient" cause, by which he meant the "the primary source of the change", or the "thing responsible" Falcon [2015], Hulswit [2004]. This indicates a world-view where even lifeless entities are some sort of "agents" that "act" upon other entities. But anthropomorphic phrases like "the moon pulls water toward it, and

[28] In total, Aristoteles considered four types of causes – "material", "formal", "efficient" and "final" cause. We already mentioned the "material" one in Section 2.1.2.3. For a further discussion, in particular of the remaining types, we refer the reader to Falcon [2015], Hulswit [2004].

this causes the bulge toward the moon" [HowStuffWorks.com, 2009], to explain the mechanisms underlying low and high tide, are completely common still today. As Hulswit [2004] points out, it is only since about the seventeenth century that cause-effect relations are – besides the "anthropomorphic" usage still common today – considered as some sort of "lifeless" *laws* which one aims to infer.

It is also interesting to look at the role of the measuring and intervening *subject* in causal models. If it is possible at all to identify a trend in science, then one such trend could be seen in the growing importance to model the investigating subject: while it did not play a significant role in Newtonian mechanics, it became important in relativity theory (in the sense of a passive observer) and in quantum mechanics (in the sense of the decision maker w.r.t. measurements). Maybe causal models, based on the notion of an intervention, can be seen as part of this trend.

2.2. Overview: outline and contributions

Here, we will first provide summaries for all remaining chapters, in Section 2.2.1. Then we will discuss the main contributions of this thesis and the parts that are due to the author of this thesis, in Sections 2.2.2 and 2.2.3, respectively. Last, we list the publications this thesis is based on in Section 2.2.4.

2.2.1. Outline

In short, this thesis studies (1) approximative and integrative inference of causal models, and (2) causal models as a basis for decision making in complex systems. We repeat the schematic illustration of this work from Chapter 1 in Figure 2.5. The focus of Chapters 3 and 4 is on topic (1), the focus of Chapter 5 is on topic (2). However, Chapters 4 and 5 also contain some elements of the respective other topic.

Often, when treating topic (2) we assume more high-level information, such as a GCM, as given, and start reasoning from there; compared to our investigation of topic (1), which naturally starts from rather low-level information, such as measurements and temporal

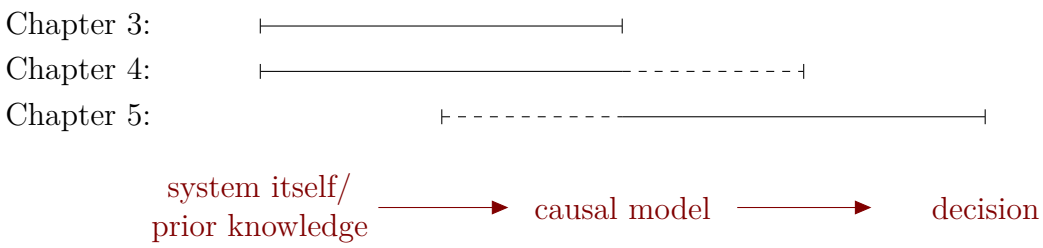


Figure 2.5.: Illustration of the structure of the remainder of this thesis. We study fragments of the "inference path" that starts at a given system and information about it and goes via causal models towards the goal of an informed decision (concerning the system), depicted in red along the x-axis. The focus of the respective chapter is depicted by a solid black line, while topics that are briefly touched are depicted by a shorter dashed line.

annotation. Recall that we presented background for topic (1) in Section 2.1.3, and for topic (2) in Section 2.1.4.

Now we give a short summary for each individual chapter:

- **Chapter 3: Causal inference from time series with hidden confounders.**
  - *Problem:* Here we study the problem of inferring a causal model of a dynamical system, which is central to areas ranging from economics to neuroscience.
  - *Integrating:* We consider a (subsample of a) time series $X = (X_t)_{t\in\mathbb{Z}}$ as given, i.e., we integrate measurements as well as knowledge of their temporal ordering.
  - *Approach:* We assume that X together with some hidden confounder $Z = (Z_t)_{t\in\mathbb{Z}}$ forms a first order vector autoregressive (VAR) process with structural transition matrix A (we give an example of a corresponding causal DAG in Figure 2.6). Then we examine under which conditions the most important parts of A are identifiable or approximately identifiable from only X. Essentially, sufficient conditions are (1) non-Gaussian, independent noise or (2) no influence from X to Z. We present two estimation algorithms that are tailored

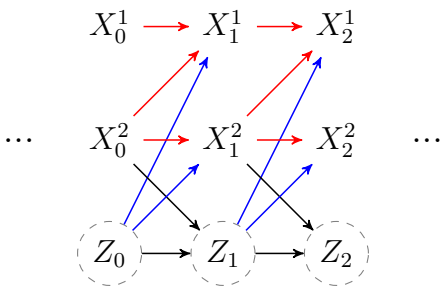


Figure 2.6.: Example causal DAG with bivariate observed time series $X = (X_t)_{t\in\mathbb{Z}}$ and univariate hidden confounder time series $Z = (Z_t)_{t\in\mathbb{Z}}$, induced by a (structural) transition matrix A. Influences within X are red, confounding by the hidden Z is in blue, and the remaining influences are in black. In Chapter 3, the goal is to infer (parts of the) causal model of such structures, in spite of Z being hidden, under the assumption of linearity. Knowing the temporal ordering means knowing the causal ordering, up to instantaneous effects, but the problem of hidden confounding remains.

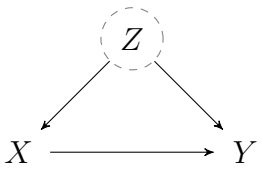


Figure 2.7.: The causal DAG that we assume in Chapter 4. The goal is to infer the effect of X on Y. We assume to know that X causally precedes Y, but the problem of hidden confounding remains.

towards conditions (1) and (2), respectively, and evaluate them on synthetic and real-world data. We present a way to check the model assumptions using only X.

- **Chapter 4: Approximate causal inference by bounding confounding in i.i.d. settings.**
  - *Problem:* The overall goal is to infer the causal effect of a treatment variable X on an outcome variable Y in i.i.d. (i.e., non-time-series) settings. In one example (Section 4.5.2) we aim at informed decision making in advertisement, where X denotes the action of sending out an advertisement letter or not.
  - *Integrating:* We assume the joint distribution of X, Y as given. We assume to know that Y does not influence X (i.e., we assume the causal DAG in

Figure 2.7). Besides that, we assume various forms of additional knowledge to be given that imply bounds on the strength of a potential hidden common cause Z of X and Y. For instance, in the advertisement example, we consider partial compliance with guidelines.

- *Approach:* Our approach is to first derive approximations (upper/lower bounds) for the causal effect, in dependence on bounds on the strength of confounding we assume as given. The approximations are derived w.r.t. a range of formalizations of "causal effect", including $P(Y|\mathrm{do}\, X{=}x)$, but also quantities such as the effect of treatment on the treated (ETT) and information theoretic quantities like information flow and causal strength. Then we discuss several scenarios where knowledge or beliefs can be integrated that in fact imply bounds on the strength of confounding, including the mentioned advertisement example.

- **Chapter 5: Decision making in cloud computing via approximate causal models.**

  - *Problem:* We consider two decision making problems that arise in cloud computing: (1) debugging and control of computing systems, and (2) bidding for "spot" resources versus buying "dedicated" resources.

  - *Integrating:* For problem (1) we integrate expert knowledge, non-causal associational information (e.g., program code), and sandbox experiments. For problem (2) we propose to integrate knowledge of the causal DAG, and partial knowledge of the conditional probabilities, distributed among the parties involved.

  - *Approach:* We formalize debugging by counterfactual probabilities and control by post-(soft-)interventional probabilities. We show that counterfactuals can approximately be calculated from a GCM (while they are originally defined only for FCMs), and based on this sketch an approach which integrates sandbox expeiments and can potentially help to address problem (1). To address problem (2), we formalize bidding by post-(soft-)interventional probabilities. We show how, in a toy scenario, cloud clients can trade off privacy

against predictability of the outcome of their buying and bidding actions, based on a simple result on approximate integration of conditionals.

2.2.2. Main contributions of this thesis

While discussing contributions in more detail in the respective chapters, here we pick and summarize the main contributions of this thesis, structured by chapter. We do not distinguish between contributions by the author of this thesis and collaborators here, but we will do so in Section 2.2.3. Besides being ordered along the "inference path" from learning to decision making which we illustrated in 2.5 on page 48, the main contributions of this thesis can also be seen as going from parts that are rather based on *"exploiting"* beaten tracks in terms of problem formulations and frameworks, to parts that are more *"explorative"* in that they try to push the boundaries of established frameworks towards relevant directions.

- **Chapter 3: Causal inference from time series with hidden confounders.** In causal learning from time series, the causal ordering implied by the temporal ordering significantly facilitates inference, but potential hidden confounding remains a major problem. Theorems 3.1 through 3.3 constitute an extensive analysis of this problem for the case of vector autoregressive (VAR) processes, which are a model frequently assumed in time series analysis [Lütkepohl, 2006]. We show that, under rather weak additional assumptions, hidden confounding can perfectly or approximately be removed, and even the "location" of confounders can be inferred. Furthermore, in Propositions 3.1 and 3.2, we prove the genericity of parts of our assumptions in a way that may be transferable to other situations.

  Theorem 3.1 can be interpreted as showing that the integration of time can – besides giving a causal ordering – also help to remove hidden confounding, as (essentially) the analogous assumptions made in i.i.d. settings only lead to a significantly weaker identifiability result [Hoyer et al., 2008].

  In Algorithms 1 and 2, we present concrete causal inference methods for time series data, tailored to the conditions in Theorems 3.1 through 3.3 that imply

(approximate) identifiability. We validate the algorithms on synthetic data, and show potentials and limitations of the first one on real data.

- **Chapter 4: Approximate causal inference by bounding confounding in i.i.d. settings.** Some causal inference methods under the title "quasi-experiments" (Section 4.2) go beyond perfectly randomized experiments, but so far only "islands" in this "continuum" between randomized experiments and observational studies are discovered. With Theorems 4.1 through 4.7, we contribute to the systematic study of this continuum. We show for a wide range of established measures of causal effect, ranging from information theoretic quantities to the effect of treatment on the treated (ETT), how they can be approximated given bounds on confounding.

  While finding real settings that imply bounds on confounding remains a challenge, for instance in Section 4.5.2, using Proposition 4.1, we show for a simplistic decision making scenario how knowledge about circumstances such as partial compliance with guidelines can be integrated to come up with bounds on confounding.

  It seems that often, aiming for approximative causal inference is more *realistic* than aiming for unique identification, and this chapter particularly contributes to the advancement of such approximative methods.

- **Chapter 5: Decision making in cloud computing via approximate causal models.** In Section 5.4.2, we propose potential first steps of a principled approach for sandbox experiment, debugging and control in cloud computing based on causal models, which can help to overcome problems of previous methods that do not handle causation explicitly (Section 5.4.3). In particular, we show how debugging questions, which are central in computer systems, can be translated into counterfactual probabilities in Section 5.4.2.4. In Proposition 5.1 we show how – theoretically – such counterfactual probabilities can approximately be derived from a GCM only, which was not known before. This proposition may be of importance beyond cloud computing, as it is rarely the case that FCMs, which are completely deterministic, can be inferred.

  In Section 5.5.2 we show how, in a toy setting, knowledge distributed among the parties involved in cloud computing can be integrated to improve bidding and allocation decisions, with the possibility to trade of prediction accuracy against

privacy. The method is based on Proposition 5.2, which indicates how in this setting and beyond, distributed knowledge in the form of marginal distributions can be integrated towards an approximate global model.

It needs to be emphasized that the practical value of the two propositions and our formalizations of the mentioned problems remains to be established. They should be seen as a thought-provoking impulse rather than a completed contribution.

2.2.3. Contributions by the author of this thesis

In what follows, the contributions that are due to the author of this thesis are listed, structured by chapters again:

- **Chapter 3: Causal inference from time series with hidden confounders.**
  - The precise formulations of all lemmas, propositions and theorems are due to the author of this thesis. Using the identifiability result underlying over-complete ICA [Kagan et al., 1973, Theorem 10.3.1] to subtract out hidden confounding – as was done in the proof of Theorem 3.1 – was suggested by Kun Zhang. The idea to heavily exploit the properties of polynomials to prove the genericity statements in Propositions 3.1 and 3.2 was suggested by Dominik Janzing. Other than that, all parts of the proofs for this chapter are due to the author of this thesis.
  - Algorithm 1;
  - and the simultaneous treatment of Granger causal inference and inference based on PCMs, are due to the author of this thesis.
- **Chapter 4: Approximate causal inference by bounding confounding in i.i.d. settings.**
  - The formulation and proofs of Theorems 4.4, 4.6 and 4.7;
  - and parts of the ideas for the various prototypical application scenarios and mathematical deviations in Section 4.5 are due to the author of this thesis, while other parts were contributed by Dominik Janzing.

- **Chapter 5: Decision making in cloud computing via approximate causal models.**
  - The formulation and proofs of the mathematical results, Propositions 5.1 and 5.2 are due to the author of this thesis.
  - The general idea to apply PCMs to problems in cloud computing was developed by the author of this thesis, together with Lucian Carata.
  - The application of machine learning algorithms to experimental data (the experiment itself was performed by Lucian Carata) is due to the author of this thesis.

2.2.4. Underlying publications

This thesis is build upon several publications/preprints. They are listed, together with their appearances, in Table 2.1.

Table 2.1.: Publications/preprints this thesis is builds on, and where they appear.

Publication	Used in
P. Geiger, K. Zhang, M. Gong, D. Janzing, and B. Schölkopf. Causal inference by identification of vector autoregressive processes with hidden components. In *Proceedings of 32th International Conference on Machine Learning (ICML 2015)*, 2015b [Geiger et al., 2015b]	Chapter 3
P. Geiger, D. Janzing, and B. Schölkopf. Estimating causal effects by bounding confounding. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*, pages 240–249, 2014 [Geiger et al., 2014]	Chapter 4, Appendix B
P. Geiger, L. Carata, and B. Schölkopf. Causal inference for cloud computing. *arXiv preprint arXiv:1603.01581*, 2016b [Geiger et al., 2016b]; P. Geiger, L. Carata, and B. Schoelkopf. Causal models for debugging and control in cloud computing. *arXiv preprint arXiv:1603.01581*, 2016a [Geiger et al., 2016a]	Chapter 5, Appendix C
P. Geiger, K. Zhang, M. Gong, D. Janzing, and B. Schölkopf. Causal inference by identification of vector autoregressive processes with hidden components. *arXiv preprint arXiv:1411.3972*, 2015a [Geiger et al., 2015a]	Appendix A

Chapter 3.

Causal inference from time series with hidden confounders

3.1. Introduction

This is the first of two chapters that focus on integrative and approximative causal inference, as indicated in Section 2.2.1 and depicted in Figure 3.1 on page 57. Specifically, in this chapter we focus on causal inference from *time series*, which, besides measurements, also integrates the temporal ordering of these measurements. Time series data being so cheap to obtain in many situations, it is surprising how valuable it is for causal inference. Its value is based on the common notion that the future does not influence the past and therefore a temporal ordering of the measurements is also a causal ordering, up to instantaneous effects (our model assumptions will be further detailed in Section 3.4.2).

Causal inference from time series has been performed in many areas such as neuroscience [Roebroeck et al., 2005]. But it is particularly relevant for economics [Lütkepohl, 2006], as causal models are more informative for decision making than purely correlational ones (Section 2.1.4), and a lot of data in economics naturally comes as time series, such as yearly gross domestic product (GDP). Maybe the most widely applied method is so-called Granger causal inference, proposed and argued for by Granger [1969]. The way Granger causal inference is usually implemented has a significant weakness: it does not account for potential hidden confounders, as will be further detailed in Sections 3.4.2

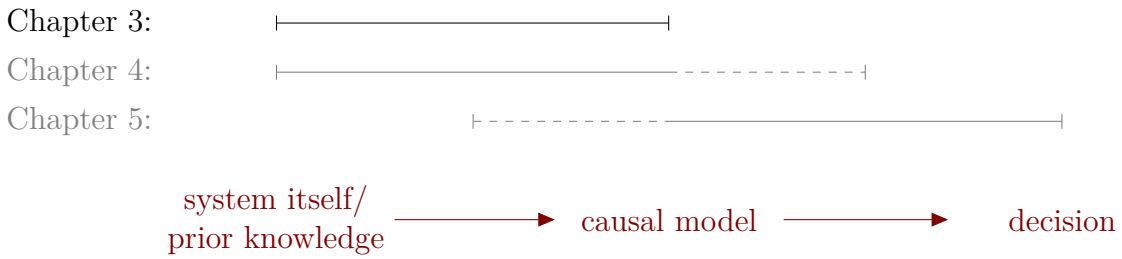


Figure 3.1.: The content of this chapter illustrated in black, relative to the rest of this thesis in gray, and the overall "inference path" in red.

and 3.4.3. We will use the framework of PCMs to analyze this weakness and to come up with ways around it. But while we will see that the problem of hidden confounding can be somewhat weakened when considering time series instead of i.i.d. data, it cannot be fully solved – a reason why we also put a focus on approximative methods which do not aim to *uniquely* identifying the correct causal model.

Parts of this chapter are based on the publication [Geiger et al., 2015b].

3.1.1. Problem statement

We assume we are given a multivariate time series sample

$$x_{1:L} = \begin{pmatrix} x_1^1 \\ \vdots \\ x_1^{K_X} \end{pmatrix}, \ldots, \begin{pmatrix} x_L^1 \\ \vdots \\ x_L^{K_X} \end{pmatrix},$$

where K_X is the dimension of x_t, $t = 1, \ldots, L$. The overall goal is to infer the causal structure underlying $x_{1:L}$, that is, to infer how the observable underlying x_t^k influences the observable underlying x_{t+s}^l, for any k, l, t, s. The more specific goal is to understand, to what extent – under what assumptions – and how – with which estimation algorithms – the causal structure can be inferred from the sample.

3.1.2. Outline of our approach

In the language of FCMs our approach can be described as follows. (For an alternative justification, based on Granger's theory, see Section 3.4.2.) We assume that $x_{1:L}$ is a finite sample of a multivariate random process $X = (X_t)_{t\in\mathbb{Z}}$ which, together with another multivariate random process $Z = (Z_t)_{t\in\mathbb{Z}}$, obeys the structural equations

$$\begin{pmatrix} X_t \\ Z_t \end{pmatrix} := \begin{pmatrix} B & C \\ D & E \end{pmatrix} \begin{pmatrix} X_{t-1} \\ Z_{t-1} \end{pmatrix} + N_t,$$

for all $t \in \mathbb{Z}$, some matrices B, C, D, E, and some i.i.d. $N_i, i \in \mathbb{Z}$.

The first stage of our investigation, Section 3.6, is on the theoretical side: we present several results that show under which conditions B and C are identifiable or approximately (i.e., up to a small number of possibilities) identifiable from only the distribution of X. Generally we assume that Z has at most as many components as X. Theorem 3.1 shows that if the noise terms are non-Gaussian and independent, and an additional genericity assumption holds true, then B is uniquely identifiable. This indicates that temporal knowledge – besides giving a causal ordering – also allows to better remove hidden confounders compared to just having i.i.d. measurements. Theorem 3.2 states that under the same assumption, those columns of C that have at least two non-zero entries are identifiable up to scaling and permutation indeterminacies (because scale and ordering of the components of Z are arbitrary). Theorem 3.3 shows that regardless of the noise distribution (i.e., also in the case of Gaussian noise), if there is no influence from X to Z and an additional genericity assumption holds, then B is identifiable from the covariance structure of X up to a small finite number of possibilities. In Propositions 3.1 and 3.2 we prove that the additional assumptions we just called generic do in fact only exclude a Lebesgue null set from the parameter space.

The second stage of our investigation, Section 3.7, is a first examination of how the above identifiability results can be translated into estimation algorithms operating on the finite sample $x_{1:L}$ of X. We propose two algorithms. Algorithm 1, which is tailored towards the conditions of Theorems 3.1 and 3.2, estimates B and C by approximately maximizing the likelihood of a parametric VAR model with a mixture of Gaussians as noise distribution. Algorithm 2, which is tailored towards the conditions of Theorem 3.3, estimates the

matrix B up to finitely many possibilities by solving a system of equations somewhat similar to the Yule-Walker equations [Lütkepohl, 2006]. Furthermore, we briefly examine how the model assumptions that we make can to some extent be checked just based on the observed sample of X. We examine the behavior of the two proposed algorithms on synthetic and real-world data.

Theorem 3.2, Theorem 3.3 as well as Algorithm 2 (and in a broader sense also Algorithm 1) can be seen as part of *approximate* causal inference, as defined Section 2.1.3.3, as they outline methods which cannot uniquely identify the correct causal model, just constrain the set of possible causal models.

3.1.3. Structure of this chapter

The remainder of this chapter is organized as follows:

- In Section 3.2 we discuss related work.
- In Section 3.3 we introduce notation and definitions for time series.
- In Section 3.4 we state the statistical and causal model that we assume throughout this chapter.
- In Section 3.5 we introduce the so-called generalized residual.
- Section 3.8 contains experiments for Algorithms 1 and 2.
- We conclude with Section 3.9.

3.2. Related Work

We discuss how the present chapter is related to previous work in similar directions. Generally it can be said, that time series with hidden confounders is a topic in causal inference, which, in spite of its relevance, has received rather little attention in the research community.

Granger causality: Probably the most widely applied approach to causal inference from time series data so far, which we refer to as *practical Granger causal analysis (or inference)* in this chapter (often just called "(linear) Granger causality"), is to simply perform a linear regression of X_t on X_{t-1}, based on the observed sample of X, and then interpret the regression matrix causally [Granger, 1969, Lütkepohl, 2006] (sometimes lags of length more than 1 are used as well). While this method may yield reasonable results in certain cases, it obviously can go wrong in others (see Section 3.4.3 for details). We believe that our approach may in certain cases lead to more valid causal conclusions.

Inference of properties of processes with hidden components: Jalali and Sanghavi [2012] also assume a VAR model with hidden components and try to identify parts of the transition matrix. However, their results are based on different assumptions: they assume a "local-global structure", i.e., connections between observed components are sparse and each latent series interacts with many observed components, to achieve identifiability. Boyen et al. [1999] - similar to us - apply a method based on expectation maximization (EM) to infer properties of partially observed Markov processes. Unlike us, they consider finite-state Markov processes and do not provide a theoretical analysis of conditions for identifiability. [Etesami et al., 2012] examine identifiability of partially observed processes that have a certain tree-structure, using so-called discrepancy measures. Gong et al. [2015] use similar assumptions as we do here to cope with a different type of hidden confounding in time series: the one that arises from subsampling. Entner and Hoyer [2010] transfer conditional-independence-based observational causal inference methods (see Section 2.1.3.2), that allow for hidden confounders, from the i.i.d. setting to the time series setting. Their method generally cannot identify the influence structure within the observed X when there are hidden confounders, while one of ours (based on Theorem 3.1) can. But this is due to their assumptions being much weaker, which makes their method applicable more broadly than ours.

Harnessing non-Gaussian noise for causal inference: Probably the most important work that uses the assumption of non-Gaussian noise for causal inference is by Shimizu et al. [2006], which considers only the a-temporal setting and does not address hidden confounders. Hyvaerinen et al. [2010] use non-Gaussian noise to infer instantaneous effects. Hoyer et al. [2008] use the theory underlying overcomplete independent component analysis (ICA) Kagan et al. [1973, Theorem 10.3.1] to derive identifiability – up to finitely many possibilities – of linear models with hidden variables, which is

somewhat similar to our Theorem 3.1. However, there are two major differences: First, they only consider models which consist of finitely many observables which are mixtures of finitely many noise variables. Therefore their results are not directly applicable to VAR models. Second, they show identifiability only up to a finite number of possibilities, while we (exploiting the autoregressive structure) prove unique identifiability. This can be interpreted as indicating, that temporal knowledge can even help for removing confounders.

Integrating several definitions of causation: Eichler [2012] provides an overview over various definitions of causation w.r.t. time series, somewhat similar to, but more comprehensive than our brief discussion in Sections 3.4.2 and 3.4.3.

3.3. Preliminaries: definitions and notation of time series

Here we introduce notation and definitions w.r.t. time series. We denote multivariate *time series*, i.e., families of random vectors over the index set $\mathbb{Z}$, by upper case letters such as X. As usual, X_t denotes the t-th member of X, and X_t^k denotes the k-th component of the random vector X_t. Slightly overloading terminology, we call the univariate time series $X^k = (X_t^k)_{t\in\mathbb{Z}}$ the *k-th component of X*. By P_X we denote the distribution of the random process X, i.e., the joint distribution of all X_t, $t \in \mathbb{Z}$.

Given a K_X-variate time series X and a K_Z-variate time series Z, $(X, Z)^\top$ denotes the $(K_X + K_Z)$-variate series

$$\left((X_t^1, \ldots, X_t^{K_X}, Z_t^1, \ldots, Z_t^{K_Z})^\top\right)_{t\in\mathbb{Z}}.$$

A K-variate time series W is a *vector autoregressive process (of order 1), or VAR process for short, with VAR transition matrix A and noise covariance matrix Σ*, if it allows a *VAR representation*, i.e.,

$$W_t = AW_{t-1} + N_t, \tag{3.1}$$

the absolute value of all eigenvalues of A is less than[1] 1, and N is an i.i.d. noise time series such that $\mathrm{cov}(N_0) = \Sigma$. We say W is a *diagonal-structural VAR process* if in the above definition the additional condition is met that $N_0^1, \ldots, N_0^K$ are jointly independent.[2]

3.4. Model assumptions

The question of inferability of the causal structure underlying the given sample $x_{1:L}$ can be reduced to the question of identifiability of parameters of the statistical model for $x_{1:L}$. Therefore, in this section we first introduce the statistical model that we consider throughout this chapter. Then we discuss in some detail why the reduction is valid. Last, we demonstrate, based on the model, how practical Granger causal inference can go wrong.

Note that maybe our strongest assumption is that of linearity and it is left to future work to what extent the results in this work can be extended to nonlinear settings.

3.4.1. Statistical model

Let K_X be arbitrary but fixed and let X be a K_X-variate time series. As stated in Section 3.1, X is the random process from which we assume we measured the given finite sample $x_{1:L}$. In particular, the random variables in X have a meaning in reality (e.g., X_3^1 is the temperature measured in room 1 at time 3) and we are interested in the causal relations between these variables. Let X be related to a K-variate VAR process W, with transition matrix A, noise time series N, and noise covariance matrix Σ, and a K_Z-variate time series Z, as follows: $W = (X, Z)^\top$ and $K_Z \leq K_X$. Furthermore, let

$$A =: \begin{pmatrix} B & C \\ D & E \end{pmatrix}, \tag{3.2}$$

[1] We require all VAR processes to be stable [Lütkepohl, 2006].

[2] Note that the notion "diagonal-structural" is a special case of the more general notion of "structural" in, e.g., [Lütkepohl, 2006].

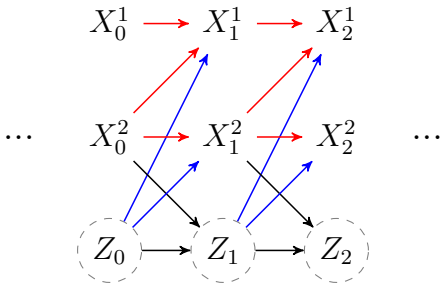


Figure 3.2.: Example of a causal DAG for a bivariate observed time series $X = (X_t)_{t\in\mathbb{Z}}$ and univariate hidden confounder time series $Z = (Z_t)_{t\in\mathbb{Z}}$ that could be induced by the (structural) transition matrix A in Equation 3.1. Keep in mind Equation 3.2. Non-zero entries of B correspond to red arrows, non-zero entries of C correspond to blue arrows, and the remaining non-zero entries correspond to black arrows.

with B a $K_X \times K_X$ matrix. We call B, the most interesting part of A, the *structural matrix underlying* X, as we will argue below that it captures the influences between the components of X. Furthermore, in case $C \neq 0$, we call Z a *hidden confounder*. (Although, rigorously speaking, if $C \neq 0$ but D is such that no Z_s^i influences more than one X_t^j, then no variable in Z would be a proper confounder according to Definition 2.5.)

3.4.2. Underlying causal model

Throughout this chapter we assume that there is an underlying system such that all variables in W correspond to actual properties of that system which are in principle measurable and intervenable. While we assume that a finite part of X, namely $X_{1:L}$, was in fact measured (Section 3.4.1), Z is completely unmeasured. Furthermore we assume that the entries of A, in particular the submatrix B, capture the actual non-instantaneous causal influences between the variables in W. This latter assumption can be justified using FCMs, and we will present this justification below, but it can also be justified using an argument by Granger [1969], which we will present afterwards.

Let us assume that W forms a causally sufficient set of variables. Furthermore, let us assume that there are no influences from present to present or present to past. Additionally, assume linearity and that influences are not beyond one time step. Together, this means we assume a FCM given by the VAR equations (3.1), with causes being on

the r.h.s. and effects on the l.h.s.[3] (In particular, these equations induce the correct (temporal) causal DAG for $(X, Z)^\top$, for which we give an example in Figure 3.2.) But this means that A does in fact capture the actual non-instantaneous causal influences between the variables in W.[4] This is one way to justify our approach (in case the requirement $K_Z \leq K_X$ and the other assumptions are met).

On the one hand, Granger [1969] proposed a definition of causation between observables which we will refer to as *Granger's ideal definition*. Assume the statistical model for the observed sample of X specified in Section 3.4.1. If we additionally assume that Z correctly models the whole rest of the universe or the "relevant" subpart of it, then according to Granger's ideal definition the non-instantaneous (direct) causal influences between the components of X are precisely given by the entries of B. But this implies that everything about B that we can infer from X can be interpreted causally, if one accepts Granger's ideal definition and the additional assumptions that are necessary (such as $K_Z \leq K_X$, which in fact may be a quite strong assumption of course). This is the other way to justify our approach.

3.4.3. How practical Granger causal inference can go wrong

The above ideal definition of causation by Granger (Section 3.4.2) needs to be contrasted with what we introduced as "practical Granger causal analysis" in Section 3.1. In practical Granger causal analysis, one just performs a linear regression of present on past on the observed X and then interprets the regression matrix causally.[5] (Often, a statistical test is applied for the null hypothesis that the respective entries in the regression matrix are 0 [Lütkepohl, 2006].) While making the ideal definition practically feasible, this may lead to wrong causal conclusions in the sense that it does not comply with the causal

[3] Note that here we ignore the fact that Pearl [2000] generally only considers models with finitely many variables while the process W is a family of infinitely many (real-valued) variables.

[4] It may be more meaningful to say that A captures the actual influences *to the extent it is identifiable from what we are given*. For simplicity we chose the other formulation though.

[5] We are aware that nonlinear models Chu and Glymour [2008] and nonparametric estimators Schreiber [2000a] have been used to find temporal causal relations. In this chapter we focus on the linear case. Also note that practical Granger causal inference is often also used with lag length higher than 1, while here we restrict to VAR processes of order 1.

structure that we would infer given we had more information.[6]

Let us give an example for this:

Example 3.1. *Let X be bivariate and Z be univariate. Moreover, assume*

$$A = \left(\begin{array}{cc|c} 0.9 & 0 & 0.5 \\ 0.1 & 0.1 & 0.8 \\ \hline 0 & 0 & 0.9 \end{array}\right),$$

and let the covariance matrix of N_t be the identity matrix. To perform practical Granger causal analysis, we proceed as usual: we fit a VAR model on only *X, in particular compute, w.l.o.g. assuming zero mean, the transition matrix by*

$$B_{pG} := \mathbb{E}(X_t X_{t-1}^\top)\mathbb{E}(X_t X_t^\top)^{-1} = \begin{pmatrix} 0.89 & 0.35 \\ 0.08 & 0.65 \end{pmatrix} \tag{3.3}$$

(up to rounding) and interpret the coefficients of B_{pG} as causal influences. Although, based on A, X_t^2 does in fact not influence X_{t+1}^1, B_{pG} suggests that there is a strong causal effect $X_t^2 \to X_{t+1}^1$ with the strength 0.35*. It is even stronger than the relation $X_t^1 \to X_{t+1}^2$, which actually exists in the complete model with the strength* 0.1*.*

3.5. The Generalized Residual: Definition and Properties

In this section we define the generalized residual and discuss some of its properties. The generalized residual is used in the proofs of the three main results of this chapter, Theorems 3.1 to 3.3.

For any $K_X \times K_X$ matrices U_1, U_2 let

$$R_t(U_1, U_2) := X_t - U_1 X_{t-1} - U_2 X_{t-2}.$$

[6] Obviously, if one is willing to assume that X is causally sufficient already, then the practical Granger causation can be justified along the lines of Section 3.4.2. For a more detailed discussion, we refer the reader to [Peters et al., 2017, Chapter 10].

We call this family of random vectors *generalized residual*. Furthermore let

$$M_1 := \mathbb{E}\left[W_t \cdot (X_t^\top, X_{t-1}^\top)\right].$$

In what follows, we list some simple properties of the generalized residual. Proofs can be found in Section A.1.

Lemma 3.1. *We have*

$$\begin{aligned} R_t(U_1, U_2) = &(B^2 + CD - U_1 B - U_2) X_{t-2} \\ &+ (BC + CE - U_1 C) Z_{t-2} \\ &+ (B - U_1) N_{t-1}^X + C N_{t-1}^Z + N_t^X, \end{aligned} \tag{3.4}$$

if $K > K_X$. In case $K = K_X$, the same equation holds except that one sets $C := D := E := 0$.

Lemma 3.2. *If (U_1, U_2) satisfies the equation*

$$(U_1, U_2) \begin{pmatrix} B & C \\ \mathbf{I} & 0 \end{pmatrix} = \left(B^2 + CD, BC + CE\right), \tag{3.5}$$

where $\mathbf{I}$ denotes the identity matrix, then $R_t(U_1, U_2)$ is independent of $(X_{t-2-j})_{j=0}^\infty$, and in particular, for $j \geq 0$,

$$\operatorname{cov}(R_t(U_1, U_2), X_{t-2-j}) = 0. \tag{3.6}$$

Let $\Gamma_i^X := \operatorname{cov}(X_t, X_{t-i})$ for all i. That is, Γ_i^X are the *autocovariance matrices* of X. Note that Equation (3.6), for $j = 0, 1$, can equivalently be written as the single equation

$$(U_1, U_2) \begin{pmatrix} \Gamma_1^X & \Gamma_2^X \\ \Gamma_0^X & \Gamma_1^X \end{pmatrix} = \left(\Gamma_2^X, \Gamma_3^X\right). \tag{3.7}$$

Keep in mind that, as usual, we say a $m \times n$ matrix has *full rank* if its (row and column) rank equals $\min\{m, n\}$.

Lemma 3.3. *Let M_1 have full rank. If (U_1, U_2) satisfies Equation (3.6) for $j = 0, 1$, then it satisfies Equation (3.5).*

Lemma 3.4. *If $K = K_X$ or if C has full rank, then there exists (U_1, U_2) that satisfies Equation (3.5).*

3.6. Theorems on identifiability and almost identifiability

This section contains the main results of the present chapter. We present three theorems on unique and approximate identifiability of B and C (defined in Section 3.4.1), respectively, given X, and briefly argue why certain assumptions we have to make can be considered as generic. Recall the definition of the matrix M_1 in Section 3.5. Note that the following results show (approximate) identifiability of B for all numbers K_Z of hidden components *simultaneously*, as long as $0 \leq K_Z \leq K_X$ (which contains the case of no hidden components as a special case).

3.6.1. Assuming non-Gaussian, independent noise

We will need the following assumptions for the theorems.

Assumptions. *We define the following abbreviations for the respective subsequent assumptions.*

A1*: All noise terms N_t^k, $k = 1, \ldots, K, t \in \mathbb{Z}$, are non-Gaussian.*

A2*: W is a diagonal-structural VAR process (as defined in Section 3.3).*

G1*: C (if it is defined, i.e., if $K > K_X$) and M_1 have full rank.*

(We will discuss the genericity of G1 in Section 3.6.3.)

The following definition of F_1 is not necessary for an intuitive understanding, but is needed for a precise formulation of the subsequent identifiability statements. Let F_1 denote the set of all K'-variate VAR processes W' with $K_X \leq K' \leq 2K_X$ (i.e. W has at most as many hidden components as observed ones), which satisfy the following

properties w.r.t. N', C', M_1' (defined similarly to N, C, M_1 in Section 3.4): assumptions A1, A2 and G1 applied to N', C', M_1' (instead of N, C, M_1) hold true.

Theorem 3.1. *If assumptions A1, A2 and G1 hold true, then B is uniquely identifiable from only P_X.*

That is: There is a map f such that for each $W' \in F_1$, and X' defined as the first K_X components of W', $f(P_{X'}) = B'$ iff B' is the structural matrix underlying X'.

A detailed proof can be found in Section A.2.1. The idea is to chose U_1, U_2 such that $R_t(U_1, U_2)$ is a linear mixture of only *finitely* many noise terms, which is possible based on Lemmas 3.1 to 3.4. Then, using the identifiability result underlying overcomplete ICA [Kagan et al., 1973, Theorem 10.3.1], the structure of the mixing matrix of $(R_t(U_1, U_2), R_{t-1}(U_1, U_2))^\top$ allows to uniquely determine B from it. Note that this is the only result on unique identifiability, while the next two results only guarantee approximate identifiability.

Again using [Kagan et al., 1973, Theorem 10.3.1], one can also show the following result. For a matrix M let $S(M)$ denote the set of those columns of M that have at least two non-zero entries, and if M is not defined, let $S(M)$ denote the empty set. A proof can be found in Section A.2.2.

Theorem 3.2. *If assumptions A1, A2 and G1 hold true, then the set of columns of C with at least two non-zero entries is identifiable from only P_X up to scaling of those columns.*

In other words: There is a map f such that for each $W' \in F_1$ with K' components, X' defined as the first K_X components of W', and C' defined as the upper right $K_X \times (K' - K_X)$ submatrix of the transition matrix of W', $f(P_{X'})$ coincides with $S(C')$ up to scaling of its elements.

3.6.2. Assuming no influence from observed to hidden components

In this section we present a theorem on the approximate identifiability of B under different assumptions. In particular, we drop the non-Gaussianity assumption. Instead, we make the assumption that Z is not influenced by X, i.e., $D = 0$.

Given $U = (U_1, U_2)$, let

$$T_U(Q) := Q^2 - U_1 Q - U_2, \tag{3.8}$$

for all square matrices Q that have the same dimension as U_1. Slightly overloading notation, we let $T_U(\alpha) := T_U(\alpha\,\mathbf{I})$ for all scalars α. Note that $\det(T_U(\alpha))$ is a univariate polynomial in α.

We will need the following assumptions for the theorem.

Assumptions. *We define the following abbreviations for the respective subsequent assumptions.*

A3*:* $D = 0$*.*

G2*: The transition matrix A is such that there exists $U = (U_1, U_2)$ such that Equation (3.5) is satisfied and $\det(T_U(\alpha))$ has $2K_X$ distinct roots.*

(We will discuss the genericity of G2 in Section 3.6.3.)

The following definition of F_2 is not necessary for an intuitive understanding, but is needed for a precise formulation of the subsequent identifiability statement. Let F_2 denote the set of all K'-variate VAR processes W' with $K_X \leq K' \leq 2K_X$, which satisfy the following properties w.r.t. N', A', C', D', M_1' (defined similarly to N, A, C, D, M_1 in Section 3.4): assumptions A3, G1 and G2 applied to N', A', C', D', M_1' (instead of N, A, C, D, M_1) hold true.

Theorem 3.3. *If assumptions A3, G1 and G2 hold true, then B is identifiable from only the covariance structure of X up to $\binom{2K_X}{K_X}$ possibilities.*

In other words: There is a map f such that for each $W' \in F_2$, and X' defined as the first K_X components of W', $f(X')$ is a set of at most $\binom{2K_X}{K_X}$ many matrices, and $B' \in f(P_{X'})$ for B' the structural matrix underlying X'.

A detailed proof can be found in Section A.2.3. The proof idea is the following: Let L denote the set of all $(U, \tilde{B})$, with $U = (U_1, U_2)$, that satisfy Equation (3.6) for $j = 0, 1$,

as well as the equation

$$T_U(\tilde{B}) = 0, \tag{3.9}$$

and meet the condition that $\det(T_U(\alpha))$ has $2K_X$ distinct roots. L is non-empty and (U, B) is an element of it, for the true B and some U, due to Lemmas 3.2 to 3.4. But L is only defined based on the covariance of X and has at most $\binom{2K_X}{K_X}$ elements (based on [J. E. Dennis et al., 1976]).

Note the similarity between Equation (3.6), or its equivalent, Equation (3.7), and the well-known Yule-Walker equation [Lütkepohl, 2006]. The Yule-Walker Equation (which is implicitly used in Equation (3.3)) determines B uniquely under some genericity assumption and given $C = 0$.

3.6.3. Discussion of the genericity assumptions

In this section we want to briefly argue why the assumptions G1 and G2 are generic. A detailed elaboration with precise definitions and proofs can be found in Section A.3. The idea is to define a natural parametrization of (A, Σ) and to show that the restrictions that assumptions G1 and G2, respectively, impose on (A, Σ) just exclude a Lebesgue null set in the natural parameter space and thus can be considered as generic.

In this section, let K such that $K_X \leq K \leq 2K_X$ be arbitrary but fixed. Let λ_k denote the k-dimensional Lebesgue measure on $\mathbb{R}^k$.

Let Θ_1 denote the set of all possible parameters (A', Σ') for a K-variate VAR processes W' that additionally satisfy assumption A2, i.e., correspond to structural W'. Let S_1 denote the subset of those $(A', \Sigma') \in \Theta_1$ for which also assumption G1 is satisfied. And let g denote the natural parametrization of Θ_1 which is defined in Section A.3.1.

Proposition 3.1. *We have $\lambda_{K^2+K}(g^{-1}(\Theta_1 \setminus S_1)) = 0$. That is, assumption G1 only excludes a set of parameters of Lebesgue measure 0 from the set of parameters that satisfy (A1 and) A2.*

A proof can be found in Section A.3.1. The proof idea is that $g^{-1}(\Theta_1 \setminus S_1)$ is essentially contained in the union of the root sets of finitely many multivariate polynomials and hence is a Lebesgue null set.

Let Θ_2 denote the set of all possible parameters (A', Σ') for the K-variate VAR processes W that additionally satisfy assumption A3, i.e., are such that the submatrix D of A is zero. Let S_2 denote the subset of those $(A', \Sigma') \in \Theta_2$ for which also assumptions G1 and G2 are satisfied. Let h denote the natural parametrization of Θ_1 which is defined in Section A.3.2. A proof for the following proposition (which is based on a similar idea as that of Proposition 3.1) can also be found in Section A.3.2.

Proposition 3.2. *We have* $\lambda_{2K^2-K_XK_Z}\left(h^{-1}(\Theta_2 \setminus S_2)\right) = 0$. *That is, assumptions G1 and G2 jointly only exclude a set of parameters of Lebesgue measure 0 from the set of parameters that satisfy A3.*

3.7. Estimation algorithms

In this section we examine how the identifiability results in Section 3.6 can be translated into estimators on finite data. We propose two algorithms.

3.7.1. Algorithm based on variational expectation maximization

Here we present an algorithm for estimating B and C which is closely related to Theorems 3.1 and 3.2. Keep in mind that the latter theorem in fact only states identifiability for $S(C)$ (the set of those columns of C that have at least two non-zero entries, defined in Section 3.6.2), up to scaling, not for the exact C. The idea is the following: We transform the model of X underlying these theorems (i.e. the general model from Section 3.4.1 together with assumptions A1, A2 and G1 from Section 3.6.1) into a parametric model by assuming the noise terms N_t^k to be mixtures of Gaussians.[7] Then we estimate all parameters, including B and C, by approximately maximizing the likelihood of the

[7] Obviously, Theorems 3.1 and 3.2 also imply identifiability of B and (up to scaling) $S(C)$ for this parametric model. We conjecture that this implies consistency of the (non-approximate) maximum likelihood estimator for that model under appropriate assumptions.

Algorithm 1 Estimate B, C using variational EM

1: **Input:** Sample $x_{1:L}$ of $X_{1:L}$.
2: Initialize the transition matrix and the parameters of the Gaussian mixture model, denoted as θ^0, set $j \leftarrow 0$.
3: **repeat**
4: **E step:** Evaluate

$$q^j(z_{1:L}, v^X_{1:L}, v^Z_{1:L}) = q^j(z_{1:L})q^j(v^X_{1:L})q^j(v^Z_{1:L}),$$

which is the variational approx. to the true posterior $q^j(z_{1:L}, v^X_{1:L}, v^Z_{1:L}|x_{1:L})$, by maximizing the variational lower bound, i.e., $q^j = \arg\max_q \mathcal{L}(q, \theta^j)$.
5: **M step:** Evaluate $\theta^{j+1} = \arg\max_\theta \mathcal{L}(q^j, \theta)$.
6: $j \leftarrow j + 1$.
7: **until** convergence
8: **Output:** The final θ^j, containing the estimated B, C.

Algorithm 2 Estimate B using covariance structure

1: **Input:** Sample $x_{1:L}$ of $X_{1:L}$.
2: Solve the linear Equation (3.7), with Γ^X_i replaced by $\hat{\Gamma}^X_i$. Let $(\hat{U}_1, \hat{U}_2)$ denote the solution.
3: Solve Equation (3.9) with $U := (\hat{U}_1, \hat{U}_2)$ for $\tilde{B}$. Let $\hat{B}_1, \ldots, \hat{B}_n$ denote the solvents.
4: **Output:** $\hat{B}_1, \ldots, \hat{B}_n$.

given sample of X using a variational expectation maximization (EM) approach similar to the one in [Oh et al., 2005]. (Directly maximizing the likelihood is intractable due to the hidden variables (Z and mixture components) that have to be marginalized out.) Let $y_{1:L}$ be shorthand for $(y_1, \ldots, y_L)$. The estimator is outlined by Algorithm 1, where (V^X_t, V^Z_t) with values (v^X_t, v^Z_t) denote the vectors of mixture components for N^X_t and N^Z_t, respectively; $q^j(z_{1:L}, v^X_{1:L}, v^Z_{1:L}|x_{1:L})$ the true posterior of $Z_{1:L}, V^X_{1:L}, V^Z_{1:L}$ under the respective parameter vector θ^j (which comprises A, Σ as well as the Gaussian mixture parameters) at step j; and $\mathcal{L}$ the variational lower bound. The detailed algorithm can be found in Section A.4. Note that, if needed, one may use cross validation as a heuristic to determine K_Z and the number of Gaussian mixture components.

3.7.2. Algorithm based on the covariance structure

Now we present an algorithm, closely related to Theorem 3.3, for estimating B up to

finitely many possibilities. It relies on the proof idea of that theorem, as we outlined it at the end of Section 3.6.2, and it is meant to be applied for cases where the conditions of that theorem are met. It uses only the estimated autocovariance structure of X. Keep in mind that $\hat{\Gamma}_i^X$ denote the sample autocovariance matrices (similar to the true autocovariances Γ_i^X defined in Section 3.5). The estimation algorithm is given by Algorithm 2.

3.7.3. Model checking

Ideally we would like to know whether the various model assumptions we make in this chapter, most importantly the one that the entries of B can in fact be interpreted causally, are appropriate. Obviously, this is impossible to answer just based on the observed sample of X. Nonetheless one can check these assumptions to the extent they imply testable properties of X.

For instance, to check (to a limited extent) the assumptions underlying Theorems 3.1 and 3.2 and Algorithm 1, i.e., the general statistical and causal model assumptions from Sections 3.4.1 and 3.4.2 together with A1, A2 and G1 from Section 3.6.1, we propose the following two tests: First, test whether $R_t(\hat{U}_1, \hat{U}_2)$ is independent of $(X_{t-2-j})_{j=0}^J$, for $(\hat{U}_1, \hat{U}_2)$ as defined in Algorithm 2, and for say $J = 2$. (If Algorithm 2 finds no $(\hat{U}_1, \hat{U}_2)$ then the test is already failed.) Second, check whether all components of X_t are non-Gaussian using e.g. the Kolmogorov-Smirnov test [Conover, 1971] for Gaussianity.

Note that under the mentioned assumptions, both properties of X do in fact hold true. Regarding the independence statement, this follows from Lemmas 3.4 and 3.2. W.r.t. the non-Gaussianity statement, this follows from the fact [Ramachandran, 1967, Theorem 7.8] that the distribution of an infinite weighted sum of non-Gaussian random variables is again non-Gaussian. It should be mentioned that the first test can also be used to check (to a limited extent) the assumptions underlying Theorem 3.3 and Algorithm 2.

3.8. Experiments

In this section we evaluate the two algorithms proposed in Section 3.7 on synthetic and real-world data and compare them to the practical Granger causation estimator. Keep in mind that the latter is defined by replacing the covariances in Equation (3.3) by sample covariances (after centering).

3.8.1. Synthetic data

We empirically study the behavior of Algorithms 1 and 2 on simulated data, in dependence on the sample length. Note that, based on theoretical considerations (see Section 3.4.3), it can be expected that the error of the practical Granger estimator is substantially bounded away from zero in the generic case, even when the sample size goes to infinity.

3.8.1.1. Algorithm 1

Here we evaluate Algorithm 1.

Experimental setup: We consider the case of a 2-variate X and a 1-variate Z, i.e., $K_X = 2, K_Z = 1$. We use sample lengths $L = 100, 500, 1000, 5000$ and for each sample length we do 20 runs. In each run we draw the matrix A uniformly at random from the stable matrices (i.e., the absolute value of all eigenvalues of A is less than 1) and then randomly draw a sample of length L from a VAR process $W = (X, Z)^\top$ with A as transition matrix and noise N_t^k distributed according to a super-Gaussian mixtures of Gaussians. Then we apply Algorithm 1 and the practical Granger causation estimator on the sample of *only* X.

Outcome: We calculated the root-mean-square error (RMSE) of Algorithm 1, i.e., $\frac{1}{20}\sum_{n=1}^{20}(B_n^{\text{est}} - B_n^{\text{true}})^2$, where $B_n^{\text{est}}, B_n^{\text{true}}$ denotes the output of Algorithm 1 and the true B, respectively, for each run n. The RMSE as a function of the sample length L is depicted in Figure 3.3, along with the RMSE of the practical Granger algorithm.

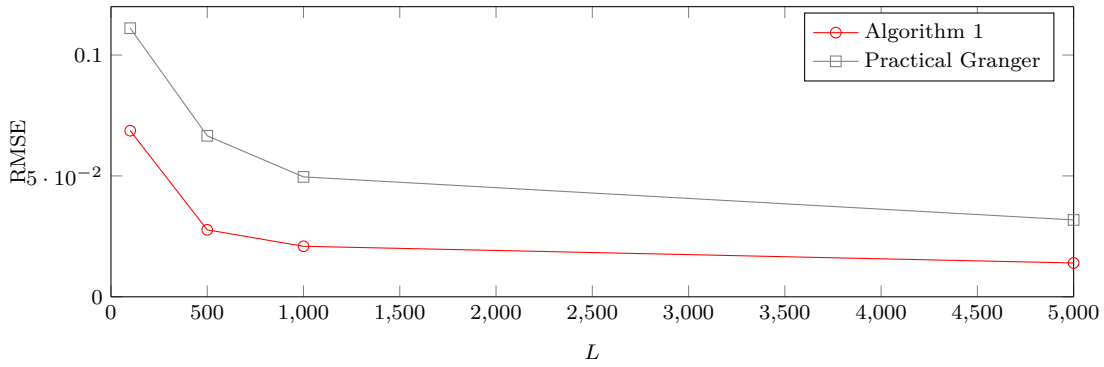


Figure 3.3.: RMSE of Algorithm 1 and the practical Granger estimator as a function of sample length L.

Discussion: This suggests that for $L \to \infty$ the error of Algorithm 1 is negligible, although it may not converge to zero. The reason for it not converging to zero is that we use variational EM which can yield a systematically wrong estimate whenever the approximative assumptions it relies on do not hold. The error of the practical Granger estimator for $L \to \infty$ is still small but substantially bigger than that of Algorithm 1.

3.8.1.2. Algorithm 2

Here we empirically establish the error of Algorithm 2, more precisely the deviation between the true B and the best out of the several estimates that Algorithm 2 outputs (recall that Theorem 3.3 only guarantees identifiability up to finitely many possibilities). Obviously in general it is unknown which of the outputs of Algorithm 2 is the best estimate. However here we rather want to establish that asymptotically, the output of Algorithm 2 in fact contains the true B. Also we compare Algorithm 2 to the practical Granger estimator, although it needs to be said, that the latter is usually not applied to univariate time series.

Experimental setup: We consider the case of 1-variate X and Z, i.e., $K_X = K_Z = 1$. We consider sample lengths $L = 10^1, 10^2, \ldots, 10^7$ and for each sample length we do 20 runs. In each run we draw the matrix A uniformly at random from the stable matrices with the constraint that the lower left entry is zero and then randomly draw a sample of length L from a VAR process $W = (X, Z)^\top$ with A as transition matrix and standard

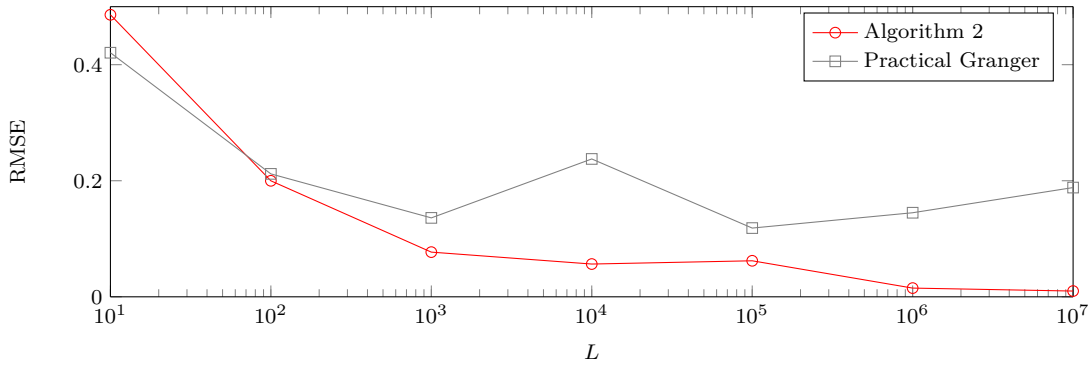


Figure 3.4.: RMSE of Algorithm 2 and the practical Granger estimator as a function of sample length L.

normally distributed noise N. Then we apply Algorithm 2 and the practical Granger causation estimator on the sample of only X.

Outcome: We calculated the root-mean-square error (RMSE) of Algorithm 2, i.e., $\frac{1}{20}\sum_{n=1}^{20}(B_n^{\text{best est}} - B_n^{\text{true}})^2$, where $B_n^{\text{best est}}, B_n^{\text{true}}$ denotes the best estimate for B returned by Algorithm 2 (i.e., the one out of the two outputs that minimizes the RMSE) and true B for each run n, respectively. The RMSE as a function of the sample length L is depicted in Figure 3.4, along with the RMSE of the practical Granger estimator.

Discussion: This empirically shows that the set of two outputs of Algorithm 2 asymptotically seem to contain the true B. However, it takes at least 1000 samples to output reasonable estimates. As expected, the practical Granger estimator does not seem to converge against the true B.

3.8.2. Real-world data

Here we examine how Algorithm 1 performs on a real-world data set.

Experimental setup: We consider a time series Y of length 340 and the three components: cheese price Y^1, butter price Y^2, milk price Y^3, recorded monthly from January 1986 to April 2014[8]. We used the following estimators: We applied practical Granger estimation to the full time series Y (i.e., considering $X = Y$) and denote the outcome by

[8]The data was retrieved from `http://future.aae.wisc.edu/tab/prices.html` on 29.05.2014.

A_{fG}. We applied practical Granger estimation to the reduced time series $(Y^1, Y^2)^\top$ (i.e., considering $X = (Y^1, Y^2)^\top$) and denote the outcome by B_{pG}. We applied Algorithm 1 to the full time series Y (i.e., considering $X = Y$), while assuming an additional hidden univariate Z, and denote the outcome by $\bar{A}_{\mathrm{fA}}$. We applied Algorithm 1 to the reduced time series $(Y^1, Y^2)^\top$ (i.e., considering $X = (Y^1, Y^2)^\top$), while assuming an additional hidden univariate Z, and denote the outcome by $\tilde{A}_{\mathrm{pA}}$. Furthermore we do a model check as suggested in Section 3.7.3, although the sample size may be too small for the independence test to work reliably.

Note that causal inference is particularly relevant in such economic settings. For instance, a policy maker may wonder about the effect of a change in the regulation of milk prices on the price of other groceries such as cheese and butter.

Outcome: The outputs are:

$$A_{\mathrm{fG}} = \begin{pmatrix} 0.8381 & 0.0810 & 0.0375 \\ 0.0184 & 0.9592 & -0.0473 \\ 0.2318 & 0.0522 & 0.7446 \end{pmatrix},$$

$$B_{\mathrm{pG}} = \begin{pmatrix} 0.8707 & 0.0837 \\ -0.0227 & 0.9559 \end{pmatrix},$$

$$\bar{A}_{\mathrm{fA}} = \begin{pmatrix} 0.8809 & 0.1812 & 0.1016 & -0.1595 \\ 0.0221 & 1.0142 & -0.0290 & -0.0492 \\ 0.2296 & 0.1291 & 0.8172 & -0.1143 \\ 1.0761 & 0.6029 & -0.7184 & 0.4226 \end{pmatrix},$$

$$\tilde{A}_{\mathrm{pA}} = \begin{pmatrix} 0.9166 & 0.0513 & -0.0067 \\ -0.0094 & 0.9828 & -0.0047 \\ -0.0031 & 0.1441 & -0.2365 \end{pmatrix}.$$

The outcome of the model check, based on a significance level of 5%, is the following: the hypothesis of Gaussianity is rejected. Also the independence hypothesis stated in Section 3.7.3 is rejected. The latter implies that the model assumptions underlying Algorithm 1 are probably wrong.

Discussion: We consider A_{fG} as ground truth. Intuitively, non-zero entries at positions $(i, 3)$ can be explained by the milk price influencing cheese/butter prices via production

costs, while non-zero entries at positions $(3, j)$ can be explained by cheese/butter prices driving the milk price via demand for milk. The explanation of non-zero entries at positions (1, 2) an (2, 1) is less clear. One can see that the upper left 2×2 submatrix of $\tilde{A}_{\mathrm{pA}}$ is quite close to that of A_{fG} (the RMSE over all entries is 0.0753), which shows that Algorithm 1 works well in this respect. Note that B_{pG} is even a bit closer (the RMSE is 0.0662). However, the upper right 2×1 matrix of $\tilde{A}_{\mathrm{pA}}$ is not close to a scaled version of the upper right 2×1 submatrix of A_{fG} (which corresponds to C). This is in contrast to what one could expect based on Theorem 3.2. $\bar{A}_{\mathrm{fA}}$ can be seen as an alternative ground truth. It is important to mention that the estimated order (lag length) of the full time series Y is 3, according to Schwarz's criterion (SC) [Lütkepohl, 2006], which would violate our assumption of a VAR process of order 1 (Section 3.4.1). The model check seems to detect this violation of the model assumptions.

3.9. Conclusions of this chapter

One of the main insights from this chapter is that while the problem of hidden confounding cannot be solved by temporal knowledge, its severity can nonetheless be weakened: When assuming linearity and non-Gaussian noise in *a-temporal (i.i.d.)* settings, then influences between the observed variables are only identifiable up to a finite number of possibilities [Hoyer et al., 2008] (to the best of our knowledge). In contrast, in Theorem 3.1 we showed that, under weak additional assumptions, in *time series*, linearity and non-Gaussianity of noise are sufficient to *uniquely* identify the influences between observed variables. This may indicate that also in other regards identifiability of the causal model can be improved by temporal knowledge – which is often cheap to obtain.

It is important to note that, while we presented concrete estimation algorithms, our analysis of identifiability does not only apply to these algorithms, but can be useful for other estimation algorithms (for the assumed scenario) as well.

Chapter 4.

Approximate causal inference by bounding confounding in i.i.d. settings

4.1. Introduction

After Chapter 3, this is the second chapter that focuses on causal inference, while it already contains an example of a decision making application, as illustrated in Figure 4.1 on page 80.

In Chapter 3, we integrated temporal knowledge for causal inference, which determines the causal ordering, up to instantaneous effects. In this chapter, we again assume the causal ordering as given, may it be based on temporal or other information. But additionally, we assume the availability of knowledge that implies bounds on hidden confounding.

The investigation is driven by the following thought: Randomized experiments constitute the gold standard for causal inference – are often expensive, unethical or impossible to perform, as pointed out in Chapter 2. The main reason for randomization is to prevent hidden confounding. But even if we do not have "perfect" randomization, which would completely prevent hidden confounding, the underlying setting may contain mechanisms that at least imply bounds on hidden confounding (we illustrated the problem of hidden confounding in Example 2.3). This thought also underlies so-called quasi-experimental designs, which we will summarize in Section 4.2. In contrast to most work in the area

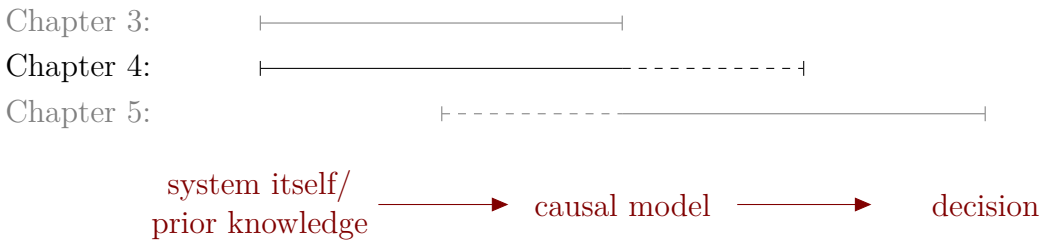


Figure 4.1.: The content of this chapter illustrated in black, relative to the rest of this thesis in gray, and the overall "inference path" in red.

of quasi-experimental designs though, our approach is based on the language of PCMs, which allows a rigorous formalization of, and reasoning about, the various scenarios we investigate.

So besides integration of additional knowledge, similar to Chapter 3, hidden confounding will play a central role here. And even more than in Chapter 3, the focus will be on *approximate* results, yielding constraints on the correct causal model instead of unique identifiability (as indicated in Section 2.1.3.3). Two fundamental differences to Chapter 3 are that here we do not consider time series, but rather i.i.d. settings, i.e., settings where the individual measurements are assumed to be *independent* samples (of the same distribution); and we consider a broader class of models than in Chapter 3, where we restricted ourselves to linear ones.

Parts of this chapter are based on the publication [Geiger et al., 2014].

4.1.1. Problem statement

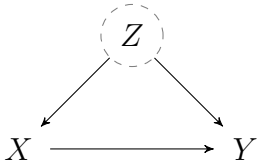


Figure 4.2.: Causal DAG for the hidden confounding scenario (gray means hidden).

We assume that we observe the variables X, Y and that Y does not influence X. That is, we assume the DAG in Figure 4.2 to be the correct causal DAG for X, Y, with Z capturing all common causes of X and Y. We assume Z to be partially or completely unobserved. Note that we allow Z, and in some cases also X, Y, to be multivariate. That is, we basically make no assumptions at this stage except that Y does not influence X, since Z can comprise arbitrarily large parts of the "rest of the universe".

Our general goal is to estimate the causal effect from X, which we also refer to as treatment variable, to Y, which we also refer to as outcome variable. Formally, this means that we want to estimate $P(Y|\mathrm{do}\, X{=}x)$ or related quantities such as the effect of treatment on the treated (ETT) [Pearl, 2000] or the causal strength from X to Y [Janzing et al., 2013]. Without further assumptions, these quantities are impossible to estimate. To give an extreme example, one can imagine observing the deterministic relationship $p(y|x) = \delta_{yx}$, with δ_{yx} denoting the Kronecker delta. This observation can be induced by two completely different underlying causal structures, the first one being that Y in fact is produced by copying X, the second one being that both X and Y are copied from Z without X having any causal effect on Y.

4.1.2. Outline of our approach

The approach we suggest to approximately infer the causal effect of X on Y in spite of hidden confounding consists of two parts: In the *first part* – Section 4.4 – we propose various possibilities to formalize the following notions:

- *Observed dependence:* the dependence of Y on X that we can observe based on $P(X, Y)$.
- *Back-door dependence:* the "spurious association" [Pearl, 2000] between X and Y due to the confounder Z.
- *Causal effect:* the causal effect of X on Y, as defined in Definition 2.6, but also notions of conditional causal effect such as the ETT.

Table 4.1.: Formalizing observed dependence, back-door dependence, causal effect and deviation measure.

Section	Notion	Formalized by ...
4.4.1	observed dependence	$\mathrm{I}(X:Y)$
	back-door dependence	$\mathrm{I}(Z:X)$, $\mathfrak{C}_{Z\to X}$
	causal effect	$\mathfrak{C}_{X\to Y}$
	deviation measure	difference
4.4.2	observed dependence	$\mathrm{I}(X:Y)$
	back-door dependence	$\mathrm{I}(Z:X)$
	causal effect	$\mathrm{I}(X\to Y\|\mathrm{do}\,Z)$
	deviation measure	difference
4.4.3	observed dependence	$p(Y\|X{=}x)$
	back-door dependence	$\mathrm{I}(X:Z)$, $\min\{\mathfrak{C}_{Z\to X},\mathfrak{C}_{Z\to Y}\}$
	causal effect	$p(Y\|\mathrm{do}\,X{=}x)$
	deviation measure	$\mathrm{D}(\cdot\|\cdot)$
4.4.4	observed dependence	$\mathbb{E}(\mathrm{d}_x \log p(Y\|X{=}x)^2)$
	back-door dependence	$\mathbb{E}((\partial_2 \log p(Y\|X{=}x,\mathrm{do}\,X{=}x))^2)$
	causal effect	$\mathbb{E}((\partial_1 \log p(Y\|X{=}x,\mathrm{do}\,X{=}x))^2)$
	deviation measure	difference
4.4.5	observed dependence	$\mathbb{E}(Y\|X{=}x') - \mathbb{E}(Y\|X{=}x)$
	back-door dependence	$\mathbb{E}(Y\|X{=}x',\mathrm{do}\,X{=}x) - \mathbb{E}(Y\|X{=}x,\mathrm{do}\,X{=}x)$
	causal effect	$\mathbb{E}(Y\|X{=}x',\mathrm{do}\,X{=}x') - \mathbb{E}(Y\|X{=}x',\mathrm{do}\,X{=}x)$
	deviation measure	difference
4.4.6	observed dependence	$\mathrm{d}_x\mathbb{E}(Y\|X{=}x)$
	back-door dependence	$\partial_1\mathbb{E}(Y\|X=x,\mathrm{do}\,X{=}x)$
	causal effect	$\partial_2\mathbb{E}(Y\|X=x,\mathrm{do}\,X{=}x)$
	deviation measure	difference

For all formalizations we present inequalities (see Table 4.1 for an overview) which turn out to always have the following prototypical form:

$$\begin{bmatrix}\text{back-door}\\ \text{dependence}\end{bmatrix} \geq \mathrm{d}\left(\begin{bmatrix}\text{observed}\\ \text{dependence}\end{bmatrix}, \begin{bmatrix}\text{causal}\\ \text{effect}\end{bmatrix}\right)$$

(where $\mathrm{d}(\cdot,\cdot)$ stands for deviation measure). In some of these results, observed dependence, back-door dependences, and causal effect are *real numbers* and $\mathrm{d}(\cdot,\cdot)$ simply stands for the usual difference which allows us to convert the prototypical form into

$$\begin{bmatrix}\text{causal}\\ \text{effect}\end{bmatrix} \geq \begin{bmatrix}\text{observed}\\ \text{dependence}\end{bmatrix} - \begin{bmatrix}\text{back-door}\\ \text{dependence}\end{bmatrix},$$

which may be more convenient for applications. In other cases, observed dependence and causal effect are *high- or infinite-dimensional objects* such as (conditional) distributions.

In order to draw conclusions on the true causal effect using the inequalities from the first part, one needs to have constraints on the back-door dependence. Therefore, in the *second part* – Section 4.5 –, we demonstrate how in various settings, one can *integrate* the structure of these settings such as to obtain constraints on the back-door dependence. Based on these constraints together with the observed dependence one can then use the results from the first part to infer bounds on the true causal effect.

Our approach can be seen as a step into the direction of establishing a formal, principled framework for causal inference methods that try to integrate structures beyond the classical randomized experiments and observational studies, in particular quasi-experimental designs. But it may well be that a general framework does not exist, as the settings are too inhomogeneous and always require some creativity for their discovery and formalization.

4.1.3. Structure of this chapter

The remainder of this chapter is structured as follows

- In Section 4.2 we discuss related work.

- Section 4.3 contains preliminaries.
- We conclude with Section 4.6.

Note that in contrast to the other chapters, we will include the central proofs directly within the chapter instead of putting them into the appendix as they are rather short and non-technical.

4.2. Related work

Several approaches have been developed to identify or approximate causal effects in i.i.d. settings in spite of hidden confounders.

Back-door/front-door criterion (see [Pearl, 2000, 2009]): We described this approach, which assumes that the complete causal DAG to be known and some variables besides X, Y to be observed, in Section 2.1.3.2 and in particular in Example 2.2. Besides the fundamental limitation – requiring knowledge of the complete causal DAG plus enough observed variables –, one drawback of this method is that it cannot be used if X is deterministically coupled to the variable that blocks the back-door path.

Instrumental variable (IV) design (see e.g. [Pearl, 2000, Angrist et al., 1996, Efron and Feldman, 1991]): In the simplest case, the causal DAG in Figure 4.2 is augmented by a parentless node Z with an arrow to X. An important example are clinical trials with partial compliance. The additional Z allows to infer bounds on the average causal effect. One drawback of this method is that it yields a convex optimization problem with the number of equations growing exponentially with the cardinality of X. Furthermore, to apply this method one needs to know $p(X, Y|Z)$ while in Section 4.2 we present a scenario where $p(Z)$ (additional to $p(X, Y)$) helps to estimate the causal effect.

Regression discontinuity (RD) design (see e.g. [Thistlewaite and Campbell, 1960, Imbens and Lemieux, 2008, Lee and Lemieux, 2010]): This framework is applicable to cases where an additional observable W mediating between Z and X is measured and X is a deterministic function of W that contains a discontinuity. Under the assumption of linearity of the remaining structural equations, the effect from X to Y, i.e. the

linear coefficient, can be identified. One limitation of this method is that it needs the discontinuity and a large slope alone does not suffice.

General quasi-experimental designs: IV, RD and similar designs are often subsumed under the name quasi-experimental designs. While lacking a precise definition (to the best our knowledge), characteristics of quasi-experimental designs are [Shadish et al., 2002]:

- the goal to infer causal effects, as with classical randomized experiments,
- a similar data collection procedure as with classical randomized experiments, in particular the ability to obtain samples for all relevant values of X (in the binary case: availability of some sort of treatment and control groups),
- no perfect control over, or randomization of, the assignment (implying less possibility for counterfactual reasoning), but some control or at least knowledge on the assignment mechanism.

Note that quasi-experiments are of particular relevance in economics [Meyer, 1995].

General framework for "non-classical" causal inference settings: It needs to be mentioned that alternative approaches to a formal framework for "non-classical" causal inference settings could be more principled than ours. For instance, such a framework could be based on ideas in [Balke and Pearl, 1994], where (1) knowledge is translated into constraints on the *complete* causal model (instead of just the strength of confounding) and then (2) the implications of the constraints on the effect of X on Y are calculated.

4.3. Preliminaries

Keep in mind that we will sometimes write $p(X)$ for the density of X, and $p(X|Y)$ for the density of X given Y, or $p(X|Y = y)$ when evaluating it at the point $Y = y$. Also keep in mind Remark 2.2 when reading expressions like $p(Y|X{=}x, \mathrm{do}\, X{=}x)$.

Throughout the chapter we will work with Z, X, Y with discrete as well as with continuous ranges. Unless noted otherwise, we make the following technical assumption regarding the distributions of the random variables in a FCM M (Definition 2.1) with

variables $X_1, \ldots, X_n$, corresponding structural equation functions $f_1, \ldots, f_n$, and causal DAG G: for each X_j, the random variable $f_j(\mathrm{pa}_j, N_j)$ has a density w.r.t. the Lebesgue measure (in the continuous case) or w.r.t. the counting measure (in the discrete case) respectively, denoted by $q_j(x_j; \mathrm{pa}_j)$ *for each value* pa_j *of* PA_j (note that we have to slightly deviate from this assumption in Section 4.5.1 though). This assumption implies the following simple lemma, which is only formulated for the case $n = 3$, since we only need this case in this thesis. A proof can be found in Section B.1.

Lemma 4.1. *Under the assumption regarding the density of* $f_j(\mathrm{pa}_j, N_j)$ *made above, the joint distribution of* X_1, X_2, X_3 *induced by a causal model* M *or any post-interventional model* $M_{\mathrm{do}\, X_i=x}$ *has a density w.r.t. the Lebesgue measure (in the continuous case) or counting measure (in the discrete case), respectively. Moreover, this density factorizes according to the causal DAG belonging to the respective model.*

We will use the following fact which immediately follows from [Pearl, 2000, Corollary 7.3.2].

Fact 4.1. *For all* x *we have*

$$p(Y|X = x, \mathrm{do}\, X{=}x) = p(Y|X = x),$$
$$\mathbb{E}(Y|X = x, \mathrm{do}\, X{=}x) = \mathbb{E}(Y|X = x).$$

4.4. The relation between observed dependence, back-door dependence and causal effect

In this section we present various possibilities to formalize the notions of observed dependence, back-door dependence and causal effect. For all formalizations we prove that the back-door dependence is equal to, or bounds from above, the deviation between the observed dependence and the actual causal effect.

Sections 4.4.1, 4.4.2 apply to X, Y, Z with finite range. Sections 4.4.3, 4.4.5 apply to X, Y, Z with arbitrary range. Sections 4.4.4 and 4.4.6 apply to X with continuous range.

4.4.1. Approximating the causal strength from X to Y

The basic quantities in this section are:
- observed dependence: $\mathrm{I}(X:Y)$,
- back-door dependence: $\mathrm{I}(X:Z)$, $\mathfrak{C}_{Z\to X}$,
- causal effect: $\mathfrak{C}_{X\to Y}$.

We consider the case of Z, X, Y having finite range. Janzing et al. [2013] proposed a definition for the causal strength of a set of arrows in a causal DAG. We briefly want to repeat this definition for the special case of measuring the strength of a single arrow. For a set of observables $V = \{X_1, \ldots, X_n\}$, a DAG G' with V as the set of nodes and a joint distribution $p(X_1, \ldots, X_n)$ and for any arrow $X_i \to X_j$ in G' we first define the distribution $p_{X_i\to X_j}$ corresponding to deleting $X_i \to X_j$ from the graph and feeding X_j with an independent copy of X_i instead, see also [Ay and Krakauer, 2007]:

$$
\begin{aligned}
p_{X_i\to X_j}(x_j|\mathrm{pa}_{X_j}^{X_i\to X_j}) &:= \sum_{x_i'} p(x_i')p(y|x_i', \mathrm{pa}_{X_j}^{X_i\to X_j}),\\
p_{X_i\to X_j}(x_k|\mathrm{pa}_{X_k}^{X_i\to X_j}) &:= p(x_k|\mathrm{pa}_{X_k}), \text{ for all } k \neq j,\\
p_{X_i\to X_j}(x_1, \ldots, x_n) &:= \prod_{k=1}^{n} p_{X_i\to X_j}(x_k|\mathrm{pa}_{X_k}^{X_i\to X_j}),
\end{aligned}
$$

where $\mathrm{pa}_{X_k}^{X_i\to X_j}$ denotes (values of) the set of parents of X_k in the modified graph G' without arrow $X_i \to X_j$ (obviously this actually only makes a difference for pa_{X_j}). Now we are able to define the *causal strength* $\mathfrak{C}_{X_i\to X_j}$ by the impact of the edge deletion:

$$
\mathfrak{C}_{X_i\to X_j} := \mathrm{D}(p(X_1, \ldots, X_n) \| p_{X_i\to X_j}(X_1, \ldots, X_n)).
$$

Let us get back to our specific confounding scenario (the causal DAG in Figure 4.2 on page 80). For general DAGs, Janzing et al. [2013] showed $\mathfrak{C}_{X\to Y} \geq I(X:Y|PA_Y \setminus X)$, that is, the information Y contains about X given its other parents is a lower bound for causal strength (they argue that this property would be desirable for other information-theoretic measures of causal strength as well). Hence in our confounding scenario we have $\mathfrak{C}_{X\to Y} \geq \mathrm{I}(X:Y|Z)$. Also keep in mind that $\mathfrak{C}_{Z\to X} = \mathrm{I}(Z:X)$ in our setting.

Lemma 4.2. *We have*

$$\mathrm{I}(X : Y|Z) \geq \mathrm{I}(X : Y) - \mathrm{I}(X : Z). \tag{4.1}$$

Proof. The statement follows from the fact that

$$\mathrm{I}(X : Y|Z) + \mathrm{I}(X : Z) = \mathrm{I}(X : Z, Y) \geq \mathrm{I}(X : Y),$$

see [Cover and Thomas, 1991]. □

We consider $\mathrm{I}(X : Y)$ as a measure of *observed dependence* between X and Y. The following theorem shows that the *back-door dependence* $\mathfrak{C}_{Z\to X}$ bounds the difference between the observed dependence and the true *causal effect* $\mathfrak{C}_{X\to Y}$.

Theorem 4.1. *We have*

$$\mathfrak{C}_{Z\to X} \geq \mathrm{I}(X : Y) - \mathfrak{C}_{X\to Y}. \tag{4.2}$$

Proof. This follows from Lemma 4.2 together with the fact that $\mathfrak{C}_{X\to Y} \geq \mathrm{I}(X : Y|Z)$ and $\mathfrak{C}_{Z\to X} = \mathrm{I}(X : Z)$ in our confounding scenario (i.e. the DAG in Figure 4.2). □

4.4.2. Approximating the information flow from X to Y

The basic quantities in this section are:
- observed dependence: $\mathrm{I}(X : Y)$,
- back-door dependence: $\mathrm{I}(X : Z)$,
- causal effect: $\mathrm{I}(X \to Y|\mathrm{do}\, Z)$.

Another information theoretic quantity to measure the causal effect of X on Y is the *information flow* proposed by [Ay and Polani, 2008]. In our setting (the causal DAG in

Figure 4.2) it is defined as

$$\begin{aligned}&\mathrm{I}(X \to Y|\mathrm{do}\, Z) := \\ &\sum_z p(z) \sum_x p(x|\mathrm{do}\, Z{=}z) \sum_y p(y|\mathrm{do}\, X{=}x, \mathrm{do}\, Z{=}z) \\ &\times \log \frac{p(y|\mathrm{do}\, X{=}x, \mathrm{do}\, Z{=}z)}{\sum_{x'} p(y|\mathrm{do}\, X{=}x', \mathrm{do}\, Z{=}z) p(x'|\mathrm{do}\, Z{=}z)}.\end{aligned}$$

Since $p(y|\mathrm{do}\, X{=}x, \mathrm{do}\, Z{=}z) = p(y|x, z)$ in our setting, we simply have $\mathrm{I}(X \to Y|\mathrm{do}\, Z) = \mathrm{I}(X : Y|Z)$.

So we can establish a theorem for the information flow similar to the one for the causal strength. It follows immediately from Lemma 4.2.

Theorem 4.2. *We have*

$$\mathrm{I}(X : Z) \geq \mathrm{I}(X : Y) - \mathrm{I}(X \to Y|\mathrm{do}\, Z). \tag{4.3}$$

4.4.3. Bounding the Kullback-Leibler divergence between $p(Y|X{=}x)$ and $p(Y|\mathrm{do}\, X{=}x)$

The basic quantities in this section are:
- observed dependence: $p(Y|X{=}x)$,
- back-door dependence: $\mathrm{I}(X : Z)$, $\min\{\mathfrak{C}_{Z\to X}, \mathfrak{C}_{Z\to Y}\}$,
- causal effect: $p(Y|\mathrm{do}\, X{=}x)$.

In some sense, $p(Y|\mathrm{do}\, X{=}x)$ is the most fundamental characterization of the *causal effect* from X to Y, while $p(Y|X{=}x)$ can be seen as the corresponding characterization of their *observed dependence*. In this section we show that the deviation between these two objects can be bounded by quantities which measure the *back-door dependence*, $\mathrm{I}(X : Z)$ and $\min\{\mathfrak{C}_{Z\to X}, \mathfrak{C}_{Z\to Y}\}$. We formalize the notion of deviation here by

$$\mathrm{D}(p(Y|X)\|p(Y|\mathrm{do}\, X)) = \sum_x p(x)\mathrm{D}(p(Y|x)\|p(Y|\mathrm{do}\, X{=}x)).$$

Theorem 4.3. *We have*

$$\mathrm{D}(p(Y|X)\|p(Y|\mathrm{do}\,X)) \leq \min\{\mathfrak{C}_{Z\to X}, \mathfrak{C}_{Z\to Y}\} \leq \mathrm{I}(X:Z).$$

Proof. First note that
$p_{Z\to X}(z,x,y) = p(z)p(x)p(y|z,x)$ and
$p_{Z\to Y}(z,x,y) = p(z)p(x|z)\sum_{z'} p(y|z,x)p(z')$.

This implies
$p(y|\mathrm{do}\,X{=}x) = p_{Z\to X}(y|X{=}x)$ and
$p(y|\mathrm{do}\,X{=}x) = p_{Z\to Y}(y|X{=}x)$.

Therefore, using the chain rule for Kullback-Leibler divergence,

$$\begin{aligned}
&\mathrm{D}(p(Y|X)\|p(Y|\mathrm{do}\,X)) \\
&= \mathrm{D}(p(Y|X)\|p_{Z\to X}(Y|X)) \\
&\leq \mathrm{D}(p(X,Y)\|p_{Z\to X}(X,Y)) \\
&= \mathfrak{C}_{Z\to X} (= \mathrm{I}(Z:X)).
\end{aligned}$$

Similarly one can derive $\mathrm{D}(p(Y|X)\|p(Y|\mathrm{do}\,X)) \leq \mathfrak{C}_{Z\to Y}$. □

The above theorem makes a statement w.r.t. the divergence between $p(Y|x)$ and $p(Y|\mathrm{do}\,X{=}x)$ *averaged* over all values x of X. But it is also possible to derive a pointwise version:

Theorem 4.4. *For all x*

$$\mathrm{D}(p(Y|x)\|p(Y|\mathrm{do}\,x)) \leq \mathrm{D}(p(Z|x)\|p(Z)),$$

with equality iff $p(z|x) = p(z)$ for all z.

Proof. By the log sum inequality we have

$$\begin{aligned} & p(y|x) \log \frac{p(y|x)}{p(y|\mathrm{do}\, x)} \\ &= \left(\sum_z p(y|x,z)p(z|x) \right) \log \frac{\sum_z p(y|x,z)p(z|x)}{\sum_z p(y|x,z)p(z)} \\ &\leq \sum_z p(y|x,z)p(z|x) \log \frac{p(y|x,z)p(z|x)}{p(y|x,z)p(z)} \qquad (4.4) \\ &= \sum_z p(y,z|x) \log \frac{p(z|x)}{p(z)}. \end{aligned}$$

Equality holds in (4.4) iff $p(y|x,z)p(z|x) = cp(y|x,z)p(z)$ for all z and some constant c, i.e. iff $p(z|x) = p(z)$ for all z. Summing over all y yields the claimed inequality. □

Note that taking the average w.r.t. X in Theorem 4.4 is another way to prove the first part of Theorem 4.3. With a similar proof we can also derive the following inequality w.r.t. the "inverse mutual information" $\mathrm{D}(p(Z)p(X)\|p(Z,X))$ (as opposed to the usual mutual information $\mathrm{I}(Z:X) = \mathrm{D}(p(Z,X)\|p(Z)p(X))$). For this purpose let us define

$$\mathrm{D}(p(Y|\mathrm{do}\, X)\|p(Y|X)) := \sum_x p(x) \sum_y p(y|\mathrm{do}\, X{=}x) \log \frac{p(y|\mathrm{do}\, X{=}x)}{p(y|X{=}x)}.$$

Corollary 4.1. *We have*

$$\mathrm{D}(p(Y|\mathrm{do}\, X)\|p(Y|X)) \leq \mathrm{D}(p(Z)p(X)\|p(Z,X)).$$

To assess which bound is relevant for a scenario, we recall that for two distributions p and q, $\mathrm{D}(p\|q)$ diverges when $q = 0$ and $p > 0$ on a set of Lebesgue measure greater than 0. If the observed dependence $p(Y|X)$ is deterministic, $p(Y|\mathrm{do}\, X)$ needs to be deterministic if $\mathrm{D}(p(Y|\mathrm{do}\, X)\|p(Y|X))$ is finite.

4.4.3.1. An example for bounding the average causal effect from X to Y

Often one is interested in estimating the *average causal effect*

$$\mathbb{E}(Y|\mathrm{do}\, X{=}x') - \mathbb{E}(Y|\mathrm{do}\, X{=}x)$$

for two values x, x' of X [Pearl, 2000], in particular because this quantity is easy to interpret. In what follows, we want to give an example how one can derive bounds on this quantity based on Theorem 4.3. It is important to mention however, that the assumptions we make are very restrictive. The purpose of the example is only to show that information theoretic bounds on the back-door dependence *can*, under appropriate assumptions, imply bounds for the average causal effect.

Let X be binary, $p(Y|x) = \mathcal{N}(\mu_x, \sigma^2)$, and $p(Y|\mathrm{do}\, X{=}x) = \mathcal{N}(\mu_{\mathrm{do}\, x}, \sigma^2_{\mathrm{do}})$, for $x = 0, 1$ (hence particularly $\mathbb{E}(Y|\mathrm{do}\, X{=}x) = \mu_{\mathrm{do}\, x}$).[1]

In this case we can calculate (have in mind that ln is the natural logarithm)

$$\begin{aligned} & p(X{=}0)(\mu_0 - \mu_{\mathrm{do}\,0})^2 + p(X{=}1)(\mu_1 - \mu_{\mathrm{do}\,1})^2 \\ & = 2\sigma^2_{\mathrm{do}}\left(\mathrm{D}(p(Y|X)\|p(Y|\mathrm{do}\, X)) - \ln\frac{\sigma^2_{\mathrm{do}}}{\sigma^2} - \frac{\sigma^2}{2\sigma^2_{\mathrm{do}}} + \frac{1}{2}\right) \\ & \leq 2\sigma^2_{\mathrm{do}}\left(\min\{\mathfrak{C}_{Z\to X}, \mathfrak{C}_{Z\to Y}\} - \ln\frac{\sigma^2_{\mathrm{do}}}{\sigma^2} - \frac{\sigma^2}{2\sigma^2_{\mathrm{do}}} + \frac{1}{2}\right). \end{aligned} \tag{4.5}$$

Now assume we fix $\min\{\mathfrak{C}_{Z\to X}, \mathfrak{C}_{Z\to Y}\}$ and σ^2_{do}. Keep in mind that μ_0, μ_1, σ^2 are observed. Then we can derive upper and lower bounds on the average causal effect $\mu_{\mathrm{do}\,1} - \mu_{\mathrm{do}\,0}$ by maximizing and minimizing it, respectively, under the constraints on the pair $(\mu_{\mathrm{do}\,1}, \mu_{\mathrm{do}\,0})$ imposed by inequality (4.5).

4.4.4. Approximating the Fisher information

The basic quantities in this section are:
- observed dependence: $\mathcal{F}_{Y|X}(x)$,
- back-door dependence: $\mathcal{F}^1_{Y|X,\mathrm{do}\, X}(x, x)$,
- causal effect: $\mathcal{F}^2_{Y|X,\mathrm{do}\, X}(x, x)$.

[1] Note, however, that both $p(Y|X{=}0)$ and $p(Y|X{=}1)$ being Gaussian actually provides some evidence for the absence of confounding since a confounder will often destroy this simple structure of $P(Y|X)$ [Janzing et al., 2011].

In the following, $\partial_i f(w_1, \ldots, w_n)$ denotes the partial derivative w.r.t. the ith argument of f evaluated at position $(w_1, \ldots, w_n)$. And $\mathrm{d}_w g(w)$ denotes the total derivative of $g(w)$ w.r.t. w at position w, in particular $\mathrm{d}_w f(w, w) = \mathrm{d}_w g(w)$ for $g(w) := f(w, w)$.

Given a family of distributions depending on continuous parameters, *Fisher information* provides a natural way to quantify the sensitivity of a probability distribution to infinitesimal parameter changes. It plays an important role for the error made when estimating the true parameter value from empirical data [Rao, 1945]. Here we quantify causal influence by the sensitivity of $p(Y|\mathrm{do}\, x)$ to small changes of x. This can be considered as a "local" measure of causal strength in the neighborhood of x. We introduce the following notation:

$$\begin{aligned}\mathcal{F}_{Y|X}(x) &:= \int (\mathrm{d}_x \log p(y|X{=}x))^2 p(y|X{=}x)\mathrm{d}y,\\ \mathcal{F}^i_{Y|X,\mathrm{do}\,X}(x,x') &:= \int (\partial_{i+1} \log p(y|X{=}x, \mathrm{do}\, X{=}x'))^2 p(y|X{=}x, \mathrm{do}\, X{=}x')\mathrm{d}y,\end{aligned}$$

for $i = 1, 2$. (Note that y in $\log p(y|X{=}x, \mathrm{do}\, X{=}x')$ counts as argument, so, for instance, $\partial_2 \log p(y|X{=}x, \mathrm{do}\, X{=}x')$ is the partial derivative w.r.t. x.)

Theorem 4.5. *For all x,*

$$\sqrt{\mathcal{F}_{Y|X}(x)} - \sqrt{\mathcal{F}^2_{Y|X,\mathrm{do}\,X}(x,x)} \leq \sqrt{\mathcal{F}^1_{Y|X,\mathrm{do}\,X}(x,x)}.$$

Proof. First note that by the chain rule

$$\mathrm{d}_x \log p(y|X{=}x, \mathrm{do}\, X{=}x) = \partial_2 \log p(y|X{=}x, \mathrm{do}\, X{=}x) + \partial_3 \log p(y|X{=}x, \mathrm{do}\, X{=}x).$$

By Fact 4.1 we have $p(y|X{=}x) = p(y|X{=}x, \mathrm{do}\, X{=}x)$ for all x, y.

Together we obtain

$$\begin{aligned}&\left(\mathbb{E}((\mathrm{d}_x \log p(y|X{=}x))^2)\right)^{\frac{1}{2}}\\ &= \left(\mathbb{E}((\partial_2 p(y|X{=}x, \mathrm{do}\, X{=}x) + \partial_3 p(y|X{=}x, \mathrm{do}\, X{=}x))^2)\right)^{\frac{1}{2}}\\ &\leq \left(\mathbb{E}((\partial_2 p(y|X{=}x, \mathrm{do}\, X{=}x))^2)\right)^{\frac{1}{2}} + \left(\mathbb{E}((\partial_3 p(y|X{=}x, \mathrm{do}\, X{=}x))^2)\right)^{\frac{1}{2}}.\end{aligned}$$

Note that the expectation is taken w.r.t. $p(y|x)$. □

4.4.5. Approximating the effect of treatment on the treated from X to Y

The basic quantities in this section are:
- observed dependence: $\mathbb{E}(Y|X{=}x') - \mathbb{E}(Y|X{=}x)$,
- back-door dependence: $\mathbb{E}(Y|X{=}x', \mathrm{do}\, X{=}x) - \mathbb{E}(Y|X{=}x, \mathrm{do}\, X{=}x)$,
- causal effect: $\mathbb{E}(Y|X{=}x', \mathrm{do}\, X{=}x') - \mathbb{E}(Y|X{=}x', \mathrm{do}\, X{=}x)$.

Following [Pearl, 2000], we define the quantity

$$\mathbb{E}(Y|X{=}x', \mathrm{do}\, X{=}x') - \mathbb{E}(Y|X{=}x', \mathrm{do}\, X{=}x)$$

as the *effect of treatment on the treated*. As the name already suggests, the idea behind this quantity is to measure the deviation between the average response of the treated subjects and the average response of these same subjects had they not been treated. The following result w.r.t. the effect of treatment on the treated follows from Fact 4.1.

Theorem 4.6. *We have for all* x, x'

$$\begin{aligned}\mathbb{E}(Y|X{=}x') - \mathbb{E}(Y|X{=}x) = \mathbb{E}(Y|X{=}x', \mathrm{do}\, X{=}x') - \mathbb{E}(Y|X{=}x', \mathrm{do}\, X{=}x) \\ + \mathbb{E}(Y|X{=}x', \mathrm{do}\, X{=}x) - \mathbb{E}(Y|X{=}x, \mathrm{do}\, X{=}x).\end{aligned}$$

Note that in mediation analysis [Pearl, 2001, Avin et al., 2005, Robins and Greenland, 1992] a similar splitting into direct and indirect effect is used. However mediation analysis addresses the problem of defining direct and indirect *causal* effects and not back-door dependences.

We briefly want to discuss the other quantities that appear in Theorem 4.6. Obviously, $\mathbb{E}(Y|X{=}x') - \mathbb{E}(Y|X{=}x)$ measures the *observed dependence* of Y on X. Now keep in mind that in $M_{\mathrm{do}\, X=x}$, X has no causal effect on Y anymore and hence Y statistically depends on X solely via Z. Therefore the difference

$$\mathbb{E}(Y|X{=}x', \mathrm{do}\, X{=}x) - \mathbb{E}(Y|X{=}x, \mathrm{do}\, X{=}x)$$

measures the strength of the *back-door dependence* of Y on X.

4.4.6. Approximating the differential effect of treatment on the treated from X to Y

The basic quantities in this section are:
- observed dependence: $\mathrm{d}_x\mathbb{E}(Y|X{=}x)$,
- back-door dependence: $\partial_1\mathbb{E}(Y|X = x, \mathrm{do}\, X{=}x)$,
- causal effect: $\partial_2\mathbb{E}(Y|X = x, \mathrm{do}\, X{=}x)$.

First note that by $\partial_i\mathbb{E}(Y|X{=}x, \mathrm{do}\, X{=}x')$ we mean $\partial_i f(x, x')$ for

$$f(x, x') := \mathbb{E}[Y|X{=}x, \mathrm{do}\, X{=}x']$$

(recall that ∂_i denotes the partial derivative w.r.t. the ith argument). In the case of continuous random variables Z, X, Y we want to consider the following quantity (if it exists i.e. if the conditional expectation is differentiable):

$$\partial_2\mathbb{E}(Y|X{=}x, \mathrm{do}\, X{=}x),$$

which we call *differential effect of treatment on the treated* or simply *differential effect* in cases where this does not lead to confusions. It is the analog to the discrete effect of treatment on the treated (see Section 4.4.5) for the case of infinitesimal interventional changes on X; we simply replaced a difference by a derivative.

Similar to Theorem 4.6 we can establish the following result. It follows from the chain rule for derivatives together with Fact 4.1.

Theorem 4.7. *For all x*

$$\mathrm{d}_x\mathbb{E}(Y|X{=}x) = \partial_1\mathbb{E}(Y|X{=}x, \mathrm{do}\, X{=}x) + \partial_2\mathbb{E}(Y|X{=}x, \mathrm{do}\, X{=}x).$$

The interpretation of this theorem is similar to the one for Theorem 4.6. Obviously, $\mathrm{d}_x\mathbb{E}(Y|X{=}x)$ is the *observed dependence*, whereas the quantity $\partial_1\mathbb{E}(Y|X = x, \mathrm{do}\, X{=}x)$ measures the *back-door dependence* of Y on X. So the observed dependence of Y on X splits into the causal effect plus the back-door dependence.

4.5. Prototypical application scenarios: integrating knowledge that bounds the back-door dependence

In this section we present several prototypical scenarios where knowledge or beliefs can be integrated that allow to derive bounds on the back-door dependence between X and Y. Together with our results from Section 4.4 these bounds help to approximately estimate the causal effect from X to Y.

4.5.1. A qualitative toy example

We want to give an example that demonstrates how human intuition concerning observed dependence and causal effect relates to the theorems from Section 4.4.

Assume there is a drug that is indicated for a specific disease. We observe some not too small number of people with the disease and see that some of them take the drug and some do not. We find that all persons who took the drug recovered on the same day whereas none of the persons not taking the drug recovered that fast. For each sick person let X denote the date he or she takes the drug and Y the date he or she recovers. Since these are only observations, we cannot exclude that there is a (hidden) confounder Z, i.e. we assume the usual causal DAG (Figure 4.2 on page 80). We estimate the distribution of Y given X by the empirical distribution, i.e. $p(y|x) = \delta_{yx}$, where δ_{yx} denotes the Kronecker delta.

Given the above setting, probably most people would argue that there has to be some effect from the drug to the immediate healing of those people who took it. However, formally and without further assumptions, $p(Y|x)$ alone does not even tell us if there is a causal link from X to Y at all. With the help of Theorem 4.3 though, we can formally reason as follows. We make the weak additional assumption that X cannot be completely determined by Z which we formalize by $\mathrm{I}(Z : X) < \mathrm{H}(X)$. It seems implausible that there exists a common cause of X and Y that determines both, the exact date X a person takes the drug and the recovering date Y. E.g. the wealth of a person may strongly influence both, the treatment he or she takes and how quickly he

or she recovers (via the general health condition), however it seems not plausible that the wealth determines the exact day of taking the drug and of recovering.

For a proof by contradiction we may assume that there is no causal effect from X to Y, i.e. $p(Y|\mathrm{do}\, X{=}x) = p(Y|\mathrm{do}\, X{=}x')$ for all x, x'. Then

$$\begin{aligned}
&\mathrm{D}(p(Y|X)\|p(Y|\mathrm{do}\, X{=}x)) \\
&= \sum_x p(x)\mathrm{D}(p(Y|X{=}x)\|p(Y|\mathrm{do}\, X{=}x)) \\
&= \sum_x p(x) \sum_y \delta_{yx} \log \frac{\delta_{yx}}{P(Y=y|\mathrm{do}\, X{=}x)} \\
&= \sum_x p(x) \log \frac{1}{p(Y{=}x|\mathrm{do}\, X{=}0)} \geq \mathrm{H}(X),
\end{aligned}$$

where the last inequality is due to Gibb's inequality [Cover and Thomas, 1991].

On the other hand, due to Theorem 4.3 we have

$$\mathrm{D}(p(Y|X)\|p(Y|\mathrm{do}\, X)) \leq \mathrm{I}(X:Z) < \mathrm{H}(X),$$

which yields the contradiction. Hence we could formally show that there has to be some causal effect from X to Y, $p(Y|\mathrm{do}\, X{=}x) \neq p(Y|\mathrm{do}\, X{=}x')$ for some x, x'. Note that the above argumentation completely transfers to any other situation where $p(y|x) = \delta_{yx}$, particularly any other range of X and Y.

4.5.2. Partial randomization scenario

We first discuss a formal scenario, then an application example, and afterwards we discuss how the scenario and our result is related to the instrumental variable design [Pearl, 2000]. Here we assume X, Y, Z to have finite range.

4.5.2.1. The formal prototype

We consider a scenario where we have measured X and Y, and where hidden variables Z and W are present and we know the distribution of W. The underlying causal structure

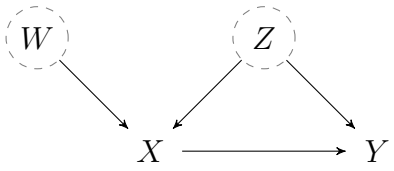


Figure 4.3.: The partial randomization causal DAG.

of all variables looks like the causal DAG depicted in Figure 4.3. We assume that W is binary. Furthermore we assume that in this scenario $\mathrm{I}(Z : X|W = 0) = 0$. The intuition behind this assumption is that W decides whether X is influenced by Z ($W = 1$) or not. This scenario implies the following inequality. Keep in mind that $|\mathrm{dom}(X)|$ denotes the size of (number of elements in) the domain of X.

Proposition 4.1. *In the given scenario we have*

$$\mathrm{I}(Z : X) \leq \log\left(|\mathrm{dom}(X)|\right) p(W{=}1). \tag{4.6}$$

Proof. We calculate

$$\begin{aligned}
\mathrm{I}(Z : X) &\leq \mathrm{I}(Z : X) + \mathrm{I}(Z : W|X) \\
&= \mathrm{I}(Z : W, X) \\
&= \mathrm{I}(Z : W) + \mathrm{I}(Z : X|W) \\
&= \mathrm{I}(Z : X|W{=}0)p(W{=}0) + \mathrm{I}(Z : X|W{=}1)p(W{=}1) \\
&= \mathrm{I}(Z : X|W{=}1)p(W{=}1) \\
&\leq \mathrm{H}(X|W = 1)p(W{=}1) \\
&\leq \log\left(|\mathrm{dom}(X)|\right) p(W{=}1).
\end{aligned}$$

□

But Theorems 4.1, 4.2 and 4.3 all establish deviations between observed dependence and causal effect bounded by $\mathrm{I}(Z : X)$ (keep in mind that in our scenario $\mathfrak{C}_{Z\to X} = \mathrm{I}(X : Z)$).

Plugging them together with Inequality 4.6, we obtain the following bounds:

$$\mathrm{I}(X:Y) - \mathfrak{C}_{X\to Y} \leq \log\left(|\mathrm{dom}(X)|\right) p(W{=}1), \tag{4.7}$$

$$\mathrm{I}(X:Y) - \mathrm{I}(X \to Y|\mathrm{do}\, Z) \leq \log\left(|\mathrm{dom}(X)|\right) p(W{=}1), \tag{4.8}$$

$$\mathrm{D}(p(Y|X)\|p(Y|\mathrm{do}\, X)) \leq \log\left(|\mathrm{dom}(X)|\right) p(W{=}1). \tag{4.9}$$

4.5.2.2. Advertisement letter example: partial randomization by partial compliance

A common application field of causal inference within decision making is advertisement [Brodersen et al., 2015]. There, the decision is about say a potential ad campaign, and the goal is formulated w.r.t. benefits from subsequent client behaviour as well as say costs of the ad campaign. To inform the decision, it is important to have an idea of the *causal effect* of the ad campaign on the subsequent behaviour of (potential) clients.

In ideal scenarios, methods like the back-door criterion (Section 2.1.3.2) or even methods which do not explicitly talk about causal semantics, such as multiarmed bandits, are applicable [Bottou et al., 2013]. In contrast, here we consider a toy scenario where only an insufficient set of variables is recorded, but some additional knowledge – about partial compliance – can be integrated to approximate the causal effect, based on Section 4.5.2.1 above.

Example 4.1. *Assume we are managers of a mail order company and want to know the effect of sending advertisement letters on the ordering behavior of the recipients, to inform our decision making regarding advertisement campaigns. We have a data set of* (X, Y) *pairs with* $X \in \{0, 1\}$ *denoting whether a letter was sent to a specific person and let* Y *denote the total costs of the products ordered by this person afterwards (while we assume the identity of the person not to be recorded). Suppose we have enough data to estimate* $p(X, Y)$*. Furthermore, assume that so far there were already imperfect guidelines based on rough intuition on whom to send letters and whom not. These guidelines introduce a potential confounder* Z *since letters were more likely send to potential customers with properties that made them also more likely to order something (if the guidelines were not complete nonsense). More specifically,* Z *could denote the recommendation for*

a specific customer based on the guidelines. But we assume that Z was not recorded (for instance because we as managers were not familiar with causal inference before).

It is known however that only some employees followed these guidelines/recommendations. Let W denote whether a letter was sent out in compliance with these guidelines ($W = 1$) or not (i.e., was send out more or less randomly). Based on an estimate of how many employees complied with the guidelines, we also have an estimate of $p(W = 1)$, i.e. the fraction of letters that was sent out in compliance with the guidelines. Based on Proposition 4.1, we know that $\mathrm{I}(Z : X) \leq \log(|\mathrm{dom}(X)|)\, p(W{=}1)$. Hence we have an upper bound on the back-door dependence of Y on X. In particular, we can apply Inequalities (4.7) to (4.9) and, under strong additional assumptions, the result w.r.t. the average causal effect from Section 4.4.3.1.

For example, by (4.7) we have $\mathrm{I}(X{:}Y) - \log(|\mathrm{dom}(X)|)\, p(W{=}1) \leq \mathfrak{C}_{X\to Y}$. Since X is binary, we have $\log(|\mathrm{dom}(X)|) = 1$ (we assume the logarithm in information theoretic quantities to be w.r.t. basis 2 here). Furthermore, $p(W{=}1){\approx}0.5$ (only half the employees followed the guidelines). Then, if we observe a strong dependence of Y on X, say $\mathrm{I}(X{:}Y){\approx}0.75$, we can conclude that $\mathfrak{C}_{X\to Y}{\gtrsim}0.25$, i.e. our advertisement letters have a significant effect on the potential customers. If one wants to know more about precisley how the effect looks like, one could apply Inequality 4.9.

It is important to emphasize, that here we utilized *partial compliance with guidelines for causal inference, while in other scenarios, another form of partial compliance may* complicate *causal inference – see the IV design (Section 4.2 and discussion below). Furthermore, note that in this simple example we only considered the causal effect of ad letters on the whole population; as a next step, one would use additional information about the recipients (which the imperfect guideline mentioned above already may have done).*

4.5.2.3. Difference to instrumental variable design

We already mentioned the instrumental variable design [Pearl, 2000] in Section 4.1. In this design it is assumed that an additional variable W is observed such that the causal structure of all variables together is as depicted in Figure 4.3, except that W is not hidden. The prototypical application scenario for this design are clinical trials with

partial compliance. Pearl [2000] describes a method to derive bounds on the average causal effect $\mathbb{E}(Y|\mathrm{do}\, X{=}1) - \mathbb{E}(Y|\mathrm{do}\, X{=}0)$. This analysis heavily depends on the range of X, Y, and W and involves convex optimization in a 15-dimensional space already for the case where all variables are binary (since Z can be assumed to attain 16 different values).

The advantage of our approach lies in the fact that the ranges of the variables may be arbitrary without increasing the complexity – for the cost of getting less tight bounds than an explicit modeling, of course. One can get bounds for the case where neither X nor Y are binary, e.g., in a drug testing scenario with different doses and descriptions of health conditions that are more complex than just reporting recovery or not. Moreover, we do not need complete knowledge of $p(Y, X|W)$ provided that we have some knowledge on W that provides upper bounds on $\mathrm{I}(X{:}Z)$.

4.5.3. A variant of the regression discontinuity design

We already mentioned the regression discontinuity design (RDD) [Thistlewaite and Campbell, 1960, Imbens and Lemieux, 2008, Lee and Lemieux, 2010] in Section 4.2. It is a quasi-experimental design that can help to estimate the causal effect from X to Y in cases where an additional variable W is measured and the underlying causal DAG of all variables together is as depicted in Figure 4.4. The design usually requires that X is a deterministic function of W that contains a discontinuity, that all remaining structural equations are linear, and that $\mathbb{E}(Z|W = w)$ is continuous in w. (Note that the causal DAG in Figure 4.4 is a special case of the general confounding scenario depicted in Figure 4.2, which can be seen by replacing Z in Figure 4.2 by $Z' := (Z, W)$.)

We now consider a scenario inspired by the RDD, which allows to bound the back-door dependence in the sense of Section 4.4.6 and thus makes Theorem 4.7 applicable to estimate the causal effect $\partial_2\mathbb{E}(Y|X{=}x, \mathrm{do}\, X{=}x)$, i.e. the differential effect of treatment on the treated. The scenario differs from the RDD in that neither a discontinuity in the structural equation for X, nor linearity of the remaining structural equations is required.

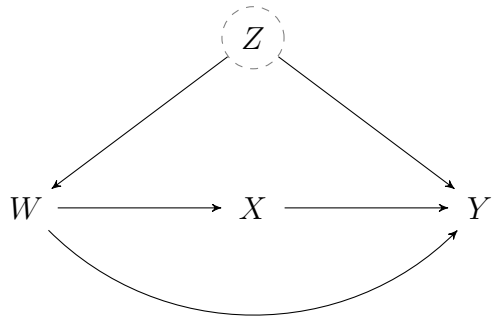


Figure 4.4.: The causal DAG for the RDD and our variant of it.

Assume the causal DAG in Figure 4.4. Furthermore assume that $X = f_X(W)$ for a function f_X that is differentiable. (This is the point where our scenario differs from RDD.) Suppose f_X is invertible, $g := f_X^{-1}$. It can easily be seen that this implies

$$\partial_1 \mathbb{E}(Y|X{=}x, \operatorname{do} X{=}x) = \partial_1 \mathbb{E}(Y|W{=}g(x), \operatorname{do} X{=}x) g'(x). \tag{4.10}$$

Note that $\partial_1 \mathbb{E}(Y|W{=}g(x), \operatorname{do} X{=}x)$ means the derivative of $\mathbb{E}(Y|W{=}w, \operatorname{do} X{=}x)$ w.r.t. w at position $(g(x), x)$. Applying Theorem 4.7 yields

$$\mathrm{d}_x \mathbb{E}(Y|X{=}x) - \partial_2 \mathbb{E}(Y|X{=}x, \operatorname{do} X{=}x) = \partial_1 \mathbb{E}(Y|W{=}g(x), \operatorname{do} X{=}x) g'(x).$$

Hence if for any position x_0 of X we assume a bound on the strength of the "back-door" dependence of Y on W, $\partial_1 \mathbb{E}(Y|W{=}g(x_0), \operatorname{do} X{=}x_0)$, and if $|g'(x_0)|$ is comparably small (which is the case when $|f'_X(g(x_0))|$ is big), then we can bound the difference between observed dependence and causal effect at position x_0.

For instance, if we consider the observed dependence $\mathrm{d}_x \mathbb{E}(Y|X{=}x)$ as a realistic scale based on which one can constrain $\partial_1 \mathbb{E}(Y|W{=}g(x), \operatorname{do} X{=}x)$, formally

$$|\partial_1 \mathbb{E}(Y|W{=}g(x), \operatorname{do} X{=}x)| \leq c |\mathrm{d}_x \mathbb{E}(Y|X{=}x)|,$$

for some c, then one can bound the modulus of the causal effect from below:

$$|\partial_2 \mathbb{E}(Y|W{=}g(x), \operatorname{do} X{=}x)| \geq (1 - c|g'(x)|) |\mathrm{d}_x \mathbb{E}(Y|X{=}x)|.$$

Obviously, one weakness of the above argument is that the estimation of the causal effect heavily depends on the bound that we assume w.r.t. the "back-door" dependence of Y on W, $\partial_1 \mathbb{E}(Y|W{=}g(x), \mathrm{do}\, X{=}x)$. However, this can be seen as a quantitative analogon to the qualitative assumption of the RDD that $\mathbb{E}(Z|W = w)$ is continuous in w.

Keep in mind that our results on Fisher information (Section 4.4.4) can be used in the case where X is not a deterministic function of W that changes rapidly but instead the conditional probability $p(X|w)$ changes fast at some $w = w_0$.

4.6. Conclusions of this chapter

In this chapter, we tried to take a step towards a general formalization of methods that lie between randomized experiments and observational studies. Rephrasing the basic idea, we integrated settings with a *bounded* "deviation from perfect experiments". We gave some examples of such settings and showed how they imply bounds on confounding (Section 4.5). And we showed how such bounds on confounding imply approximations to the true causal effect we are interested in (Section 4.4).

Chapter 5.

Decision making in cloud computing via approximate causal models

5.1. Introduction

In Chapters 3 and 4, we mainly focused on integrative and approximative learning of PCMs with only a small digression to decision making in Chapter 4. The investigation took place on a rather abstract level, meaning that it was not motivated by specific problems from specific domains, but rather we tried to develop approaches that in principle can be applied to all possible domains (that satisfy the respective conditions).

In contrast, in this chapter the main focus is on decision making using PCMs, while learnig of PCMs is discussed only briefly, as illustrated by Figure 5.1 on page 105. Furthermore, the investigation is driven by rather specific technical and economical problems in cloud computing [Armbrust et al., 2010]. Cloud computing is a computing paradigm as well as a business model that has become increasingly popular in recent years. It allows to rent computing resources on-demand, and to use them efficiently by sharing them in a smart way, in particular using auctions to sell unused resources.

For the first time within the main part of this thesis, here we will use the notions of counterfactuals and transportability, which can be defined based on PCMs (although we already mentioned counterfactuals in the introductory Section 2.1.5). Similarly to Chapter 4, a focus will be on *approximations*.

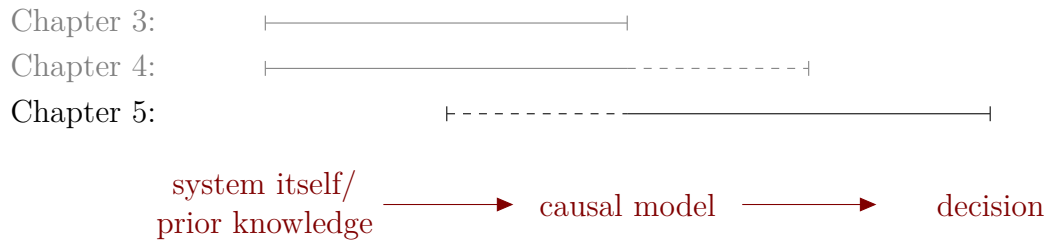


Figure 5.1.: The content of this chapter illustrated in black, relative to the rest of this thesis in gray, and the overall "inference path" in red.

Keep in mind that this chapter is made up of initial ideas rather than fully elaborated approaches. Nonetheless, most notably Section 5.5 goes into a direction – approximate causal reasoning for economic decision making problems – which seems promising.

Parts of this chapter are based on the pre-prints [Geiger et al., 2016b] and [Geiger et al., 2016a].

5.1.1. Problem outline

Several new challenges arise from the paradigm of cloud computing. On a technical level, it is a problem to understand, control and debug the involved computing systems up to the size of several data centers, with as much automation as possible, to make them behave in a desired way. We will go into more detail on this in Section 5.4.1. On an economical level, while auctions for "spot" resources help providers to use resources more efficiently, the unpredictability of their prices and performance complicates bidding and buying decisions for clients. We will go into more detail on this in Section 5.5.1.

In the absence of exact models, it is natural to try to address such problems using data-driven methods [Padala et al., 2009, Ostrowski et al., 2011, Snee et al., 2015, Chiang et al., 2014, Zheng et al., 2009]. However, standard machine learning usually applies in settings where the underlying system is invariant, often based on the assumption that

samples are i.i.d., and does not make predictions about the effect of interventions, which is important though for debugging, control and integration of heterogeneous data.

5.1.2. Contributions

The present chapter thus takes first steps towards addressing challenges of cloud computing using *causal* models. Inferring causal models from (observational) data is notoriously hard, and convincing applications of causal modeling to real world problems are scarce. The present chapter is no exception in that the main focus is conceptual rather than empirical. Our main contributions are:

- We present two theoretical results for approximations in causal modeling, Propositions 5.1 and 5.2, which are of relevance for the subsequent cloud problems and possibly beyond, in Section 5.3. It needs to be emphasized that the practicability of these theoretical results remains to be proved.
- In Section 5.4, we suggest first steps towards causal models and approximate counterfactuals as a principled approach for addressing cloud control and performance debugging problems, integrating sandbox experiments.
- In Section 5.5, we use approximate integration of causal knowledge to enable cloud clients to better predict performance and costs, while preserving privacy, in a toy setting.

It needs to be emphasized again that the practical value of the two propositions and our formalizations of the mentioned problems remains to be established. They should be seen as a thought-provoking impulse rather than a completed contribution.

5.1.3. Structure of this chapter

The remainder of this chapter is structured as follows:

- in Section 5.3, we give a brief introduction to cloud computing;

- Section 5.3 contains the definition of counterfactuals (in addition to our two theoretical results);
- Section 5.6 contains simplistic real-world and simulated experiments for our two approaches, as well as a preliminary causal model of a more realistic cloud system;
- in Section 5.7, we discuss related work;
- and we conclude the chapter with Section 5.8.

5.2. Background in cloud computing

Traditionally, both businesses and individuals have used dedicated local computers, or computer networks, for storing, managing and processing data. However, this can be inefficient in several ways: the overhead of maintaining such an infrastructure is high, and one needs to buy enough computers to handle peak loads, while during normal operation most will remain unutilized

Cloud computing significantly changes this, by allowing computing resources to be rented on demand. A company, the *cloud provider*, is now responsible for running all the hardware, keeping it upgraded and sharing it amongst multiple clients. Such an infrastructure can be run in a highly efficient manner: tens or hundreds of *virtual machines (VMs)*, i.e., emulations of computer systems, chartered by different clients, run on a single *physical server* and share its resources such as central processing units (CPUs), memory and network. Note that we refer to a system as being *in production*, if this system does actual work for clients and visitors, and if contracts have to be met w.r.t. this system (in contrast, e.g., to an experimental system).

5.3. Two approximations in causal modeling

5.3.1. Structural counterfactuals and an approximation

Let M_0 be an FCM over a set V of variables, and let U denote the set of independent background variables in M_0 (as described in Definition 2.1). Let E, X, Y be (sets of) variables in V. The *structural counterfactual probability of Y being y, had X been x, given evidence $E = e$*, can be defined [Pearl, 2000] based on M_0 as[1]

$$p(Y_{\mathrm{do}\, X=x} = y|e) := \sum_u p(y|\mathrm{do}\,(x), u)p(u|e). \tag{5.1}$$

Even though computer systems are "more deterministic" than many other systems, due to interactions with the environment and missing information, one usually can only infer a GCM, and not an FCM, of a computer system. Without an FCM though, counterfactual probabilities (Equation (5.1)) are generally not uniquely determined, i.e., they cannot be derived from a GCM. Let us give an example.

Example 5.1 (GCMs do not determine counterfactual probabilities)**.** *Let $V = \{X, Y\}$ for binary X, Y, and consider the GCM M with DAG $X \to Y$ and conditionals $p_X(0) = \frac{1}{2}$ and $p_{Y|X}(0|x) = p_Y(0) = \frac{1}{2}$. M is induced by two very different FCMs. On the one hand, the FCM M_0 with structural equations*

$$\begin{aligned} X &:= U_X, \\ Y &:= U_Y, \end{aligned}$$

and $U_X \sim U_Y \sim Uniform(\{0, 1\})$, where $Uniform(\{0, 1\})$ denotes the uniform distribution on $\{0, 1\}$, induces M (see Remark 2.1 for what we mean by "induce" here). On the other hand, the FCM M_0' with structural equations

$$\begin{aligned} X &:= U_X, \\ Y &:= X \text{ XOR } U_Y, \end{aligned}$$

[1] Note that [Pearl, 2000] in his definition uses functions instead of (deterministic) conditionals.

and $U_X \sim U_Y \sim Uniform(\{0,1\})$ *induces* M*. But in* M_0 *we have*

$$\begin{aligned}
&p(Y_{\mathrm{do}\, X=1} = 0 | X = 0, Y = 0) \\
&= \sum_{u_Y} p(Y = 0 | \mathrm{do}\, X = 1, u_Y) p(u_Y | X = 0, Y = 0) \\
&= \sum_{u_Y} p(Y = 0 | u_Y) p(u_Y | X = 0, Y = 0) \\
&= 1 \cdot 1 + 0 \cdot 0 = 1,
\end{aligned}$$

while in M_0' *we have*

$$\begin{aligned}
&p(Y_{\mathrm{do}\, X=1} = 0 | X = 0, Y = 0) \\
&= \sum_{u_Y} p(Y = 0 | \mathrm{do}\, X = 1, u_Y) p(u_Y | X = 0, Y = 0) \\
&= \sum_{u_Y} p(Y = 0 | X = 1, u_Y) p(u_Y | X = 0, Y = 0) \\
&= 0 \cdot 1 + 1 \cdot 0 = 0.
\end{aligned}$$

This gives an extreme example of counterfactual probabilities not being determined by a GCM.

For a more detailed discussion of this phenomenon we refer the reader to [Peters et al., 2017].

Now we show that nonetheless counterfactual probabilities can be calculated *approximately*, and one can *know*, from only the GCM, how wrong the approximation is at most – on average. This will be important for our approach to debugging in Section 5.4, and, as we belief, for other areas as well.

Let M be a GCM and let Z be the set of its root variables (variables with no parents in the causal DAG). For any (sets of) variables X, Y, E in M we define the *approximate structural counterfactual* or *approximate counterfactual* as[2]

$$\tilde{p}(Y_{\mathrm{do}\, X=x} = y | e) := \sum_{w} p(y | \mathrm{do}\,(x), w) p(w | e), \tag{5.2}$$

[2]The idea of a counterfactual definition based on only the GCM has been mentioned in [Pearl, 2000, Section 7.2.2], but not been further investigated. Depending on the specific setting and the available information, there may be more suitable approximations to encode counterfactual-like probabilities.

where $W := Z \setminus X$.

Proposition 5.1. *Let M_0 be an FCM that induces a GCM M, and let Z denote the root variables in M. For all (sets of) variables E, X, Y we have*

$$\mathrm{D}(p(Y_{\mathrm{do}\, X=x}|E)\|\tilde{p}(Y_{\mathrm{do}\, X=x}|E)) \leq \mathrm{H}(E|Z), \tag{5.3}$$

where $p(Y_{\mathrm{do}\, X=x}|e)$ is defined w.r.t. M_0 and $\tilde{p}(Y_{\mathrm{do}\, X=x}|e)$ w.r.t. M.

We prove (using monotonicity of the KL divergence and properties of entropy) a generalization of Proposition 5.1 – Proposition C.1 – in Section C.1.[3]

Example 5.2. *To give some intuition about the approximate counterfactual and the proposition, let us first consider the following two special cases: If M is already an "FCM" in the sense that all its variables are completely determined by the root nodes, then we have $H(E|Z) = 0$, and thus, based on Equation (5.3), both quantities coincide, which seems natural. If the evidence comprises the root nodes, $Z \subset E$, then the approximation amounts to the simple conditional $p(y|\mathrm{do}\,(x), w)$ (where w is the part of e the corresponds to W), similar as if we had evidence on all background variables in an FCM.*

Note that for the M in Example 5.1, the approximate counterfactual does not help much. It can be calculated as

$$\tilde{p}(Y_{\mathrm{do}\, X=1} = 0|X = 0, Y = 0) = p(Y = 0|\mathrm{do}\,(x)) = p(Y = 0) = \frac{1}{2}$$

As is easy to see, this implies the KL divergence between $\tilde{p}(Y_{\mathrm{do}\, X=1}|X = 0, Y = 0)$ and the true $p(Y_{\mathrm{do}\, X=1}|X = 0, Y = 0)$ under both, M_0 and M_0' of Example 5.1, to be 1. This KL divergence coincides with the upper bound to the KL divergence in Proposition 5.1, since $H(X, Y|Y) = 1$ in Example 5.1.

The practical meaningfulness of the approximate counterfactual probability, in particular for decision making, remains subject to debate. We will briefly comment on it in Remark 5.1 below.

[3]Note that, if we chose the set Z in Proposition C.1 such that it is as "close" (in the causal diagram) to Y as possible, this could yield better approximations than simply letting Z be the root nodes, as done in $\bar{p}(Y_{\mathrm{do}\, X=x} = y|e)$. We leave this as a question for future work.

5.3.2. Approximate integration of causal knowledge

The following result will be important for Section 5.5 since it can be used to preserve some amount of *privacy*. Consider random variables $C, X_0, \ldots, X_K, Z$. A typical causal structure which satisfies the assumptions we make below is depicted in Figure 5.5 on page 125. Here we introduce what can be seen as an approximation to "transportability", as introduced by Bareinboim and Pearl [2012], in the following simple case: we would like to know $p(z)$, we do know the mechanism $p(z|x_0, \ldots, x_K)$ plus, from a different source, $p(x_k, c)$ for all k, but we do *not* know $p(x_0, \ldots, x_K)$. Define the approximation

$$\bar{p}(z) := \sum_{x_0,\ldots,x_K,c} p(z|x_0, \ldots, x_K) \prod_k p(x_k|c)p(c). \tag{5.4}$$

Proposition 5.2. *If* $Z \perp\!\!\!\perp C|X_0, \ldots, X_K$, *then* $\mathrm{D}(p(Z)\|\bar{p}(Z)) \leq \sum_k \mathrm{H}(X_k|C)$.

Note that based on the proposition, again, we can *know* how wrong the approximation is at most, using only the available information $p(x_k|c), p(c)$. A proof (again using monotonicity of the KL divergence and properties of entropy), can be found in Section C.2.

Example 5.3. *To get an intuition, consider the case that all* X_k *are fully determined by* C*: then* $\bar{p}(z)$ *and* $p(z)$ *coincide, which is reflected by* $\sum_k \mathrm{H}(X_k|C)$ *being 0. As already mentioned, an example of a causal model which implies the condition of the proposition is depicted in Figure 5.5 on page 125.*

While here we apply the proposition for a predictability-privacy problem in Section 5.5, it is more generally applicable where joint distributions are not available. In particular, while in Section 5.5 we will focus on approximate integration for *privacy* reasons, an even more frequent reason may be that only (insufficient) marginals are *known*. Keep in mind that stronger statements on the set of possible $p(z)$ under the available information may exist, e.g., based on ideas in Balke and Pearl [1994].

5.4. Problem 1 – models for control and debugging – and our approach

We start with the problem statement (Section 5.4.1), followed by our approach (Section 5.4.2). Then we illustrate our approach in detail based on several toy scenarios and discuss advantages over previous work (Section 5.4.3).

5.4.1. Problem statement

Cloud computing involves technical systems of the highest complexity, which have to be controlled and debugged, ideally in a (semi-)automatic way. More specifically, the *control problem* can be stated as follows: During the operation of a cloud server many "decisions" automatically have to be made regarding how resources, such as complete computers, or parts, such as CPU time, are allocated among the various applications or virtual machines (VMs) of clients. The goal is to optimize this automatic decision making, based on some given utility function, encoding, e.g., energy consumption, guarantees given to customers, or simply profit.

The *(performance) debugging problem* (closely related to "performance attribution") can be formulated as follows: the general goal is to understand which component of a system contributes to what extent to the measured performance. Based on this, it can be decided which components have to be modified to improve the performance. To give an example, a cloud computing client may wonder whether the high latency of his web server is caused from concurrent programs within his VM (which he could directly intervene), or by other, concurrent VMs on the same physical cloud sever. We will come back to this example in Section 5.4.3, where we address a toy scenario, as well as Section 5.6.2, where we give an example of a preliminary but realistic causal model that can help in such a situation. Note that we presently focus on debugging for *individual observations*, i.e., on the unit-level (see also Section 2.1.5).

Usually, plenty of heterogeneous knowledge and data is available about the involved systems: expert knowledge, formal program code and system specifications (often con-

taining non-causal associational knowledge), data from the very system or similar ones, and data from sandbox experiments.

5.4.2. Outline of an approach

We now sketch several steps of a unified approach based on causal models, which can potentially help to address the control and the debugging problem. In what follows, we will refer to the cloud system "in production", i.e., the fully configured system with a specific set of applications, as the *"target system"*. Note that, depending on the specific setup, some steps may be canceled.

5.4.2.1. Step 1A: inference of causal diagram and some mechanisms

Given: the various information sources described below.

Procedure: Keep in mind that the inference procedure we describe here is usually not based on the target system itself, since some details of it (such as the specific VMs running on it) are varying quickly, but instead on *past* experience with other systems of equal or similar configuration. In particular, usually not all details of the target system are known during this step, so that some mechanisms stay underdetermined, but can be inferred later during Step 1B. As usual, the main sources for causal inference are randomized interventional experiments, observational data (deploying observational causal inference methods 2.1.3.2) and expert knowledge. A necessary condition to harness the first two sources is the decision about - and performance of - *measurements* of the system, for which we propose to use tools discussed by Carata et al. [2014], Snee et al. [2015].

Note the important fact that many aspects of computer systems (hardware and software) are - *by design* - modular, i.e., separable into individually manipulable input-output mechanisms, which is a central assumption in causal models, as we mentioned in Section 2.1.5. To give a simple example: to see if erroneous behavior is caused by the network, one can unplug the network cable and check if the error occurs nonetheless – a procedure which generally would not change any other mechanism, such as the CPU or keyboard. Furthermore, the same (or similar) mechanisms occur in different systems,

which is very helpful for extrapolation from experiments. Note that there is an additional source of information which is specific to computer systems: a lot of knowledge about non-causal associations, such as which program calls which other program during execution, is available, often in a well-formatted way (e.g. program code or system architecture specifications). Such information could be translated into hypotheses on causal association (or be used for measurement selection), in a (semi-)automatic way.

The output of this procedure is a causal diagram G of the target system, together with those mechanisms, i.e., conditionals in the causal model M of the target system, which can be inferred based on past experience. For those mechanisms which cannot be known based on past experience, but only when the target system is revealed (e.g., the specific VMs running on it), but which *cannot be explored directly on the target system* either (since tentative configurations may violate contracts with clients [Chiang et al., 2014, Zheng et al., 2009]), we discuss the integration of sandbox experiments in Step 1B below, which should then complete the causal model M.

5.4.2.2. Step 1B: design and integration of sandbox experiments

Given: an additional cloud system, the *"experimental system"*, equivalent in hardware to the target system, the causal diagram G of the target system, some variable X (e.g. performance of some VM) in G, and the identity (e.g., VM) but not all properties of the mechanism that produces X, and whose unknown properties should be inferred during the experiment.

Procedure: The knowledge of G allows to integrate sandbox experiments in a principled way:

1. Derive all direct influences of X from G, i.e., the parents PA_X (which could include resources such as CPU time or size of requests received from the internet).
2. *Design* the sandbox experiment on the experimental system such that (1) the experimental system has the same conditional $p(x|\mathrm{pa}_X)$ as mechanism for X (e.g., by simply running the same VM on the experimental system as is planned to run on the target system) and (2) all variables in PA_X are randomly varied.

3. Based on the gathered data, regress X on PA_X and plug the inferred conditional $p(x|pa_X) = p(x|\mathrm{do}\, pa_X)$ as mechanism for X into M. This is possible since all parents of X were "intervened" and regressed upon.

Without going further into detail, it needs to be mentioned that the transfer of the conditional between experimental and target system can be seen as a simple example of "transportation" of causal relations as defined by Pearl and Bareinboim [2011b].

5.4.2.3. Step 1C: control

Given: causal model M of the target system, some utility u, which is variable in M or a function of one or several variables in M, and some variable X (e.g. concurrent workload, CPU time, network bandwidth) in M, which should be controlled such as to optimize u (or $p(u)$).

Procedure: As M predicts the effect on u of modifying any of its mechanisms, it can be used to find the mechanism, or "policy", $p(x|\mathrm{pa}_X) = \pi(x|pa_X)$, which maximizes u.

5.4.2.4. Step 1D: observation-level performance debugging

Given: causal model M of the target system, a variable Y in M that measures the performance, a performance debugging query Q, and an (individual) observation $Y = y, F = f$, where F contains all observables besides Y. (Since we move on the level of individual observations instead of populations, we term this step "observation-level performance debugging".)

Procedure: For the performance debugging query Q, we assume the following form: "In the current situation, would it improve performance Y from the current y to y', if we would set X to x', given side information $F = f$?" The side information f may contain an observation x of X. Stated this way, it seems natural to translate this query into a query for the structural counterfactual probability $p(Y_{\mathrm{do}\, X=x'} = y'|y, f)$.[4] Then, based on Section 5.3.1 and in particular Proposition 5.1, we can calculate the approximate

[4]Clearly, there are other ways to formalize attribution and debugging.

answer $\tilde{p}(Y_{\mathrm{do}\, X=x'} = y'|y, f)$ from the GCM M, if $\mathrm{H}(E|Z)$ is small, where Z is a set of root nodes.

Remark 5.1 (The value of (approximate) counterfactuals for performance debugging)**.** *A remark is due regarding the notion of a counterfactual and its application to performance debugging. In the narrow sense, a counterfactual statement is always a statement about the past and so it is neither falsifiable, nor can it help for any (falsifiable recommendations regarding) future decision.*

In contrast, here we have in mind a broader notion of a counterfactual: a situation where one observes a system with a poor performance and asks how the performance could be debugged when the system remains in the "same" *state, or visits the same or similar states again. (In the language of causal models, "state" means the tuple of background variables.) This question is relevant in situations where the debugging action can be performed quickly after the observation of poor performance, and where ones assumes that the* state changes comparably slowly*, i.e., the state varies smoothly with time.*[5] *Alternatively, the question can be relevant if one has a good "subjective" judgement about the similarity of the state between two points in time – if the judgement is based on pbjective observables though, a non-counterfactual form of reasoning may be more appropriate.*

Situations where counterfactual reasoning may be useful arise, in particular, whenever one does not assume to "know" the population-level distribution of the state well enough (but one beliefs in the structural equations), for instance, because it varies with time, and instead one wants to reason on the observation-level, i.e., unit-level. Because on the population-level, there are better ways for decision making than counterfactual reasoning, see Step 1C.

We propose one way to formalize performance debugging questions, and to answer them, based on one possible formalization of counterfactual probabilities proposed by Pearl [2000]. It remains an open question whether there are better formalizations than ours for the debugging questions we consider, and whether the general notion of a counterfactual probability, as well as its formalization by Pearl [2000], are sensible. For a discussion, see also [Peters et al., 2017].

[5] It seems that a more thorough analysis of this argument might be fruitful, as it could theoretically justify the frequent usage of counterfactual reasoning in everyday life. We leave this to future work.

Note that an additional issue, which we are not able to settle here, is how close our approximation of a counterfactual comes to the true counterfactual in practice.

5.4.3. Application to toy scenarios and discussion of potential advantages over previous approaches

For researchers familiar with causal inference, some of the steps described above may seem trivial. However, all current approaches to the described problems we are aware of are lacking a *principled* (formal) language, with concepts such as causal sufficiency, for such things as integration of sandbox experiments and performance debugging.

We will now give toy examples to make the approach outlined in Step 1B through Step 1D more concrete, and simultaneously show the advantages of our approach based on causal models over some previous approaches. (For examples of applications of Step 1A, see Sections 5.6.1 and 5.6.2.) Keep in mind that, clearly, the approach we outlined does not completely solve the problem: the inference of knowledge it relies on remains a challenge as with all other approaches. However, our approach may be less prone to errors and more data-efficient.

- **Step 1B:** Integrating sandbox experiments without a principled approach [Chiang et al., 2014, Zheng et al., 2009], can lead to errors: e.g., if not all parents (direct causes) of a variable X are varied during the experiment and regressed upon afterwards or, say, X is regressed on its causal children. Any methodology that does not include reasoning about concepts such as causal effect, causal sufficiency or randomization is prone to such mistakes. Let us give a toy example of how our approach works for sandbox experiments, and how other approaches can go wrong in terms of variation and regression.

  Example 5.4 (Design and integration of sandbox experiments, and possible mistakes)**.** *Imagine we are the cloud provider and we want to decide whether we can put some VM A on some cloud server, where already other concurrent VMs are running. Let $L \in \{0, 1\}$ denote the performance of (the main application running inside) A, with $L = 0$ denoting good, and $L = 1$ bad performance. For instance, L could denote some latency. Assume that Figure 5.2 depicts the correct causal*

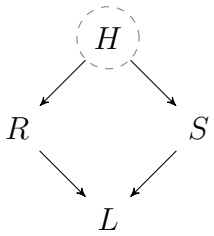


Figure 5.2.: Causal diagram when running VM A on the target system. Not varying or regressing on S during the sandbox experiment can lead to wrong predictions of performance L on the target system, especially when some hidden source H (say internet users) introduces strong correlations between R and S.

DAG of the target system, i.e., when A would be running on the mentioned cloud sever. In particular, the performance depends on two factors, say amount of requests $R \in \{0, 1\}$ *coming into A from the internet, on the one hand, and usage* $S \in \{0, 1\}$ *of the CPU of the cloud sever by the concurrent VMs, on the other, where 0 stands for "low" and 1 for "high". And in turn,* R, S *depend on H which may denote the state of the internet users, which send requests to A but potentially also to concurrent VMs and therefore also influence S. (Alternatively, H could denote a* parameter *for the behaviour of the internet users, i.e., for the distribution of their states.)*

Assume the true mechanism underlying L to be

$$L := R \text{ AND } S,$$

where AND *denotes the logical "AND". I.e., the performance is bad iff A has to serve many requests* ($R = 1$) *and at the same time CPU usage by concurrent VMs is high* ($S = 1$). *Furthermore, assume that on the target system, we have* $R \approx S$. *For instance, this could be due to the fact that A and concurrent VMs serve internet users in the same time zone. Additionally, assume* $p(R = 0) = p(S = 0) = \frac{1}{2}$.

Suppose we have inferred the causal DAG in Figure 5.2 based on Step 1A. We now want to infer the mechanism underlying the performance L, so, following Step 1B (taking L as X), we would perform a sandbox experiment where we would vary both, R and S, and afterwards regress on both, R and S. We would correctly infer the mechanism $L := R$ AND S. *Additionally knowing* $p(r, s)$ *(say from previous*

experience, or from reports by the cloud clients) we would correctly predict the probability of bad performance of A on the target system, $p(L = 1)$*, to be*

$$\sum_{r,s} p(L = 1|r, s)p(r, s) = 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = \frac{1}{2}$$

In contrast, without such a principled approach, two things can happen.

If in the sandbox experiment, only R is varied and regressed upon, while S is kept to a constant 0 *(because it was not properly inferred or communicated as an influence factor, or simply because on the experimental system no concurrent VMs are emulated), then* $p(L = 1)$ *would be wrongly predicted as*

$$\sum_{r} p(L = 1|r, 0)p(r) = 0 \cdot \frac{1}{2} + 0 \cdot \frac{1}{2} = 0.$$

And even if in the sandbox experiment, S would be varied according to the correct $p(s)$ *on the target system (e.g., because the concurrent VMs of the target system would be emulated well on the experimental system), but if one would* forget about regressing on S*, then still one would wrongly predict* $p(L = 1)$ *to be*

$$\sum_{r} p(L = 1|r, s)p(r)p(s) = 0 \cdot \frac{1}{4} + 0 \cdot \frac{1}{4} + 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{4} = \frac{1}{4}.$$

Clearly, this was only a simplistic toy example, but to the best of the knowledge of the author, such problems have not been thematized in the literature [Chiang et al., 2014, Zheng et al., 2009] yet.

- **Step 1C:** Causal models provide a principled tool for control of cloud systems that allows to integrate various forms of information, such as results of sandbox experiments obtained in 5.4.2.2. Furthermore, compared to, e.g., [Padala et al., 2009], which is based on adaptive control, an advantage of using causal models is that they allow to encode and integrate knowledge about which mechanisms vary and which stay invariant.

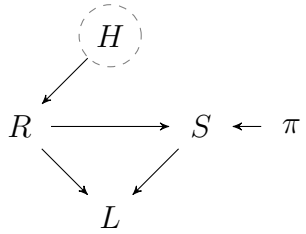


Figure 5.3.: Causal diagram when running A on a system controlled by policy π. It is similar to the system in Figure 5.2, except that now S is influenced by the choice of the policy (as well as the current R which serves as an input to the policy), and therefore we add π to the diagram and draw an arrow to S. Note that handling the policy, which is rather a parameter than a variable, in such a way is similar to the use of so-called "selection diagrams" in Pearl and Bareinboim [2011b], where the mechanisms that vary are marked by special nodes with arrows to them.

Example 5.5 (Control based on causal models)**.** *Consider A, the same VM as in Example 5.4, with performance L. Recall that there we inferred the mechanism for L to be $L := R$ AND S.*

Now assume that we consider a different target system than in Example 5.4, namely, a system that involves a policy $\pi(r|s)$ that controls the amount S of CPU that is occupied by VMs other than A. We depict the causal DAG in Figure 5.3.

Suppose the goal is as follows: keep the probability of poor performance below $\frac{1}{2}$, i.e., $p(L = 1|\pi) \leq \frac{1}{2}$, while allocating as little CPU as possible to A, i.e., minimizing $p(S = 0|\pi)$ (so that more CPU can be used by other VMs). Furthermore, assume $p(R = 0) = \frac{1}{2}$, as in Example 5.4.

Using the causal DAG and "plugging in" our knowledge of the mechanisms, it is easy to see that the optimal policy is $\pi(S = 1|r) = 1$, i.e., always occupy the CPU by other VMs. Because then

$$p(L = 1|\pi) = \sum_{r,s} p(L = 1|r, s)\pi(s|r)p(r) \tag{5.5}$$

$$= p(L = 1|0, 1)\frac{1}{2} + p(L = 1|1, 1)\frac{1}{2} \tag{5.6}$$

$$= 0 + \frac{1}{2} = \frac{1}{2} \tag{5.7}$$

so the goal w.r.t. performance L is still met. This shows how causal models provide a principled tool to integrate sandbox experiments, based on Step 1B, to perform control, as proposed in Step 1C (the S here corresponds to the X there).

Let us mention a potential advantage of control based on causal models in case cloud systems are time-varying. Assume H denotes a parameter for the behavior of the internet users (we indicated this meaning in Example 5.4.2.2). Suppose H varies for some reason, say due to an ad campaign, in an unpredictable way. We know that the behavior of the internet users influences L only via R, since the rest of the cloud system is not affected by the internet. This knowledge is encoded in the causal DAG in Figure 5.3. Based on this, we have

$$p(l|r, s, h) = p(l|r, s).$$

So we have formally reasoned that even if H varies, the mechanism $p(l|r, s)$ stays the same. Hence, to derive the new optimal policy π, all one has to do is to infer the new $p(r)$ and plug it into Equation 5.5 (and optimize for π). Furthermore, we can be certain that we identified *the new system and the new optimal policy (given our assumptions are correct). This sort of reasoning has been analyzed, on a more general level, by Pearl and Bareinboim [2011b] (but they do not apply it to control settings).*

In contrast, approaches to (adaptive) control for cloud computing which are not based on modularity and such reasoning [Padala et al., 2009] may try to infer the complete information, $p(r)$ as well as $p(l|r, s)$, from scratch upon a variation of H, assuming it to be a completely new "environment" (recall Section 2.1.5 where we discussed the connection between causation and modularity). And even if such approaches utilize the invariant $p(l|r, s)$ after a variation of H in one way or another, they are usually missing the language to reason about the identifiability *of the new system (after the variation in H), as we did above based on causal models.*

It needs to be emphasized that here we considered an overly simplistic scenario. In more complex and realistic scenarios, there are much more mechanisms involved that could potentially vary or stay invariant, respectively. See Section 5.6.2 for an example of a causal DAG of a more realistic but still simple cloud system.

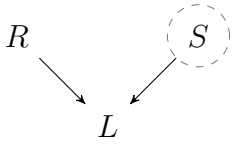


Figure 5.4.: Causal diagram for observation-level performance debugging in a toy setting. S is unobserved, but nonetheless we assume $p(l|r,s)$ to be known, may it be that the provider publishes it, or the client knows it from own (sandbox) experiments.

- **Step 1D:** We now give en example for how observation-level debugging can be performed based on Step 1D. This approach can be seen as complementary to other methods for this problem [Ostrowski et al., 2011], where errors may arise from confusing causation with correlation, or where it is more difficult to integrate heterogeneous knowledge such as sandbox experiments.

  Example 5.6 (Observation-level performance debugging)**.** *Note that, while this is a toy scenario, the assumptions we make in this example regarding what is known/observed and what not are close to realistic [Snee et al., 2015].*

  Similar as in Example 5.4, consider a VM with performance (latency) L running on a cloud system, with $R \in \{0,1\}$ denoting the amount of incoming requests, and $S \in \{0,1\}$ the amount of, say, CPU time allocated to concurrent VMs (0 stands for "low" and 1 for "high"). Here, denote the VM by B. In contrast to Example 5.4, assume the causal DAG depicted in Figure 5.4. Furthermore, let $L \in \{0,1,2,3\}$ and the structural equation for L be given by

  $$L := R + S + U_L, \tag{5.8}$$

  with $p(U_L = 0) = \frac{1}{2}$. Suppose $p(S = R) = p(S = 0) = \frac{1}{2}$, where $p(S = 0) = \frac{1}{2}$ may be seen as encoding some prior belief.

  Now assume that the client whom B belongs to wonders, whether it would improve the latency L to a desired 0 in the current situation where she observes $L = 2, R = 1$, if she decreased the amount of incoming requests to a lower level, i.e., if she set R to 0. (Note that "current situation" can include the nearby future, if the unobserved variables vary comparably slowly, see Remark 5.1.) She does not *observe S due to neither the cloud provider nor other clients publishing this information. This is*

a realistic assumption in cloud computing. Based on Step 1D, she translates this question into a query for the counterfactual probability $p(L_{\mathrm{do}\,R=0} = 0|R = 1, L = 2)$.

Suppose that while S *is not published,* $p(l|r, s)$ *is known, may it be that the provider publishes it, or the client knows it from own (sandbox) experiments. That is,* $p(r), p(s)$ *and* $p(l|r, s)$ *are give, but not the structural Equation 5.8 itself. Now, although the structural Equation 5.8 would be needed to calculate the counterfactual* $p(L_{\mathrm{do}\,R=0} = 0|R = 1, L = 2)$ *exactly (see Example 5.1) she can calculate the* approximate *counterfactual probability defined in Equation 5.2 as*

$$\begin{aligned}
&\tilde{p}(L_{\mathrm{do}\,R=0} = 0|R = 1, L = 2) \\
&= \sum_s p(L = 0|\mathrm{do}\,R = 0, s)p(s|R = 1, L = 2) \\
&= \sum_s p(L = 0|\mathrm{do}\,R = 0, s)p(L = 2|R = 1, s)p(s|R = 1)\frac{1}{p(L = 2|R = 1)} \\
&= \frac{1}{4},
\end{aligned}$$

where we plugged in R, S *for the set of root variables* Z*, which yields* S *as* W*, and as evidence* E *we took* (R, L) *with value* $(1, 2)$*. Based on this, she concludes that the probability that setting* R *to* 0 *helps for decreasing latency* L *to* 0 *is rather small (in the current situation).*

Note that the true counterfactual probability (Equation 5.1 in Section 5.3.1) in this specific case is given by

$$\begin{aligned}
&p(L_{\mathrm{do}\,R=0} = 0|R = 1, L = 2) \\
&= \sum_{u_R,u_S,u_L} p(L = 0|\mathrm{do}\,R = 0, u_R, u_S, u_L)p(u_R, u_S, u_L|R = 1, L = 2) \\
&= \sum_{u_L,s} p(L = 0|\mathrm{do}\,R = 0, s, u_L)p(s, u_L|R = 1, L = 2) \\
&= 0 + p(L = 0|\mathrm{do}\,R = 0, S = 0, U_L = 1)p(S = 0, U_L = 1|R = 1, L = 2) \\
&\quad + p(L = 0|\mathrm{do}\,R = 0, S = 1, U_L = 0)p(S = 1, U_L = 0|R = 1, L = 2) + 0 \\
&= 0,
\end{aligned}$$

which would lead to the even stronger conclusion that setting R *to* 0 *for decreasing*

L to 0 would not work at all.

Note that the upper bound of Proposition 5.1 here takes the value

$$\begin{aligned} \mathrm{H}(R, L|R, S) &= \mathrm{H}(L|R, S) \\ &= \sum_{r,s} p(r, s)\mathrm{H}(L|r, s) \\ &= 1. \end{aligned}$$

Recall that we picked $p(U_L = 0) = \frac{1}{2}$, i.e., rather strong noise. For less noise, the approximation would be even better and the bound smaller.

Note that generally, one could try to *learn* (in the sense of machine learning) things such as how to perform and integrate the experiment [Snee et al., 2015], but one would always have to rely on prior assumptions, which may then be more difficult to encode.

5.5. Problem 2 – cost predictability versus privacy – and our approach

We start with the problem statement (Section 5.5.1), followed by our approach (Section 5.5.2). Then we present a toy example (Section 5.5.3), and some additional remarks (Section 5.5.4).

5.5.1. Problem statement

Here we consider an economical aspect of cloud computing. Currently, one common way for clients to purchase cloud resources from a provider is via an auction mechanism for "spot" (i.e., short-term) resources, which can be described in a simplified way as follows: The customer enters a bid, e.g., for an hour of usage. Once the price determined by the provider (based on supply, demand, and other private factors) drops below the bid, the customer gets the resource, usually as long as her bid exceeds the price (within the hour). This approach has several advantages, in particular for the provider: he can sell resources which are unused but which fluctuate a lot (due to guarantees given to "dedicated" or

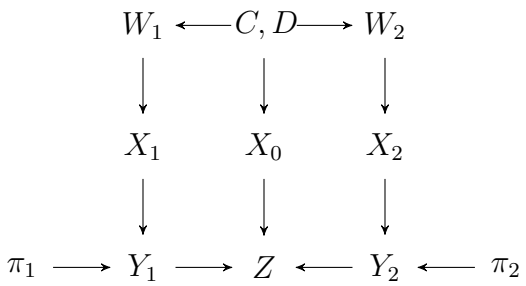


Figure 5.5.: Causal diagram G_2. D is hidden.

"on-demand" customers). But clients can profit as well: the spot resources are usually significantly cheaper than the long-term dedicated resources.

An obvious drawback of spot resources is that this kind of mechanism comes with a high *uncertainty* for the clients: it is hard to tell how the prices will evolve in the future, and, in particular, purchased resources can be terminated in an unforeseeable way, which is, to some extent, due to the *unpredictability of the other clients*. Therefore, if the client does not want to take these risks which can significantly harm his/her business, they often avoid this mechanism.

5.5.2. Sketch of an approach

In what follows, we present a very first step towards addressing the problem based on PCMs. We assume that there is one provider, and clients $1, \ldots, K$. By "stakeholders" we refer to provider and clients together. For each point in time (say, the beginning of an hour), let X_k denote client k's demand for the next hour, Y_k the cloud product that the client buys from the provider, W_k the information based on which the client decides her demand (e.g., hour of the day), which may not always be fully known though, and π_k her policy determining which cloud product Y_k to buy, given her demand X_k. Let X_0 denote the provider's pricing parameter at that time point (which may depend, e.g., on energy costs), and let Z denote the outcome of the provider's mechanism applied to the Y_k. (Generally, Z can include the price as well as say termination of spot resources; for simplicity, let it only denote the cost/price for the moment, which can comprise the indirect costs resulting from loss of visitors through termination.)

We assume the following simple mechanism (which is a simplified version of the auction described above): all clients k always get the product they want, but the subsequent price vector Z varies and is not known in advance. The causal diagram G_2 for the complete causal structure, for the case $K = 2$, is depicted in Figure 5.5. The role of C will be explained below, while D denotes the hidden part of the confounder (C, D).

Our approach to the uncertainty problem, towards more predictable prices and subsequent reduced costs, is based on the idea that clients may not want to share all, but are willing to share some of their information between each other. More specifically, we propose the following two-step procedure which allows the clients to *trade off privacy versus predictability interests*, by jointly picking a variable C such that $p(X_k|C)$ allows an approximate prediction of Z which still preserves some privacy.[6]

5.5.2.1. Step 1A: jointly picking C

First, all stakeholders k pick their candidates for C (possibly based on a given list and some "privacy budget"), balancing their privacy interests against minimizing $\mathrm{H}(X_k|C)$. If the intersection of their candidates is non-empty, they reveal $\mathrm{H}(X_k|C)$ for all k and joint candidates C.[7] They pick the C that minimizes $\sum_k \mathrm{H}(X_k|C)$ to optimize the predictability, based on Proposition 5.2.

5.5.2.2. Step 2B: prediction and individual decision

Now all clients k reveal their $p(x_k|c)$. $p(c)$ is assumed to be common knowledge. Furthermore, all $p(y_k|x_k, \pi_k)$ are either known a priori (based on the possible products the provider offers) or revealed now. The provider reveals $p(z|x_0, y_0, \ldots, y_K)$ and $p(x_0|c)$. Now $\bar{p}(z|\pi_1, \ldots, \pi_K)$ can be calculated, based on Equation 5.4. More specifically, we

[6] An extreme approach would be to directly infer a joint model for all clients from their joint data (i.e., considering all clients as a "single client"). Here we assume that this is not possible, due to heterogeneous data, privacy interests, etc.

[7] If the intersection is empty, the procedure is canceled without result, and the stakeholders proceed in the classical, non-collaborative way.

have

$$\begin{aligned}\bar{p}(z|\pi_1,\ldots,\pi_K) &= \sum_{x_0,\ldots,x_K} p(z|x_0,\ldots,x_K;\pi_1,\ldots,\pi_K)\prod_{k=0}^{K} p(x_k|c)p(c)\\ &= \sum_{x_0,\ldots,x_K}\left(\sum_{y_1,\ldots,y_K} p(z|x_0,y_1,\ldots,y_K)\prod_{k=1}^{K} p(y_k|x_k;\pi_k)\right)\prod_{k=0}^{K} p(x_k|c)p(c)\end{aligned}$$

Then, based on Proposition 5.2, the clients narrow down the set of possible $p(z|\pi_1,\ldots,\pi_K)$ to those for which

$$\mathrm{D}(p(Z|\pi_1,\ldots,\pi_K)\|\bar{p}(Z|\pi_1,\ldots,\pi_K)) \leq \sum_k \mathrm{H}(X_k|C).$$

Based on this constraint on $p(z|\pi_1,\ldots,\pi_K)$, each client k decides on their strategy π_k, e.g., based on game-theoretic considerations.

5.5.3. Application to toy scenario

To illustrate the approach, let us give an example.

Example 5.7. *A cloud provider, Clark, offers to his clients, Alice ($k = 1$) and Bob ($k = 2$), monthly (dedicated) large resources ($Y_k = 2$), rather expensive, or hourly spot small ($Y_k = 0$) and large ($Y_k = 1$) resources, which are usually cheaper. However, if Alice and Bob happen to both order large spot resource for the same hour, the cost for both of them ($[Z]_1, [Z]_2$) is significantly higher than the hourly rate for the monthly large resource, since Clark may have to buy a new resource, or he may have to cancel one of his client's applications, causing the loss of web site visitors. Now assume Alice and Bob, during Step 1A, pick the hourly weather forecast, which is* 0 *for sunny and* 1 *for cloudy, for C, since it is public information anyway that both their web sites are* weather related*: Alice runs a website for* outdoor *activities, Bob one for* indoor *activities, both in the same region. And the remaining uncertainty w.r.t. their demand (X_k being 0 for "small" or 1 for "high"), i.e., $H(X_k|C)$, is small. The causal diagram for this scenario is G_2 depicted in Figure 5.5. Based on this, Alice and Bob can conclude that they will rarely require a large resource at the same time, and they can go for spot resources as their respective (dominant) strategies π_k.*

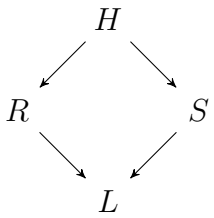


Figure 5.6.: Causal diagram G_1.

5.5.4. Discussion

In some cases, the provider could infer the joint distribution of all X_k, based on past data, which would contain all relevant information. However, the complete system is so complex that it is unlikely to be *stationary*. Note that during each step, already some information is revealed, but this is transparent to the stakeholders. Limitations of our approach are that (1) the clients may not even be willing to reveal their $p(x_k)$, or (2) X_k may not be predictable or the model may be wrong (although humans and organizations usually do plan ahead).

It needs to be emphasized, that here we completely ignore strategic aspects, which can lead to problems in our proposed approach. Such aspects could be analyzed, e.g., based on game theory.

5.6. Experiments

5.6.1. Control and debugging problem on simple but real cloud system

Here we test small parts of our approach in Section 5.4.2 on a very simple, but real cloud system: a physical server running a specific application (a web server) together with some concurrent workload (another web server). The system we consider has the same causal DAG as two of the examples in Section 5.4.3. And while the scenarios are generally similar, the system we consider here is simpler, for experimental purposes.

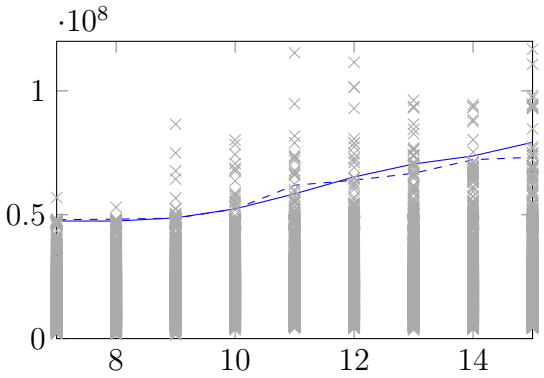


Figure 5.7.: X-axis: Number of simultaneous requests $S = s$. Y-axis: 99th percentile of prediction $\hat{p}(l|\mathrm{do}\, s)$ (dashed blue) is close to 99th percentile (solid blue) of ground truth test data from $p(l|\mathrm{do}\, s)$ (subsample in gray).

A source H keeps sending simultaneous request to application and concurrent workload (drawn from a multivariate correlated Poisson distribution), of which R are received by the application and S by the concurrent workload. Then, for each request, the latency (performance) of the application is measured in nanoseconds by L.

We examine how well Step 1A works. First, we infer the causal diagram G_1 depicted in Figure 5.6, as well as an estimate of $p(r, s, l)$ from observational samples of the system, based on Step 1A, and together denote them by (incomplete) M_1. Then, from M_1, using back-door adjustment Pearl [2000], we derive a prediction $\hat{p}(l|\mathrm{do}\, s)$ for $p(l|\mathrm{do}\, s)$. Besides Step 1A, this tests the applicability of Step 1C, when thinking of a simple controller that outputs a constant for S (e.g. by putting the application on another machine with such a concurrent workload), as well as Step 1D which relies on post-interventional distributions (of an updated model, though). The outcome is depicted in Figure 5.7, where we use the 99th percentile as statistic, which is common in cloud computing. It shows that the prediction is close to the ground truth test data, both in magnitude and in trend.

5.6.2. Example of a more realistic cloud system

The experiments in the previous Section 5.6.1 were performed on an overly simplistic system. Here we want to give an example of a preliminary, partial causal model (causal DAG plus some knowledge on the mechanisms, e.g., additivity) of a more *realistic* system

to which our approach in Section 5.4.2, in particular the performance debugging in Step 1D, is meant to be applied. Note that this is merely for illustration purposes, we do *not test* any hypothesis here.[8]

We consider a cloud sever running serveral VMs. We focus on one specific VM, call it A for the moment. Inside the VM A, a web sever B (more specifically: "lighttpd") runs. We consider the following observed and hidden variables, among others, measured inside and outside A[9]:

- "req_size": size of the file requested by an internet user from the web server B;
- "local_load": resource-consuming activity of other applications in A, besides B;
- "concurrent_vm_count": number of VMs running concurrently with A on the physical sever (outside A);
- "srv_lat": latency of the web server B, which can be seen as part of the *objective* which needs to be minimized.

We depict the partial causal model in Figure 5.8. It is taken from Carata [2016], who also gives descriptions for all other variables in the figure not discussed here. Note that this is a model of an *experimental* system, while on a system in production, some variables, such as "local_load" and "concurrent_vm_count", could be influences by a (hidden) common cause, similar to H in the previous experiment in Section 5.6.1.

This model was inferred as described in Step 1A, in an iterative and sequential way, based on non-causal associational knowledge about the program execution structure (known from the program code) as well as the general system architecture, further expert knowledge, and independence tests on sampled data. As can be seen, often the integrated knowledge allows to draw conclusions on the *additivity* of mechanisms, which can be based on the fact the runtime of one program essentially is the sum of the runtimes of its subroutines.

[8] The inference of the causal model of – and the application of our approach to – such a complex system turned out to be more difficult than expected. Therefore, no evaluation of our approach applied to this system can be reported at this stage.

[9] In cloud computing, it is important to distinguish between inside and outside of A, since, for privacy reasons, often only things inside A can be known to the client that A belongs to.

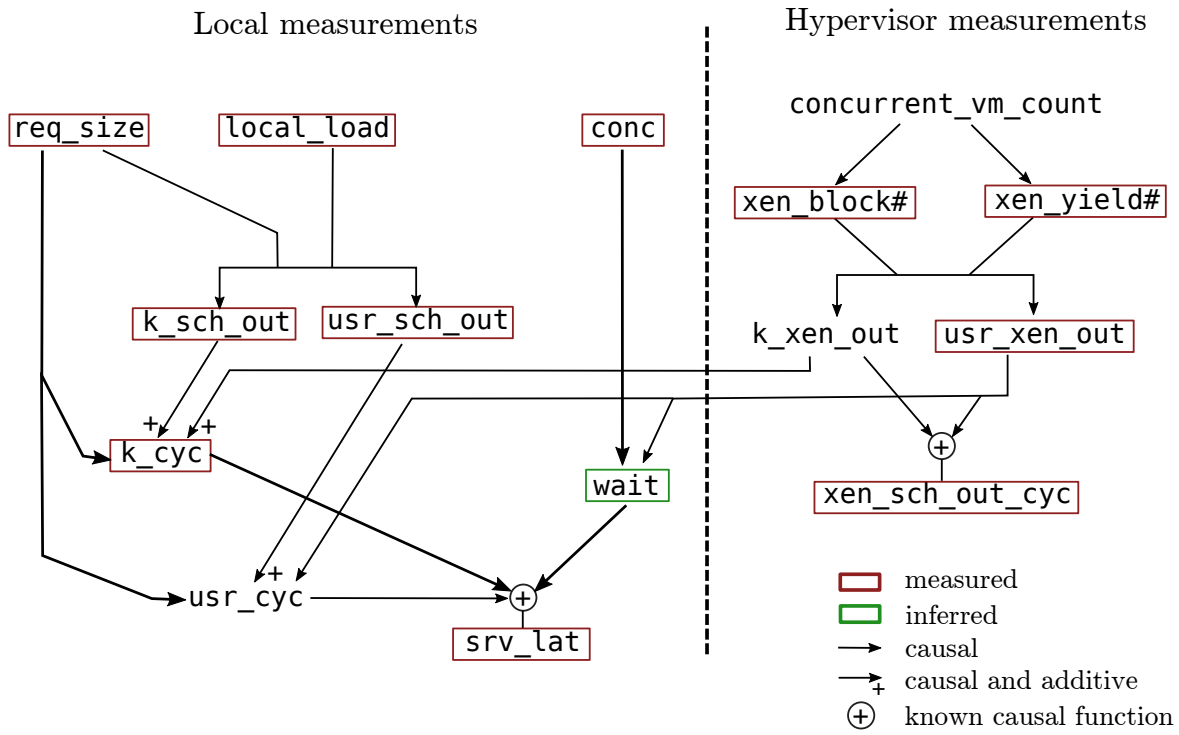


Figure 5.8.: Example of a preliminary causal DAG of a cloud system. Variables on the left side are measured within the VM A, that runs together with other VMs on the cloud server. The right side contains measurements outside the VM (the "hypervisor" is the program that is responsible for allocating the cloud server's resources among the VMs). The objective is to minimize the latency of some web server B running in A, denoted by "srv_lat", while keeping utilization by other VMs, denoted by "concurrent_vm_count", as high as possible. Possible *manipulations* include reducing the workload within the VM A, denoted by "local_load", versus changing the number of concurrent VMs. If the causal model is good, it can help to pick the optimal manipulations. The figure is taken from [Carata, 2016] which also gives descriptions of the remaining variables not described here.

Such a model could help for decision making in various ways, for instance for the performance debugging problem mentioned in Sections 5.4.1 and 5.4.3: A cloud client, the owner of A, may observe a high latency ("srv_lat") of his web server B, together with some other variables. He wonders if, in this situation, the high latency is caused by other programs within his VM A ("local_load"), by other, concurrent VMs ("concurrent_vm_count") running on the same physical cloud sever, or simply by large requests ("req_size") coming in at that moment. Based on this, he could conclude whether he should intervene on "local_load", which may be the simplest, or rather intervene on "concurrent_vm_count" say by changing to another cloud product, such as a dedicated server, which may be more expensive.

5.6.3. Predictability-privacy problem on simulated data

For our approach in Section 5.5.2 to work, $\bar{p}(z)$ has to approximate $p(z)$ reasonably well. Here we examine to what extent this is the case in a simulated version of the toy example in Section 5.5.3, additionally testing how tight the bound in Proposition 5.2 is. Compared to the toy example, we restrict ourselves to spot resources, i.e., $\{0, 1\}$ for Y_k, and assume the following specific mechanisms: The policy π_k is for both to simply purchase their demand ($Y_k := X_k$), Clark's pricing is "cheap" ($Z = 0$) versus "very expensive for one of them since both want large" ($Z = 1$), in particular $Z := Y_1$ AND Y_2. Furthermore,

$$X_k := C \text{ XOR } D \text{ XOR } N_{X_k},$$

where D is some confounder which Alice and Bob do not want to reveal.

Now for "each" $0 \leq r \leq 0.5$, we draw 1000 samples of $C \sim \mathit{Bernoulli}(0.5 - r)$, $D \sim \mathit{Bernoulli}(r)$ to find out how wrong $\bar{p}(z)$ gets when increasing the confounder D that is not revealed or adjusted for, and $N_{X_k} \sim \mathit{Bernoulli}(0.2 - 0.2r)$ (to also examine a little variation in the noise strength). The outcome is depicted in Figure 5.9. It shows that $\bar{p}(z)$ is a good estimate in this simple setting (which is also due to the fact that already $p(x_1), p(x_2)$ alone reveal something about $p(x_1, x_2)$). It also shows that (in this setting), the bound from Proposition 5.2 may be improvable, as the dashed red line is far away from the solid red line.

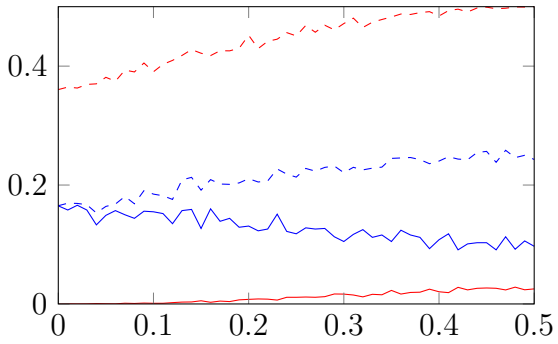


Figure 5.9.: X-axis: Parameter r (the higher, the more influence from D). Y-axis: $\bar{p}(z)$ (dashed blue) is close to $p(z)$ (solid blue) even when D gets strong and C weakens; $\frac{1}{4}\mathrm{D}(p(Z)\|\bar{p}(Z))$ (solid red), $\frac{1}{4}\sum_k \mathrm{H}(X_x|C)$ (dashed red).

5.7. Related work

Regarding Section 5.3, approximations to non-identifiable quantities in causal models were examined by Balke and Pearl [1994]. While their technique does not seem directly applicable to the setup of Proposition 5.1, it may allow to derive stronger statements, i.e., further narrowing down the set of possible $p(z)$, than Proposition 5.2, which could be examined in future work. We discussed some related work for Section 5.4, i.e., the control and debugging problem, in Section 5.4.3. Additionally, maybe the work closest to our investigation in that section is [Lemeire et al., 2007], which suggests to use causal models for performance modeling of programs, but does not consider counterfactuals, or more complex computing systems. Generally, the utilization of modularity based on PCMs in that section is strongly inspired by the theory of "transportability" of causal relations developed by Pearl and Bareinboim [2011b], however, that theory has not been applied to (cloud) computing problems so far. The relation between causality and control is also considered in [Bottou et al., 2013]. Regarding Section 5.5, [Angel et al., 2014] can be seen as related in that they allow the provider to hide their exact costs while still making some information of the costs available to others. The work [McSherry and Talwar, 2007] investigates privacy-preserving mechanisms, but does not consider the integration of the revealed information to an (estimate) of a causal model.

5.8. Conclusions of this chapter

This chapter assayed how causal inference – in principle – can help with technological and economical problems in cloud computing. Guided by these problems, we presented two theoretical results for approximate causal inference, and reported initial experimental results. The application of causal inference in this domain is, to the best of our knowledge, the first of its kind. We believe the potential in this area is very significant, both for applications and for methodological work. Problems in computing systems rarely fit the classical settings that machine learning excels at.

In particular, for issues such as integration of sandbox experiments, (formally) reasoning about concepts such as causation, causal sufficiency or randomization seems crucial, and methodology which neglects this, such as classical machine learning, may be prone to errors. Another concept which plays an important role in causal modeling (but, of course, also in some other areas) is that of identifiability, which helps to "critically" reason about what can and what cannot be inferred based on the given. We used it for the control problem for cases that only some "modules" of the system vary.

A causal perspective with its focus on predicting the effect of interventions may be a crucial component in future developments, ideally combined with aspects of game theory and mechanism design, to extend our approach for the predictability-privacy trade-off. Next steps would involve extending the experimtens on real cloud systems, such as the system for which a preliminary model was derived in Section 5.6.2, and based on this, advancing the approach we sketched in Section 5.4.

Chapter 6.

Conclusions

The goal of this investigation was to better understand causation; to develop methods for inferring probabilistic causal models (PCMs) in realistic scenarios, integrating the vast amounts of empirical data available today as well as high-order knowledge, accepting approximations instead of unique identification; and to use PCMs for informed decision making in technical, economical and natural systems. In this conclusion, we briefly want to discuss some accomplishments and limitations, starting with this thesis in particular, and ending with PCMs in general.

6.1. Conclusions on individual chapters

We began our main investigation in Chapter 3 with a systematic analysis of the problem of hidden confounding in time series, in terms of theory and practical methods, assuming the model of a vector autoregressive (VAR) processes. We showed how, under rather weak additional assumptions, the true causal structure is uniquely or approximately identifiable despite hidden confounding, owing to the integration of the temporal structure. In spite of VAR processes being used so frequently as a model, a clear limitation of our investigation is the assumption of linearity. But our results may also serve as a basis when relaxing this assumption.

In Chapter 4, we tried to formalize the intuition that there are structures and prior knowledge *beyond* perfect experiments (and purely observational studies) that can be

integrated for causal inference and subsequent decision making. Roughly speaking, we translated settings into a degree of "deviation from a perfect experiment" (i.e., strength of hidden confounding), and showed how this degree of deviation can mathematically imply approximations to the true causal effect (the less deviation, the better the constraints on the true causal effect). A drawback of this contribution is that there may be more principled approaches to the problem. Furthermore, we did not evaluate our approach on real data.

Chapter 5 was driven by decision making problems in cloud computing. We performed a first investigation of how causal models can help to address key challenges in this field. We established two theoretical results about approximative causal reasoning, with one of them trying to make counterfactual reasoning based on PCMs more accessible for debugging of cloud computing systems and real world scenarios in general. A limitation of our investigation was that realistic systems are much more involved compared to the toy settings we analyzed, and so we were only able to report initial results of a small part of our method on a simple real systems. In particular, a more thorough investigation of the practicability of counterfactual reasoning for decision making (possibly based on assuming a smooth state process) as well as of approximate integration of causal knowledge remains to be performed.

6.2. A broader view on this thesis

Let us make some more general remarks about this thesis. Owing to the involvedness of the concept of causation, we tried to cover many aspects, from inference on a rather abstract level to more specific applications. (And in the background chapter we also discussed the meaning of causation, quoting several definitions, commenting on difficulties, and putting the concept of causation into context – instead of just sticking to a standard treatment based on say PCMs and interventions which may swipe conceptual difficulties under the rug.) A disadvantage of this rather broad coverage, based on the fact that any investigation is selective, is that some topics were only touched superficially, i.e, some parts of this thesis contain initial ideas rather than completed contributions.

Also from the perspective of general research practice this thesis covered a rather broad

range, form parts that are more based on *"exploiting"* beaten tracks but may be lacking the connection to a specific real world problem, to parts that try to *"explore"* and push the boundaries of established frameworks towards relevant directions. Chapter 3, with its concrete mathematical contributions within the frameworks of VAR processes, Granger causality and PCMs can be seen as en example of "exploitation" – although it needs to be emphasized that in spite of the practical relevance of hidden confounding in time series, it is surprising how little this issue has been theoretically analyzed so far. An example of "exploration" can be seen in Chapter 5, where we tried to establish a rather new application field for causal modeling, based on formalizing relevant questions in terms of causal modeling language, establishing initial mathematical results to answer them approximately, seeking to overcome limitations of previous approaches which are based on modeling correlation rather than causation.

6.3. Causal models in this thesis and beyond

We also want to make some broader comments on accomplishments and limitations of PCMs, while still keeping in mind their role in this thesis.

The language of PCMs constitutes a powerful tool for studying causal inference with *mathematical rigor*. For instance, it helps to precisely phrase the problem of hidden confounding in times series, which formed the starting point for Chapter 3. Generally, a great advantage of PCMs are their ability to *formally reconcile observational and experimental settings*. A downside of the formal rigor, as with any formal language and axiomatic approach, is that it can delude into studying models, which are sometimes arbitrary and meaningless, instead of studying reality itself. It is also not clear, to what extent formal methods can lead towards a genuinely deeper causal *understanding* of the world, it seems that it always depends on how well an informal thought can be translated into a formal model. Additionally, it is worth mentioning that PCMs focus on the "influence" aspect of causation, and neglect – to some extent – other ways in which the concept of causation is applied. In this sense, it may be said that PCMs *help for reasoning about causation on the level of correlations and interventions*, while they are of limited help on other levels.

The rigor of PCMs also reveals how strong *assumptions and "inputs"* are necessary for causal inference to lead to meaningful "outputs". For instance, the back-door criterion by Pearl [2000] may work without any underlying assumptions, but it requires to know most of the underlying causal DAG. To know the causal DAG though, one may have to perform randomized experiments first, which in turn would render causal reasoning pointless. In this thesis, building on previous work in this direction, we tried to overcome this problem by *integrating* the rich conceptual context in which causation is embedded in our world – *a priori and a posteriori* – to come up with "inputs" to causal inference that are easier to obtain but still relevant. Examples include temporal knowledge in Chapter 3, kowledge about partial compliance in Chapter 4 and, to a limited extent, specifications of engineered systems in Chapter 5. PCMs proved a powerful language to express and reason about such "alternative inputs" and allowed us to in fact establish several theoretical results in this direction, which we complemented with practical methods to some extent. But clearly, the results are just small steps.

Regarding *decision making* based on PCMs the following can be said: Within the inference process that leads to a hypothesis class or Bayesian prior that relates past observations with outcomes of future actions, the language of PCMs seems to help to make steps more explicit and give them a semantic. For instance, causal reasoning helped to integrate sandbox experiments in Sections 5.4.2 and 5.4.3. However, in concrete cases, often the difficulty of say coming up with a causal DAG – in particular the conditional independences it entails – remains.

There are other limitations of approaches based on PCMs, including this thesis. For instance, often one wants to *find* the factors that cause say an undesired situation which one want to change, such as low performance of computer systems. But it is only *after a candidate factor was found* and measured, that one can *use PCMs* to first verify or falsify the candidate factor, and then inform decision making. Related to this is the problem of defining meaningful variables, which is completely external to PCMs.

All in all it can be said that causal inference is a difficult topic – which is worth studying though, in light of the importance of understanding causes and predicting effects in this world.

Appendix A.

Proofs and detailed algorithm for Chapter 3

Here we present proofs for Chapter 3, as well as an elaboration of Algorithm 1 in that chapter.

Parts of this chapter are based on the appendix of [Geiger et al., 2015a].

A.1. Proofs for Section 3.5

For this section keep in mind the definitions of W, X, Z, N, N^X, N^Z and A, B, C, D, E from Section 3.4.1 as well as M_1 from Section 3.5.

Proof of Lemma 3.1. The case $K = K_X$ is obvious, so we only prove the case $K > K_X$.

In particular, keep in mind that

$$\begin{pmatrix} X_t \\ Z_t \end{pmatrix} = A \begin{pmatrix} X_{t-1} \\ Z_{t-1} \end{pmatrix} + N_t,$$

and

$$A = \begin{pmatrix} B & C \\ D & E \end{pmatrix}. \tag{A.1}$$

Hence based on

$$\begin{pmatrix} X_t \\ Z_t \end{pmatrix} = A^2 \begin{pmatrix} X_{t-2} \\ Z_{t-2} \end{pmatrix} + AN_{t-1} + N_t,$$

and

$$A^2 = \begin{pmatrix} B^2 + CD & BC + CE \\ DB + ED & DC + E^2 \end{pmatrix}.$$

we get

$$X_t = (B^2 + CD)X_{t-2} + (BC + CE)Z_{t-2} + BN_{t-1}^X + CN_{t-1}^Z + N_t^X, \tag{A.2}$$

$$X_{t-1} = BX_{t-2} + CZ_{t-2} + N_{t-1}^X. \tag{A.3}$$

Based on the definition of the generalized residual $R_t(U_1, U_2)$ in Section 3.5 and equations (A.2) and (A.3), we have

$$\begin{aligned}
&R_t(U_1, U_2) \\
&= X_t - U_1 X_{t-1} - U_2 X_{t-2} \\
&= (B^2 + CD)X_{t-2} + (BC + CE)Z_{t-2} + BN_{t-1}^X + CN_{t-1}^Z + N_t^X \\
&\quad - U_1(BX_{t-2} + CZ_{t-2} + N_{t-1}^X) - U_2 X_{t-2} \\
&= (B^2 + CD - U_1 B - U_2)X_{t-2} + (BC + CE - U_1 C)Z_{t-2} \\
&\quad + (B - U_1)N_{t-1}^X + CN_{t-1}^Z + N_t^X.
\end{aligned}$$

□

Proof of Lemma 3.2. Equation (3.5) together with Equation (3.4) implies

$$R_t(U_1, U_2) = (B - U_1)N_{t-1}^X + CN_{t-1}^Z + N_t^X.$$

Based on $\|A\| < 1$, we have [Lütkepohl, 2006]

$$\begin{pmatrix} X_t \\ Z_t \end{pmatrix} = W_t = \sum_{i=0}^{\infty} A^i N_{t-i} = \sum_{i=0}^{\infty} A^i \begin{pmatrix} N_{t-i}^X \\ N_{t-i}^Z \end{pmatrix}.$$

This implies that

$$(X_{t-2-j})_{j=0}^{\infty} \perp\!\!\!\perp N_{t-1}^X, N_{t-1}^Z, N_t^X.$$

□

Proof of Lemma 3.3. Keep in mind that

$$\begin{aligned} M_1 &= \mathbb{E}\left[\begin{pmatrix} X_t \\ Z_t \end{pmatrix}(X_t^T, X_{t-1}^T)\right] \\ &= \begin{pmatrix} \mathbb{E}[X_t X_t^T] & \mathbb{E}[X_t X_{t-1}^T] \\ \mathbb{E}[Z_t X_t^T] & \mathbb{E}[Z_t X_{t-1}^T] \end{pmatrix}. \end{aligned}$$

Based on Equation (3.4), we have for $j = 0, 1$

$$\begin{aligned} 0 &= \mathrm{cov}(R_t(U_1, U_2), X_{t-2-j}) \\ &= (B^2 + CD - U_1 B - U_2)\mathrm{cov}(X_{t-2}, X_{t-2-j}) \qquad \text{(A.4)} \\ &\quad + (BC + CE - U_1 C)\mathrm{cov}(Z_{t-2}, X_{t-2-j}) \\ &= (B^2 + CD - U_1 B - U_2)\mathbb{E}[X_t X_{t-j}^T] + (BC + CE - U_1 C)\mathbb{E}[Z_t X_{t-j}^T]. \qquad \text{(A.5)} \end{aligned}$$

We can write Equation (A.5) as the following system of linear equations

$$\Big(B^2 + CD - U_1 B - U_2, BC + CE - U_1 C\Big) \begin{pmatrix} \mathbb{E}[X_t X_t^T] & \mathbb{E}[X_t X_{t-1}^T] \\ \mathbb{E}[Z_t X_t^T] & \mathbb{E}[Z_t X_{t-1}^T] \end{pmatrix} = 0,$$

that is

$$\Big(B^2 + CD - U_1 B - U_2, BC + CE - U_1 C\Big) M_1 = 0.$$

Since we assumed that M_1 has full rank, we can conclude

$$B^2 + CD - U_1 B - U_2 = 0 \quad \wedge \quad BC + CE - U_1 C = 0.$$

□

Proof of Lemma 3.4. C is a $K_X \times K_Z$ matrix of full rank, with $K_Z \leq K_X$, hence C has full row rank. Hence $\begin{pmatrix} B & C \\ \mathbf{I} & 0 \end{pmatrix}$ has full row rank. Thus, there is a (U_1, U_2) which solves Equation (3.5).

□

A.2. Proofs for Sections 3.6.1 and 3.6.2

Recall assumptions A1, A2, A3, G1, G2 and the definition of F_1, F_2 in Sections 3.6.1 and 3.6.2 and the definition of W, X, Z and A, B, C, D, E from Section 3.4.1.

A.2.1. Proof of Theorem 3.1

Keep in mind that by a *representation* of a random vector Y we mean a matrix Q together with a random vector $F = (f_1, \ldots, f_r)$ with independent components, such that $Y = QF$.

To prove Theorem 3.1 we need the following seminal result which is contained in [Kagan et al., 1973, Theorem 10.3.1]. It allows to exploit non-Gaussianity of noise terms to achieve a certain kind of identifiability. The theorem will be at the core of the proof of Theorem 3.1.

Theorem A.1. *Let $Y = QF$ and $Y = RG$ be two representations of a p-dimensional random vector, where Q and R are constant matrices of order $p \times r$ and $p \times s$ respectively, and $F = (f_1, \ldots, f_r)$ and $G = (g_1, \ldots, g_s)$ are random vectors with independent components. Then the following assertion holds. If the i-th column of Q is not proportional to any column of R, then F_i is normal.*

We proceed with the proof of Theorem 3.1.

Proof of Theorem 3.1. **Ansatz:**

We prove that given P_X, the structural matrix B underlying X is determined uniquely.

Choosing (U_1, U_2):

Based on assumption G1 and Lemmas 3.4 and 3.2, there always exists (U_1, U_2) such that

$$\operatorname{cov}(R_t(U_1, U_2), X_{t-2-j}) = 0. \tag{A.6}$$

Pick one such (U_1, U_2).

Deriving a representation for $\begin{pmatrix} R_t(U_1, U_2) \\ R_{t-1}(U_1, U_2) \end{pmatrix}$:

Based on Lemma 3.3, we know that

$$B^2 + CD - U_1 B - U_2 = 0 \quad \wedge \quad BC + CE - U_1 C = 0,$$

and thus, based on Equation (3.4),

$$R_t(U_1, U_2) = N_t^X + CN_{t-1}^Z + (B - U_1)N_{t-1}^X.$$

Observe that

$$
\begin{aligned}
&\begin{pmatrix} R_t(U_1,U_2) \\ R_{t-1}(U_1,U_2) \end{pmatrix} \\
&= \begin{pmatrix} N_t^X + CN_{t-1}^Z + & (B-U_1)N_{t-1}^X & \\ & N_{t-1}^X & +CN_{t-2}^Z + (B-U_1)N_{t-2}^X \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{I} & C & (B-U_1) & 0 & 0 \\ 0 & 0 & \mathbf{I} & C & (B-U_1) \end{pmatrix} \begin{pmatrix} N_t^X \\ N_{t-1}^Z \\ N_{t-1}^X \\ N_{t-2}^X \\ N_{t-2}^Z \end{pmatrix} \\
&=: Q\tilde{N}_t.
\end{aligned}
$$

This is one representation of $\begin{pmatrix} R_t(U_1,U_2) \\ R_{t-1}(U_1,U_2) \end{pmatrix}$.

Based on Theorem A.1 and the structure of Q, $B-U_1$ is identifiable from $\begin{pmatrix} R_t(U_1,U_2) \\ R_{t-1}(U_1,U_2) \end{pmatrix}$. This can be seen as follows.

Identifying $B-U_1$ from $\begin{pmatrix} R_t(U_1,U_2) \\ R_{t-1}(U_1,U_2) \end{pmatrix}$:

Knowing P_X, we also know $P_{(R_t(U_1,U_2),R_{t-1}(U_1,U_2))}$ which in particular determines the class of all possible representations of $\begin{pmatrix} R_t(U_1,U_2) \\ R_{t-1}(U_1,U_2) \end{pmatrix}$. Pick one representation

$$
\begin{pmatrix} R_t(U_1,U_2) \\ R_{t-1}(U_1,U_2) \end{pmatrix} = Q'\tilde{N}_t'
$$

out of this class. W.l.o.g. let Q' be such that all its columns are pairwise linearly independent.

Theorem A.1 implies that each column of Q' is a scaled version of some column of Q and vice versa.

Now define the $K_X \times K_X$ matrix $V := (v_1, \ldots, v_{K_X})$ as follows.

For each $j = 1, \ldots, K_X$:

If Q' has a column with a non-zero entry at position $K_X + j$ and a non-zero entry in the upper half, let this column be denoted by q_j and define

$$v_j := \left[\frac{1}{[q_j]_{K_X+j}} q_j \right]_{1:K_X},$$

where $[q]_{k_1,\ldots,k_l}$ denotes the l-dimensional vector consisting of k_1st to k_lth entry of a vector q, and $k : l$ is shorthand for $k, k+1, \ldots, l$. Otherwise, if Q has no such column, then set

$$v_j := 0.$$

We have $V = B - U_1$. This can be seen as follows:

Let w_j denote the jth column of $B - U_1$.

For each $j = 1, \ldots, K_X$:

Either we have $w_j \neq 0$. Then the corresponding column in Q, i.e. $\begin{pmatrix} w_j \\ e_j \end{pmatrix}$, where e_j denotes the jth unit vector, is the only column with a non-zero entry at position $K_X + j$ and a non-zero entry in the upper half. Thus Q' contains a scaled version of $\begin{pmatrix} w_j \\ e_j \end{pmatrix}$ and no other column with a non-zero entry at position $K_X + j$ and a non-zero entry in the upper half. We denoted this column by q_j and defined $v_j = \left[\frac{1}{[q_j]_{K_X+j}} q_j \right]_{1:K_X}$. Since $\left[\frac{1}{[q_j]_{K_X+j}} q_j \right]_{K_X+j} = 1 = \left[\begin{pmatrix} w_j \\ e_j \end{pmatrix} \right]_{K_X+j}$, we know that $\frac{1}{[q_j]_{K_X+j}} q_j = \begin{pmatrix} w_j \\ e_j \end{pmatrix}$ and hence $v_j = w_j$.

Or we have $w_j = 0$. Then Q and hence also Q' contains no column with a non-zero entry at position $K_X + j$ and a non-zero entry in the upper half. Then by definition we have $v_j = 0$ and thus again $v_j = w_j$.

Hence $V = B - U_1$.

Putting all together:

We defined U_1 solely based on P_X and an arbitrary choice and then, for the fixed U_1, uniquely determined $B - U_1$, again only based on P_X. Hence $B = U_1 + (B - U_1)$ is uniquely determined by P_X.

□

A.2.2. Proof of Theorem 3.2

Here we prove Theorem 3.2.

Proof of Theorem 3.2. Keep in mind the proof of Theorem 3.1. There we showed that the matrix

$$Q = \begin{pmatrix} \mathbf{I} & C & (B - U_1) & 0 & 0 \\ 0 & 0 & \mathbf{I} & C & (B - U_1) \end{pmatrix}$$

is identifiable from P_X up to scaling and permutation of its columns, for some U_1. This implies that we can identify the matrix

$$Q_1 = \begin{pmatrix} \mathbf{I} & C \end{pmatrix}$$

up to scaling and permutation of its columns, simply by picking those columns of any scaled and permuted version of Q_1, that only have non-zero entries in the upper half.

But this in turn implies that we can identify the set of columns of C with at least two non-zero entries up to scaling of those columns. Just pick from any scaled and permuted version of Q_1 those columns with at least two non-zero entries.

□

A.2.3. Proof of Theorem 3.3

Following standard terminology [J. E. Dennis et al., 1976], for any $n \times n$-matrices $F_1, \ldots, F_m$ and Y we call

$$M(Y) := F_0 Y^m + F_1 Y^{m-1} + \ldots + F_m$$

a *matrix polynomial of degree* m. We say a matrix Y_0 is a *right solvent* or simply *solvent* of $M(Y)$, if $M(Y_0) = 0$. We say $\lambda \in \mathbb{C}$ is a *latent root* of $M(Y)$, if, slightly overloading notation, $M(\lambda) := M(\lambda \mathbf{I})$ is not invertible.

To prove Theorem 3.3 we need the following result which is a version of [J. E. Dennis et al., 1976, Corollary 4.1].

Theorem A.2. *Let* $M(Y) := F_0 Y^m + F_1 Y^{m-1} + \ldots + F_m$ *be any matrix polynomial, where* $F_1, \ldots, F_m$ *are* $n \times n$ *square matrices. If* $M(\lambda)$ *has* mn *distinct latent roots, then it has at most* $\binom{mn}{n}$ *different right solvents.*

(Note that this assertion is also stated in the conclusion section of [Pereira, 2003] but without proof it seems.)

Proof. In this proof we assume the paper [J. E. Dennis et al., 1976] as context. That is, *all definitions and equations we refer to in this proof are meant w.r.t. that paper.*

Let S be a solvent of $M(Y)$. By the corollary containing Equation (1.4), we have $M(\lambda) = Q(\lambda)(\mathbf{I}\,\lambda - S)$, with $Q(\lambda)$ a matrix polynomial of degree $m-1$. By assumption, we know that $\det(M(\lambda)) = \det(Q(\lambda))\det(\mathbf{I}\,\lambda - S)$ has mn distinct roots. Since $\det(Q(\lambda))$ has at most $(m-1)n$ different roots, we know that $\det(\mathbf{I}\,\lambda - S)$ has to have n different roots. Hence S has n distinct eigenvalues and is uniquely determined by its n eigenpairs, i.e. pairs $(a, \mathbb{C}\,v)$ such that $Sv = av$.

Keep in mind that a latent pair of $M(\lambda)$ is a scalar a together with a ray $\mathbb{C}\,v$ for some vector $v \neq 0$ such that $M(a)v = 0$. Let L denote the set of latent pairs of $M(\lambda)$. Based on Equation (1.4), each eigenpair of a solvent S is a latent pair of $M(\lambda)$. Hence for each solvent S, the tuple of n eigenpairs that uniquely determines this solvent has to be a

subset of size n of L. Therefore, the number of solvents is less or equal than $\binom{|L|}{n}$, in case L is finite.

Consider the $mn \times mn$ matrix C^B defined by Equation (3.2). Theorem 3.1, applied to C^B (see remark above Equation (3.2)), states that

$$\det(C^B - \lambda I) = (-1)^{mn} \det(M(\lambda)).$$

Hence $\det(C^B - \lambda I)$ has exactly mn distinct roots.

Now assume that $|L| > mn$, i.e., $M(\lambda)$ has more than mn latent pairs. Then there have to be two latent pairs $(a, \mathbb{C}\, v)$ and $(a, \mathbb{C}\, v')$ with $\mathbb{C}\, v \neq \mathbb{C}\, v'$. Based on Theorem 3.2, part (i), this implies that C^B, as defined by Equation (3.2), has two linearly independent vectors as eigenvectors to the same eigenvalue a. Thus the eigenvalue a has geometric and hence also algebraic multiplicity at least 2. This implies that $\det(C^B - \lambda I)$ has exactly mn distinct roots and at least one of the roots, namely a, has algebraic multiplicity at least 2. This is a contradiction to the fact that $\det(C^B - \lambda I)$ has degree mn.

□

Proof of Theorem 3.3. Keep in mind that assumption A3 reads $D = 0$.

Let S_1 denote the set of $U = (U_1, U_2)$ such that

$$\operatorname{cov}(R_t(U_1, U_2), X_{t-2-j}) = 0. \tag{A.7}$$

Let S_2 denote the set of $U = (U_1, U_2)$ such that $\det(T_U(\alpha))$ has $2K_X$ distinct roots.

Based on the assumption G2, there exists $U = (U_1, U_2)$ such that the equation

$$(U_1, U_2) \begin{pmatrix} B & C \\ \mathbf{I} & 0 \end{pmatrix} = \left(B^2, BC + CE\right) \tag{A.8}$$

is satisfied and $\det(T_U(\alpha))$ has $2K_X$ distinct roots. This U is in S_2 and based on Lemma 3.2 it is also in S_1. Hence $S := S_1 \cap S_2$ is non-empty.

Note that S is defined only based on P_X.

Pick one $U = (U_1, U_2)$ out of S.

Let

$$L := \{\tilde{B} : T_U(\tilde{B}) = 0\}.$$

Based on Theorem A.2, L has at most $\binom{2K_X}{K_X}$ elements. And since $U \in S_1$, assumption G1 together with Lemma 3.3 implies that $B \in L$, for the true B.

Hence B is determined by P_X up to $\binom{2K_X}{K_X}$ possibilities.

□

A.3. Discussion of the genericity assumptions: an elaboration of Section 3.6.3

This section is an elaborated version, including proofs, of Section 3.6.3.

We want to argue why the assumptions G1 and G2 stated in Sections 3.6.1 and 3.6.2 are generic. Keep in mind the definitions of W, X, Z, N, N^X, N^Z and A, B, C, D, E, Σ from Section 3.4.1 as well as M_1 from Section 3.5. The idea is to define a natural parametrization of (A, Σ) and to show that the restrictions that assumptions G1 and G2, respectively, impose on (A, Σ) just exclude a Lebesgue null set in the natural parameter space.

Have in mind that Theorems 3.1 and 3.3 state (almost) identifiability of B from P_X induced by any W in F_1 and F_2, respectively. In particular, such W can have *arbitrary numbers of components* K, as long as $K_X \leq K \leq 2K_X$. However, for the sake of simplicity, we show the genericity of assumptions G1 and G2 only under the assumption of an arbitrary but *fixed* K. Therefore, in this section, let K such that $K_X \leq K \leq 2K_X$ be arbitrary but fixed. As usual, let $K_Z = K - K_X$.

Let λ_k denote the k-dimensional Lebesgue measure on $\mathbb{R}^k$. Let vec denote the column stacking operator and vec^{-1} its inverse. The dimension of the domain of vec can always be understood from the context. For a vector q, let $[q]_{k_1,\ldots,k_l}$ denote the l-dimensional

vector consisting of k_1st to k_lth entry of q. Moreover, let $k : l$ be shorthand for $k, k+1, \ldots, l$.

A.3.1. Genericity assumption in Theorems 3.1 and 3.2

Let Θ_1 denote the set of all possible parameters (A', Σ') for a K-variate VAR processes W' that additionally satisfy assumption A2, i.e., correspond to structural W'. Let S_1 denote the subset of those $(A', \Sigma') \in \Theta_1$ for which also assumption G1 is satisfied.

(The relation between S_1 as defined above and F_1 as defined in Section 3.6.1 is the following: for any process W' with parameters (A', Σ'), $W' \in F_1$ iff W' satisfies assumption $A1$ (i.e., its noise components are non-Gaussian) and additionally $(A', \Sigma') \in S_1$.)

To parametrize Θ_1 in a practical way, let $g = (g_1, g_2) : \mathbb{R}^{K^2+K} \to \mathbb{R}^{K^2} \times \mathbb{R}^{K^2}$ be defined by

$$g_1(v) := \mathrm{vec}^{-1}([v]_{1:K^2}),$$
$$g_2(v) := \mathrm{diag}([v]_{K^2+1:K^2+K}),$$

for all $v \in \mathbb{R}^{K^2+K}$. Hence g_1 is the natural parametrization of A and g_2 for Σ.

We repeat the proposition already stated in Section 3.6.3:

Proposition 1. *We have* $\lambda_{K^2+K}\left(g^{-1}(\Theta_1 \setminus S_1)\right) = 0$.

Let $\Phi_1 := g^{-1}(\Theta_1)$. Since $g|_{\Phi_1} : \Phi_1 \to \Theta_1$ is a linear bijective function, the above statement can be interpreted as $\Theta_1 \setminus S_1$ being very small and thus G1 being a requirement that is met in the generic case.

A.3.1.1. Proof of Proposition 3.1

The proof idea for Proposition 3.1 is that $g^{-1}(\Theta_1 \setminus S_1)$ is essentially contained in the union of the root sets of finitely many multivariate polynomials and hence is a Lebesgue null set. Before we give a rigorous proof, we first need introduce some definitions and establish two lemmas.

Lemma A.1. *For any n and any non-zero multivariate polynomial $q(x_1, \ldots, x_n)$, the set*

$$L := \{(x_1, \ldots, x_n) \in \overset{n}{\mathbb{R}} : q(x_1, \ldots, x_n) = 0\}$$

is a null set w.r.t. the n-dimensional Lebesgue measure on $\mathbb{R}^n$.

Proof. We prove the statement via induction over n.

Basis:

Let $n = 1$. Let $q(x_1)$ be any non-zero polynomial. By the fundamental theorem of algebra it follows immediately that it has at most $\deg(q)$ real roots. Hence L is a Lebesgue null set.

Inductive step:

Now assume the statement holds for all multivariate polynomials in less than n variables. Let $q(x_1, \ldots, x_n)$ be any n-variate non-zero polynomial. We can consider q as a univariate polynomial in x_1, denoted by $r(x_1; x_{2:n})$, with coefficients $r_i(x_{2:n})$ that are multivariate polynomials in $x_{2:n}$, i.e.

$$q(x_1, \ldots, x_n) = r(x_1; x_{2:n}) = r_0(x_{2:n}) + r_1(x_{2:n})x_1 + \ldots + r_l(x_{2:n})x_1^l,$$

for some l.

There has to be some j such that $r_j(x_{2:n})$ is not the zero polynomial, since otherwise $q(x_1, \ldots, x_n)$ would be the zero polynomial. Let

$$L' := \{(x_2, \ldots, x_n) \in \overset{n-1}{\mathbb{R}} : r_j(x_2, \ldots, x_n) = 0\}.$$

By induction, we know that $\lambda_{n-1}(L') = 0$. Hence $r(x_1; x_{2:n})$ is a non-zero polynomial for all $x_{2:n} \in \mathbb{R}^{n-1} \setminus L'$. In particular, due to the fundamental theorem of algebra, for all $x_{2:n} \in \mathbb{R}^{n-1} \setminus L'$, the set $L_{x_{2:n}} := \{x_1 \in \mathbb{R} : r(x_1; x_{2:n}) = 0\}$ is finite (has at most $n - 1$ elements).

Note that, since q is continuous, $L = q^{-1}(\{0\})$ is closed and thus measurable. Let 1 denote the indicator function. In particular, 1_L is measurable. Furthermore, note that

$1_L(x_{1:n}) = 1_{L_{x_{2:n}}}(x_1)$ for all $x_{1:n}$. Therefore and due to Fubini's theorem (for completed product spaces) we have

$$\begin{aligned}
\lambda_n(L) &= \int_{\mathbb{R}^n} 1_L(x_1, \ldots, x_n)\ \mathrm{d}x_{1:n} \\
&= \int_{\mathbb{R}^{n-1}} \int_{\mathbb{R}} 1_{L_{x_{2:n}}}(x_1)\ \mathrm{d}x_1\ \mathrm{d}x_{2:n} \\
&= \int_{\mathbb{R}^{n-1} \setminus L'} \int_{\mathbb{R}} 1_{L_{x_{2:n}}}(x_1)\ \mathrm{d}x_1\ \mathrm{d}x_{2:n} \\
&= \int_{\mathbb{R}^{n-1} \setminus L'} \lambda_1(L_{x_{2:n}})\ \mathrm{d}x_{2:n} \\
&= \int_{\mathbb{R}^{n-1} \setminus L'} 0\ \mathrm{d}x_{2:n} \\
&= 0.
\end{aligned}$$

□

Let $\Psi_1 := g^{-1}(S_1)$.

For a $I \times J$ matrix

$$M = \begin{pmatrix} m_{11} & \ldots & m_{1J} \\ & \vdots & \\ m_{I1} & \ldots & m_{IJ} \end{pmatrix},$$

let $[M]_{ij} := m_{ij}$ and $[M]_{i_1:i_2, j_1:j_2} := (m_{ij})_{i_1 \le i \le i_2, j_1 \le j \le j_2}$.

Keep in mind the following equations for the autocovariance matrices $\Gamma_i := \mathbb{E}[\tilde{W}_t \tilde{W}_{t-i}^\top]$ of any VAR process $\tilde{W}$ with parameters $(\tilde{A}, \tilde{\Sigma})$ [Lütkepohl, 2006]:

$$\mathrm{vec}(\Gamma_0) = (\mathbf{I} - \tilde{A} \otimes \tilde{A})^{-1} \mathrm{vec}(\tilde{\Sigma}), \tag{A.9}$$

$$\Gamma_i = \tilde{A}^i \Gamma_{i-1}. \tag{A.10}$$

In this subsection, given any $\phi \in \Phi_1$, let W^ϕ be some K-variate VAR process with parameters $g(\phi)$, and let X^ϕ denote the first K_X and Z^ϕ denote the remaining $K - K_X$ components of W^ϕ.

And also for this subsection, for any $\phi \in \Phi_1$ and $i \ge 0$, let $\Gamma_i(\phi) := \mathbb{E}[W_t^\phi (W^\phi)_{t-i}^\top]$.

Recall the definition of M_1 from Section 3.5. Here we explicitly consider M_1 as a function on Φ_1. That is, for any $\phi \in \Phi_1$ let

$$M_1(\phi) := \mathbb{E}\left[W_t^\phi((X_t^\phi)^\top, (X_{t-1}^\phi)^\top)\right].$$

Later we want to shot that the set of $\phi \in \Phi_1$ for which $M_1(\phi)$ does not have full rank is a Lebesgue null set. It suffices to show that $M_1(\phi)$ has a fixed square submatrix $M_2(\phi)$ such that the set of $\phi \in \Phi_1$ for which $M_2(\phi)$ is not invertible is a Lebesgue null set, since the former set is contained in the latter. For this purpose let us define

$$M_2(\phi) := \begin{cases} \mathbb{E}\left[W_t^\phi\left((X_t^\phi)^\top, [X_{t-1}^\phi]_{K_X-K_Z:K_X}\right)\right], & \text{if } K > K_X \\ \mathbb{E}\left[W_t^\phi(W_t^\phi)^\top\right] (= \Gamma_0(\phi)), & \text{if } K = K_X \end{cases}.$$

That is, M_2 is a $K \times K$ square matrix with a subset of the columns of M_1 as columns (keep in mind that $[X_{t-1}^\phi]_{K_X-K_Z:K_X}$ are the $(K_X - K_Z)$-th to K_X-th components of X_{t-1}^ϕ).

Let

$$f(\phi) := \det(M_2(\phi)). \tag{A.11}$$

Lemma A.2. *There is some $\phi \in \Phi_1$ such that $f(\phi) \neq 0$.*

Proof. We only treat the cases $K = K_X$ and $K = K_X+1$. The cases $K_X+1 < K \leq 2K_X$ can be treated similarly.

The case $K = K_X$:

Let $\tilde{A} := \frac{1}{2}\,\mathbf{I}$ and $\tilde{\Sigma} := \mathbf{I}$ and let $\phi := g^{-1}(\tilde{A}, \tilde{\Sigma})$. Based on Equation (A.9) this immediately implies

$$M_2(\phi) = \Gamma_0(\phi) = \frac{4}{3}\,\mathbf{I}, \tag{A.12}$$

and hence $f(\phi) = \det(M_2(\phi)) \neq 0$.

The case $K = K_X + 1$:

Let $\tilde{\Sigma} := \mathbf{I}$ and

$$\tilde{A} := \left(\begin{array}{ccc|cc} \frac{1}{2} & & & & \\ & \ddots & & & \\ & & \frac{1}{2} & & \\ \hline & & & \frac{1}{2} & \frac{1}{2} \\ & & & & \frac{1}{2} \end{array}\right) =: \left(\begin{array}{c|c} \tilde{A}_1 & 0 \\ \hline 0 & \tilde{A}_2 \end{array}\right),$$

and let $\phi := g^{-1}(\tilde{A}, \tilde{\Sigma})$ denote the corresponding parameter vector. Now we want to calculate $\Gamma_0(\phi), \Gamma_1(\phi)$. For this purpose, observe that we can split W^ϕ into the two independent VAR processes

$$\begin{aligned} Y^1 &:= ([X^\phi]_1, \ldots, [X^\phi]_{K_X - 1})^\top, \\ Y^2 &:= ([X^\phi]_{K_X}, Z)^\top. \end{aligned}$$

Equation (A.9) applied to Y^1 implies

$$\mathrm{vec}(\mathbb{E}[Y_t^1 (Y_t^1)^\top]) = (\mathbf{I} - \tilde{A}_1 \otimes \tilde{A}_1)^{-1} \mathrm{vec}(\mathbf{I}) = \frac{4}{3} \mathrm{vec}(\mathbf{I}),$$

that is

$$\mathbb{E}[Y_t^1 (Y_t^1)^\top] = \frac{4}{3} \mathbf{I}.$$

On the other hand, Equation (A.9) applied to Y^2 yields

$$\mathrm{vec}(\mathbb{E}[Y_t^2 (Y_t^2)^\top]) = (\mathbf{I} - \tilde{A}_2 \otimes \tilde{A}_2)^{-1} \mathrm{vec}(\mathbf{I}) = \frac{4}{27} \begin{pmatrix} 9 & 3 & 3 & 5 \\ & 9 & & 3 \\ & & 9 & 3 \\ & & & 9 \end{pmatrix} \mathrm{vec}(\mathbf{I}),$$

that is

$$\mathbb{E}[Y_t^2 (Y_t^2)^\top] = \frac{4}{27} \begin{pmatrix} 14 & 3 \\ 3 & 9 \end{pmatrix}.$$

Thus

$$\Gamma_0(\phi) = \mathbb{E}[W_t^\phi (W^\phi)_t^\top] = \begin{pmatrix} \mathbb{E}[Y_t^1 (Y_t^1)^\top] & 0 \\ 0 & \mathbb{E}[Y_t^2 (Y_t^2)^\top] \end{pmatrix}$$
$$= \begin{pmatrix} \frac{4}{3} & & & & & \\ & \ddots & & & & \\ & & \frac{4}{3} & & & \\ & & & \frac{56}{27} & \frac{4}{9} & \\ & & & \frac{4}{9} & \frac{4}{3} & \end{pmatrix},$$

and

$$\Gamma_1(\phi) = \tilde{A}\Gamma_0(\phi) = \begin{pmatrix} \frac{2}{3} & & & & & \\ & \ddots & & & & \\ & & \frac{2}{3} & & & \\ & & & \frac{34}{27} & \frac{8}{9} & \\ & & & \frac{2}{9} & \frac{2}{3} & \end{pmatrix}.$$

Hence

$$M_2(\phi) = \begin{pmatrix} \frac{4}{3} & & & & & \\ & \ddots & & & & \\ & & \frac{4}{3} & & & \\ & & & \frac{56}{27} & \frac{34}{27} & \\ & & & \frac{4}{9} & \frac{2}{9} & \end{pmatrix}.$$

Hence ϕ is such that $f(\phi) = \det(M_2(\phi)) \neq 0$.

□

Proof of Proposition 3.1. Recall that $\Phi_1 = g^{-1}(\Theta_1)$, $\Psi_1 = g^{-1}(S_1)$, $f(\phi) = \det(M_2(\phi))$, and how S_1 is related to f.

First, show that f is a rational function:

Keep in mind that each entry of $g_1(\phi)$ is a linear function in ϕ.

For any $\phi \in \Phi_1$, let $G(\phi) := \mathbf{I} - g_1(\phi) \otimes g_1(\phi)$. Note that each entry of $G(\phi)$ is a multivariate polynomial in ϕ. We have for $i = 0, 1$, and for all $\phi \in \Phi_1$, using Equation (A.9) and Cramer's rule,

$$\begin{aligned}\Gamma_i(\phi) &= g_1(\phi)^i \mathrm{vec}^{-1}(G(\phi)^{-1}\mathrm{vec}(g_2(\phi))) \\ &= g_1(\phi)^i \mathrm{vec}^{-1}(\det(G(\phi))^{-1}\mathrm{adj}(G(\phi))\mathrm{vec}(g_2(\phi))) \\ &= \det(G(\phi))^{-1} g_1(\phi)^i \mathrm{vec}^{-1}(\mathrm{adj}(G(\phi))\mathrm{vec}(g_2(\phi))).\end{aligned}$$

(Note that the definition of Φ_1 implies that $\|g_1\| < 0$ and thus $\det(G) \neq 0$ on Φ_1.)

Keep in mind that for any matrix Q, the determinant $\det(Q)$ as well as all entries of the adjugate $\mathrm{adj}(Q)$, are multivariate polynomials in the entries of Q. In particular each entry of $g_1(\phi)^i \mathrm{vec}^{-1}(\mathrm{adj}(G(\phi))\mathrm{vec}(g_2(\phi)))$ is a multivariate polynomial in ϕ.

Now observe that on Φ_1 we have

$$\begin{aligned}f &= \det(M_2) \\ &= \det\left(\left([\Gamma_0]_{1:K,1:K_X}, [\Gamma_1]_{1:K,K_X-K_Z:K_X}\right)\right) \\ &= \det\left(\left(\left[\det(G)^{-1}\mathrm{vec}^{-1}(\mathrm{adj}(G)\mathrm{vec}(g_2))\right]_{1:K,1:K_X}, \right.\right. \\ &\quad \left.\left.\left[\det(G)^{-1} g_1 \mathrm{vec}^{-1}(\mathrm{adj}(G)\mathrm{vec}(g_2))\right]_{1:K,K_X-K_Z:K_X}\right)\right) \\ &= \det(G)^{-K} \det\left(\left(\left[\mathrm{vec}^{-1}(\mathrm{adj}(G)\mathrm{vec}(g_2))\right]_{1:K,1:K_X}, \right.\right. \\ &\quad \left.\left.\left[g_1 \mathrm{vec}^{-1}(\mathrm{adj}(G)\mathrm{vec}(g_2))\right]_{1:K,K_X-K_Z:K_X}\right)\right)\end{aligned}$$

For all $\phi \in \mathbb{R}^{K^2+K}$, let

$$\begin{aligned}&r(\phi) := \det(G(\phi))^K, \\ &q(\phi) := \det\left(\left(\left[\mathrm{vec}^{-1}(\mathrm{adj}(G)\mathrm{vec}(g_2))\right]_{1:K,1:K_X}, \left[g_1 \mathrm{vec}^{-1}(\mathrm{adj}(G)\mathrm{vec}(g_2))\right]_{1:K,K_X-K_Z:K_X}\right)\right).\end{aligned}$$

Based on the above argument, $q(\phi), r(\phi)$ are multivariate polynomials (mappings from $\mathbb{R}^{K^2+K}$ to $\mathbb{R}$). Hence in particular, $f = \frac{q}{r}$ is a rational function on Φ_1.

Second, show that $\lambda_{K^2+K} \circ g^{-1}(\Theta_1 \setminus S_1) = 0$:

In what follows, we only discuss the case $K > K_X$. The case $K = K_X$ works similarly and is even simpler.

Let $\tilde{C}(\phi)$ denote the upper right submatrix of $g_1(\phi)$ of dimension $K_X \times K_Z$. Keep in mind that $\Psi_1 = g^{-1}(S_1)$ is the set of those $\phi \in \Phi_1 = g^{-1}(\Theta_1)$, for which $\tilde{C}(\phi)$ and $M_1(\phi)$ have full rank.

Let H denote the set of those $\phi \in \Phi_1$, for which $\det\left(\left[\tilde{C}(\phi)\right]_{1:K_Z,1:K_Z}\right) = 0$. Since $\det\left(\left[\tilde{C}(\phi)\right]_{1:K_Z,1:K_Z}\right)$ is a non-zero multivariate polynomial in ϕ, based on Lemma A.1 we have $\lambda_{K^2+K}(H) = 0$.

Let H' denote the set of those $\phi \in \Phi_1$, for which $q(\phi) = 0$. Based on Lemma A.2 we know that there is some ϕ such that $q(\phi) \neq 0$. Hence based on Lemma A.1 we have $\lambda_{K^2+K}(H') = 0$.

If any ϕ is in Φ_1 but neither in H nor in H', then $\det\left(\left[\tilde{C}(\phi)\right]_{1:K_Z,1:K_Z}\right) \neq 0$ and $q(\phi) \neq 0$, and thus $\tilde{C}(\phi)$ and $M_1(\phi)$ have full rank. That is, $H^C \cap (H')^C \cap \Phi_1 \subset \Psi_1$. Therefore

$$\begin{aligned}
\lambda_{K^2+K}\left(g^{-1}(\Theta_1 \setminus S_1)\right) &= \lambda_{K^2+K}\left(g^{-1}(\Theta_1) \setminus g^{-1}(S_1)\right) \\
&= \lambda_{K^2+K}\left(\Phi_1 \setminus \Psi_1\right) \\
&\leq \lambda_{K^2+K}\left(\Phi_1 \setminus (H^C \cap (H')^C \cap \Phi_1)\right) \\
&\leq \lambda_{K^2+K}\left(\Phi_1 \setminus (H^C \cap (H')^C)\right) \\
&= \lambda_{K^2+K}\left(\Phi_1 \setminus (H \cup H')^C\right) \\
&= \lambda_{K^2+K}\left(\Phi_1 \cap (H \cup H')\right) = 0.
\end{aligned}$$

□

A.3.2. Genericity assumptions in Theorem 3.3

Let Θ_2 denote the set of all possible parameters (A', Σ') for the K-variate VAR processes W that additionally satisfy assumption A3, i.e., are such that the submatrix D of A is zero. Let S_2 denote the subset of those $(A', \Sigma') \in \Theta_2$ for which also assumption G1 and G2 is satisfied.

To parametrize Θ_2 in a practical way, let $h = (h_1, h_2) : \mathbb{R}^{2K^2 - K_X K_Z} \to \mathbb{R}^{K^2} \times \mathbb{R}^{K^2}$ be defined by

$$h_1(v) := \begin{pmatrix} \mathrm{vec}^{-1}([v]_{1:K_X^2}) & \mathrm{vec}^{-1}([v]_\alpha) \\ 0 & \mathrm{vec}^{-1}([v]_\beta) \end{pmatrix},$$
$$h_2(v) := \mathrm{vec}^{-1}([v]_{K^2 - K_X K_Z + 1:2K^2 - K_X K_Z}),$$

for all $v \in \mathbb{R}^{K^2+K}$, where

$$\alpha := K_X^2 + 1 : K_X^2 + K_X K_Z,$$
$$\beta := K_X^2 + K_X K_Z + 1 : K^2 - K_X K_Z.$$

Hence h_1 is the natural parametrization of A and h_2 for Σ.

We repeat the proposition already stated in Section 3.6.3:

Proposition 2. *We have* $\lambda_{2K^2 - K_X K_Z}\left(h^{-1}(\Theta_2 \setminus S_2)\right) = 0$.

Let $\Phi_2 := h^{-1}(\Theta_2)$. Since $h|_{\Phi_2} : \Phi_2 \to \Theta_2$ is a linear bijective function, the above statement can be interpreted as $\Theta_2 \setminus S_2$ being very small and thus the combination of G1 and G2 being a requirement that is met in the generic case.

A.3.2.1. Proof of Proposition 3.2

The proof idea for Proposition 3.2 - similar as for Proposition 3.1 - is that $h^{-1}(\Theta_2 \setminus S_2)$ is essentially contained in the union of the root sets of finitely many multivariate polynomials and hence is a Lebesgue null set. To give a rigorous proof of Proposition 3.2, we first need to introduce some definitions which are very similar to those in Section A.3.1, and establish a lemma.

Recall that $T_{(U_1,U_2)}(Q) = Q^2 - U_1 Q - U_2$ (see Section 3.6.2).

Within this section, given any $\phi \in \Phi_2$, let W^ϕ be some K-variate VAR process with parameters $h(\phi)$, and let X^ϕ denote the first K_X and Z^ϕ denote the remaining $K - K_X$ components of W^ϕ. And also for this section, for any $\phi \in \Phi_2$ and $i \geq 0$, let $\Gamma_i(\phi) := \mathbb{E}[W_t^\phi (W^\phi)_{t-i}^\top]$.

Recall the definition of M_1 from Section 3.5. Here we explicitly consider M_1 as a function on Φ_2. That is, for any $\phi \in \Phi_2$ let

$$M_1(\phi) := \mathbb{E}\left[W_t^\phi((X_t^\phi)^\top, (X_{t-1}^\phi)^\top)\right].$$

Lemma A.3. *Let $q_0(x_1, \ldots, x_m), \ldots, q_n(x_1, \ldots, x_m)$ be multivariate polynomials (elements of in $\mathbb{R}[x_1, \ldots, x_m]$). Let*

$$q(\alpha; x_1, \ldots, x_m) := q_0(x_1, \ldots, x_m) + q_1(x_1, \ldots, x_m)\alpha + \ldots + q_n(x_1, \ldots, x_m)\alpha^n,$$

i.e. a univariate polynomial in α (an element of $\mathbb{R}[\alpha]$) parametrized by $(x_1, \ldots, x_m)$. If $q(\alpha; x_1, \ldots, x_m)$ has n distinct roots for one $(x_1, \ldots, x_m) \in \mathbb{R}^m$, then

$$\{(x_1, \ldots, x_m) \in \overset{m}{\mathbb{R}} : q(\cdot; x_1, \ldots, x_m) \text{ does not have } n \text{ distinct roots}\}.$$

is a null set w.r.t. the m-dimensional Lebesgue measure on $\mathbb{R}^m$.

Proof. Given two polynomials $r(\alpha), s(\alpha)$, let $S(r, s)$ denote their Sylvester matrix [Dickenstein and Emiris, 2010, Weisstein, 2015]. Keep in mind that all entries of the Sylvester matrix $S(r, s)$ are either 0 or coincide with a coefficient of r or s. Hence in particular, all entries of $S(r, s)$ are polynomials in the coefficients of r and s.

Given a non-zero polynomial $p(\alpha) = p_0 + p_1\alpha + \ldots + p_{\deg(p)}\alpha^{\deg(p)}$, let $\Delta(p)$ denote its discriminant, i.e.

$$\Delta(p) := p_{\deg(p)}^{2\deg(p)-2} \prod_{i<j} (\alpha_i - \alpha_j)^2,$$

where $\alpha_1, \ldots, \alpha_{\deg(p)}$ are the $\deg(p)$ complex roots of p, with potential multiplicities.

Keep in mind the following equation [Dickenstein and Emiris, 2010, Weisstein, 2015] that relates discriminant and Sylvester matrix: for all polynomials $p(\alpha)$ we have

$$(-1)^{\frac{1}{2}\deg(p)(\deg(p)-1)} p_{\deg(p)} \Delta(p) = \det(S(p, p')), \tag{A.13}$$

where $p'(\alpha)$ is the derivative of $p(\alpha)$ w.r.t. α.

Let

$$s(x_1, \ldots, x_m) := \det(S(q(\cdot; x_1, \ldots, x_m), q'(\cdot; x_1, \ldots, x_m))),$$

which is a multivariate polynomial in $(x_1, \ldots, x_m)$ based on the fact that the coefficients of $q(\cdot; x_1, \ldots, x_m)$ are multivariate polynomial in $(x_1, \ldots, x_m)$ and the determinant of the Sylvester matrix is a multivariate polynomial in the coefficients of $q(\cdot; x_1, \ldots, x_m)$.

By assumption there is one $(x_1, \ldots, x_m) \in \mathbb{R}^m$ such that $q(\cdot; x_1, \ldots, x_m)$ has n distinct roots. Based on Equation (A.13) and the definition of $\Delta(q(\cdot; x_1, \ldots, x_m))$, this implies that for this $(x_1, \ldots, x_m)$, $s(x_1, \ldots, x_m) \neq 0$. Based on Lemma A.1, $s(x_1, \ldots, x_m) \neq 0$ for all $(x_1, \ldots, x_m) \in \mathbb{R}^m \setminus L$, for some Lebesgue null set L.

Using Equation (A.13) again, we know that $\Delta(q(\cdot; x_1, \ldots, x_m)) \neq 0$ for all $(x_1, \ldots, x_m) \in \mathbb{R}^m \setminus L$. Hence $q(\cdot; x_1, \ldots, x_m)$ has n distinct roots for all $(x_1, \ldots, x_m) \in \mathbb{R}^m \setminus L$.

□

Proof of Proposition 3.2. **Prerequisties:**

Keep in mind that $\Phi_2 = h^{-1}(\Theta_2)$ and $\Psi_2 = h^{-1}(S_2)$.

Let

$$\begin{pmatrix} \tilde{B}(\phi) & \tilde{C}(\phi) \\ 0 & \tilde{E}(\phi \end{pmatrix} := \tilde{A}(\phi) := h_1(\phi).$$

Let H denote the set of those $\phi \in \Phi_2$, for which $\tilde{C}(\phi)$ and $M_1(\phi)$ have full rank. Let H' denote the set of those $\phi \in \Phi_2$, for which $\tilde{A}(\phi)$ is such that there exists $U = (U_1, U_2)$ such that the equation

$$(U_1, U_2) \begin{pmatrix} \tilde{B}(\phi) & \tilde{C}(\phi) \\ \mathbf{I} & 0 \end{pmatrix} = \left(\tilde{B}(\phi)^2, \tilde{B}(\phi)\tilde{C}(\phi) + \tilde{C}(\phi)\tilde{E}(\phi)\right), \tag{A.14}$$

or equivalently

$$(U_1, U_2) \begin{pmatrix} \tilde{C}(\phi) & \tilde{B}(\phi) \\ 0 & \mathbf{I} \end{pmatrix} = \left(\tilde{B}(\phi)\tilde{C}(\phi) + \tilde{C}(\phi)\tilde{E}(\phi), \tilde{B}(\phi)^2\right) \tag{A.15}$$

is satisfied.

Keep in mind that $\Psi_2 = H \cap H'$.

Similar as in the proof of Proposition 1, it can be shown that

$$\lambda_{2K^2-K_XK_Z}(\Phi_2 \setminus H) = 0. \tag{A.16}$$

It remains to show the same for H'.

The case $K_Z = K_X$:

Let L_C denote the set of those $\phi \in \Phi_2$, for which $\tilde{C}(\phi)$ is not invertible. As usual (see the proof of Proposition 3.1), Lemma A.1 implies that L_C has Lebesgue measure zero.

For all $\phi \in \mathbb{R}^{2K^2-K_XK_Z}$, define $U(\phi) = (U_1(\phi), U_2(\phi))$ as follows:

On $\mathbb{R}^{2K^2-K_XK_Z} \setminus L_C$ let

$$\begin{aligned}
(U_1, U_2) &:= \left(\tilde{B}\tilde{C} + \tilde{C}\tilde{E}, \tilde{B}^2\right) \begin{pmatrix} \tilde{C}^{-1} & -\tilde{C}^{-1}\tilde{B} \\ 0 & \mathbf{I} \end{pmatrix} && \text{(A.17)} \\
&= \left(\tilde{B} + \tilde{C}\tilde{E}\tilde{C}^{-1}, -\tilde{B}^2 - \tilde{C}\tilde{E}\tilde{C}^{-1}\tilde{B} + \tilde{B}^2\right) && \text{(A.18)} \\
&= \left(\tilde{B} + \tilde{C}\tilde{E}\tilde{C}^{-1}, -\tilde{C}\tilde{E}\tilde{C}^{-1}\tilde{B}\right) && \text{(A.19)} \\
&= \left(\tilde{B} + \tilde{C}\tilde{E}\det(\tilde{C})^{-1}\mathrm{adj}(\tilde{C}), -\tilde{C}\tilde{E}\det(\tilde{C})^{-1}\mathrm{adj}(\tilde{C})\tilde{B}\right) && \text{(A.20)} \\
&= \det(\tilde{C})^{-1}\left(\det(\tilde{C})\tilde{B} + \tilde{C}\tilde{E}\mathrm{adj}(\tilde{C}), -\tilde{C}\tilde{E}\mathrm{adj}(\tilde{C})\tilde{B}\right), && \text{(A.21)}
\end{aligned}$$

where, as usual, adj denotes the adjugate of a matrix. Otherwise, on L_C, let $(U_1, U_2) := (0, 0)$ (or anything else since this case does not matter).

On $\mathbb{R}^{2K^2-K_XK_Z} \setminus L_C$ we have

$$\begin{aligned}
&\det(T_U(\alpha)) \\
&= \det(\alpha^2\,\mathbf{I} - U_1\alpha - U_2) \\
&= \det(\tilde{C})^{-K_X} \det\left(\det(\tilde{C})\alpha^2\,\mathbf{I} - \alpha\left(\det(\tilde{C})\tilde{B} + \tilde{C}\tilde{E}\mathrm{adj}(\tilde{C})\right) + \tilde{C}\tilde{E}\mathrm{adj}(\tilde{C})\tilde{B}\right).
\end{aligned}$$

Keep in mind that for any matrix Q, the determinant $\det(Q)$ as well as all entries of the adjugate $\mathrm{adj}(Q)$, are multivariate polynomials in the entries of Q. (And obviously the entries of $\tilde{A}(\phi), \tilde{B}(\phi), \tilde{C}(\phi), \tilde{E}(\phi)$ are multivariate polynomials in ϕ.)

Hence

$$\begin{aligned}\tilde{q}(\alpha,\phi) := \det\Big(&\det(\tilde{C}(\phi))\alpha^2\,\mathbf{I} - \alpha\Big(\det(\tilde{C}(\phi))\tilde{B}(\phi) + \tilde{C}(\phi)\tilde{E}(\phi)\mathrm{adj}(\tilde{C}(\phi))\Big)\\ &+\tilde{C}(\phi)\tilde{E}(\phi)\mathrm{adj}(\tilde{C}(\phi))\tilde{B}(\phi)\Big)\end{aligned} \tag{A.22}$$

is a multivariate polynomial in $(\alpha,\phi) \in \mathbb{R}\times\mathbb{R}^{2K^2-K_XK_Z}$. And in particular, considering ϕ as parameter vector,

$$q(\alpha;\phi) := \tilde{q}(\alpha,\phi)$$

is a univariate polynomial in α, whose coefficients are all multivariate polynomials in ϕ. Note that $q(\alpha;\phi)$ has degree $2K_X$ for all $\phi \in \mathbb{R}^{2K^2-K_XK_Z}\setminus L_C$, since it is up to a constant, which does not depend on α, equal to $\det(\alpha^2\,\mathbf{I} - U_1\alpha - U_2)$.

We want to apply Lemma A.3 to $q(\alpha;\phi)$. For this purpose we need to show that there is a $\phi \in \mathbb{R}^{2K^2-K_XK_Z}$, such that $q(\alpha;\phi)$ has $2K_X$ distinct roots.

Let ϕ be such that

$$\tilde{B}(\phi) = \mathrm{diag}(1,3,5,\ldots,2K_X-1), \tag{A.23}$$

$$\tilde{C}(\phi) = \mathbf{I}, \tag{A.24}$$

$$\tilde{E}(\phi) = \mathrm{diag}(2,4,6,\ldots,2K_X). \tag{A.25}$$

For this ϕ we have

$$\begin{aligned}q(\alpha;\phi) &= \det(\alpha^2\,\mathbf{I} - \alpha(\mathrm{diag}(1,3,5,\ldots,2K_X-1) + \mathrm{diag}(2,4,6,\ldots,2K_X))\\ &\quad + \mathrm{diag}(1,3,5,\ldots,2K_X-1)\mathrm{diag}(2,4,6,\ldots,2K_X))\\ &= (\alpha^2 - (1+2)\alpha + 1\cdot 2)(\alpha^2 - (3+4)\alpha + 3\cdot 4)\cdot\ldots\\ &\quad \cdot(\alpha^2 - (2K_X - 1 + 2K_X)\alpha + (2K_X-1)2K_X)\\ &= (\alpha-1)(\alpha-2)(\alpha-3)(\alpha-4)\cdot\ldots\cdot(\alpha-(2K_X-1))(\alpha-2K_X+2)\end{aligned}$$

hence $q(\alpha;\phi)$ has the distinct roots $1,2,\ldots,2K_X$.

Now Lemma A.3 implies that $q(\alpha;\phi)$ has $2K_X$ distinct roots for all $\phi \in \mathbb{R}^{2K^2-K_XK_Z} \setminus L$, for some L with $\lambda_{2K^2-K_XK_Z}(L) = 0$.

Keep in mind that

$$\det(T_{U(\phi)}(\alpha)) = \det(C)^{-1} q(\alpha;\phi)$$

for all $\phi \in \mathbb{R}^{2K^2-K_XK_Z} \setminus L_C$. Hence $\det(T_{U(\phi)}(\alpha))$ has $2K_X$ distinct roots for all $\phi \in \mathbb{R}^{K^2+K} \setminus (L \cup L_C)$. Moreover, for all $\phi \in \mathbb{R}^{2K^2-K_XK_Z} \setminus (L \cup L_C)$, $U(\phi)$ satisfies Equation (A.14) by its definition. This implies $(L \cup L_C)^C \subset H'$ and in particular $(H')^C \subset L \cup L_C$, where $(\cdot)^C$ denotes the complement of a set, as usual. Hence $\lambda_{2K^2-K_XK_Z}((H')^C) = 0$.

Using the fact that $\Psi_2^C = (H \cap H')^C = H^C \cup (H')^C$ and Equation (A.16) we can calculate

$$\begin{aligned}
\lambda_{2K^2-K_XK_Z}\left(h^{-1}(\Theta_2 \setminus S_2)\right) &= \lambda_{2K^2-K_XK_Z}(\Phi_2 \setminus \Psi_2) \\
&= \lambda_{2K^2-K_XK_Z}(\Phi_2 \cap \Psi_2^C) \\
&= \lambda_{2K^2-K_XK_Z}(\Phi_2 \cap (H^C \cup (H')^C)) \\
&= \lambda_{2K^2-K_XK_Z}((\Phi_2 \cap H^C) \cup (\Phi_2 \cap (H')^C)) \\
&\leq \lambda_{2K^2-K_XK_Z}(\Phi_2 \cap H^C) + \lambda_{2K^2-K_XK_Z}(\Phi_2 \cap (H')^C) \\
&= 0.
\end{aligned}$$

Second, the case $K_Z < K_X$:

This case works similarly as the case $K_Z = K_X$.

Let $\mathbf{I}_m$ denote the $m \times m$ identity matrix and $0_{m\times n}$ the $m \times n$ zero matrix. For the sake of a simple notation, here we suppress the dependence on ϕ. Let

$$d := \operatorname{diag}(2, 4, 6, \ldots, 2(K_X - K_Z))$$

and

$$\begin{aligned}
\hat{B} &:= \tilde{B}, \\
\hat{C} &:= \left(\begin{array}{c|c} \mathbf{I}_{K_X - K_Z} & \\ 0_{K_Z \times (K_X - K_Z)} & \tilde{C} \end{array} \right). \\
\hat{E} &:= \begin{pmatrix} d & 0_{(K_X - K_Z) \times K_Z} \\ 0_{K_Z \times (K_X - K_Z)} & \tilde{E} \end{pmatrix}.
\end{aligned}$$

Note that $\hat{B}, \hat{C}, \hat{E}$ all have dimension $K_X \times K_X$.

Now the argument is similar as for the case $K_Z = K_X$, except that we replace $\tilde{B}, \tilde{C}, \tilde{E}$ by $\hat{B}, \hat{C}, \hat{E}$.

Let us briefly comment on two points.

First, similar as for the case $K_Z = K_X$, whenever $\hat{C}$ is invertible, we define

$$(U_1, U_2) := \left(\hat{B}\hat{C} + \hat{C}\hat{E}, \hat{B}^2 \right) \begin{pmatrix} \hat{C}^{-1} & -\hat{C}^{-1}\hat{B} \\ 0 & \mathbf{I} \end{pmatrix}. \tag{A.26}$$

(The argument for $\tilde{C}$ to almost always have full rank and thus $\hat{C}$ almost always being

invertible carries over from the case $K_Z = K_X$.) This implies that (U_1, U_2) satisfies

$$
\begin{aligned}
&(U_1, U_2) \left(\begin{array}{c|c|c} \mathbf{I}_{K_X - K_Z} & \tilde{C} & \tilde{B} \\ 0_{K \times (K_X - K_Z)} & 0 & \mathbf{I} \end{array}\right) \\
&= (U_1, U_2) \begin{pmatrix} \hat{C} & \hat{B} \\ 0 & \mathbf{I} \end{pmatrix} \\
&= \left(\hat{B}\hat{C} + \hat{C}\hat{E}, \hat{B}^2\right) \\
&= \left(\tilde{B} \left(\begin{array}{c} \mathbf{I}_{K_X - K_Z} \\ 0_{K_Z \times (K_X - K_Z)} \end{array}\right) \middle| \tilde{C}\right) \\
&\quad + \left(\begin{array}{c|c} \mathbf{I}_{K_X - K_Z} & \tilde{C} \\ 0_{K_Z \times (K_X - K_Z)} & \end{array}\right) \begin{pmatrix} d & 0_{(K_X - K_Z) \times K_Z} \\ 0_{K_Z \times (K_X - K_Z)} & \tilde{E} \end{pmatrix}, \tilde{B}^2 \Bigg) \\
&= \left(\left(\tilde{B} \begin{pmatrix} \mathbf{I}_{K_X - K_Z} \\ 0_{K_Z \times (K_X - K_Z)} \end{pmatrix} \middle| \tilde{B}\tilde{C}\right) + \left(\begin{array}{c|c} d & \\ 0_{K_Z \times (K_X - K_Z)} & \end{array}\right) \tilde{C}\tilde{E}\right), \tilde{B}^2 \Bigg) \\
&= \left(\left(\tilde{B} \begin{pmatrix} \mathbf{I}_{K_X - K_Z} \\ 0_{K_Z \times (K_X - K_Z)} \end{pmatrix} + \begin{pmatrix} d \\ 0_{K_Z \times (K_X - K_Z)} \end{pmatrix} \middle| \tilde{B}\tilde{C} + \tilde{C}\tilde{E}\right), \tilde{B}^2 \Bigg) \\
&= \left(\tilde{B} \begin{pmatrix} \mathbf{I}_{K_X - K_Z} \\ 0_{K_Z \times (K_X - K_Z)} \end{pmatrix} + \begin{pmatrix} d \\ 0_{K_Z \times (K_X - K_Z)} \end{pmatrix}, \tilde{B}\tilde{C} + \tilde{C}\tilde{E}, \tilde{B}^2 \right),
\end{aligned}
$$

whenever $\hat{C}$ is invertible. Hence (U_1, U_2) also satisfies Equation (A.15), whenever $\hat{C}$ is invertible.

Second, keep in mind how we, in the case $K_Z = K_X$, constructed the sample ϕ such that $q(\alpha; \phi)$ had $2K_X$ distinct roots. We used equations (A.23) to (A.25). Note that the way we constructed $\hat{B}, \hat{C}, \hat{E}$ here, there has to be a ϕ such that these equations hold true for $\hat{B}, \hat{C}, \hat{E}$ instead of $\tilde{B}, \tilde{C}, \tilde{E}$. Now with the analogous calculation as in the case $K_Z = K_X$, it follows that for this ϕ, $q(\alpha; \phi)$ has $2K_X$ distinct roots.

□

A.4. Algorithm 1 in detail

Here we describe Algorithm 1 introduced in Section 3.7.1 in detail. The approach is similar to the one in [Oh et al., 2005].

A.4.1. The Likelihood and its approximation

Here we assume the general model specified in Section 3.4.1 and additionally that for each $i = 1, ..., K$ the density p_{n_i} of the noise term N_t^i is a mixture of p_i Gaussians, i.e., $p_{n_i} = \sum_{c=1}^{p_i} \pi_{i,c} \mathcal{N}(n_i | \mu_{i,c}, \sigma_{i,c}^2)$, where $\pi_{i,c} \geq 0$, $\sum_{c=1}^{p_i} \pi_{i,c} = 1$. In what follows, we denote the values of the sample $X_{1:L}$ by $x_{1:L}$, the values of the hidden variables $Z_{1:L}$ by $z_{1:L}$, and the values of the vectors $V_{1:L}^X, V_{1:L}^Z$ that select the mixture component of $N_{1:L}^X, N_{1:L}^Z$ by $v_{1:L}^X, v_{1:L}^Z$.

We can write down the complete-data likelihood as

$$p(x_{1:L}, z_{1:L}, v_{1:L}^X, v_{1:L}^Z) = \left[\prod_{l=1}^{L} p(v_l^X) p(v_l^Z)\right] p(z_1 | v_1^Z) \left[\prod_{l=2}^{L} p(z_l | z_{l-1}, x_{l-1}, v_l^Z)\right] p(x_1 | v_1^X) \left[\prod_{l=2}^{L} p(x_l | x_{l-1}, z_{l-1}, v_l^X)\right], \tag{A.27}$$

where

$$p(v_l^X) = \prod_{i=1}^{K_X} p(v_{l,i}^X) = \prod_{i=1}^{K_X} \pi_{i+K_Z, v_{l,i}^X}, \tag{A.28}$$

$$p(v_l^Z) = \prod_{i=1}^{K_Z} p(v_{l,i}^Z) = \prod_{i=1}^{K_Z} \pi_{i, v_{l,i}^Z}, \tag{A.29}$$

$$p(x_l | x_{l-1}, z_{l-1}, v_l^X) = \mathcal{N}(x_l | B x_{l-1} + C z_{l-1} + \mu_{v_l^X}, \Sigma_{v_l^X}), \tag{A.30}$$

$$p(z_l | z_{l-1}, x_{l-1}, v_l^Z) = \mathcal{N}(z_l | E z_{l-1} + D x_{l-1} + \mu_{v_l^Z}, \Sigma_{v_l^Z}), \tag{A.31}$$

$$\mu_{v_l^X} = (\mu_{K_Z+1,v_{l,1}^X}, ..., \mu_{K,v_{l,K_X}^X})^\intercal, \mu_{v_l^Z} = (\mu_{1,v_{l,1}^Z}, ..., \mu_{K_Z,v_{l,K_Z}^Z})^\intercal, \tag{A.32}$$

$$\Sigma_{v_l^X} = \text{diag}(\sigma^2_{K_Z+1,v_{l,1}^X}, ..., \sigma^2_{K,v_{l,K_X}^X}),\ \Sigma_{v_l^Z} = \text{diag}(\sigma^2_{1,v_{l,1}^Z}, ..., \sigma^2_{K_Z,v_{l,K_Z}^Z}). \tag{A.33}$$

Instead of maximizing the marginal likelihood $p(x_{1:L})$, we maximize the EM lower bound of $p(x_{1:L})$, which leads to the EM algorithm. In the E-step, the posterior of the hidden variables $p(z_{1:L}, v_{1:L}^X, v_{1:L}^Z|x_{1:L})$ is intractable because the number of Gaussian mixtures grows exponentially with the length of the time series. Thus, approximations must be made to make the problem tractable. We use a factorized approximate posterior

$$p(z_{1:L}, v_{1:L}^X, v_{1:L}^Z|x_{1:L}) \approx q(z_{1:L}|x_{1:L})q(v_{1:L}^X, v_{1:L}^Z|x_{1:L})$$

to approximate the true posterior based on the mean-field assumption. Then the variational EM lower bound can be written as

$$\begin{aligned}
\mathcal{L} = &\sum_{v_{1:L}^X, v_{1:L}^Z} q(v_{1:L}^X, v_{1:L}^Z|x_{1:L}) \int dz_{1:L}\ q(z_{1:L}|x_{1:L}) \ln p(x_{1:L}, z_{1:L}, v_{1:L}^X, v_{1:L}^Z) \\
&- \sum_{v_{1:L}^X, v_{1:L}^Z} q(v_{1:L}^X, v_{1:L}^Z|x_{1:L}) \ln q(v_{1:L}^X, v_{1:L}^Z|x_{1:L}) - \int dz_{1:L}\ q(z_{1:L}|x_{1:L}) \ln q(z_{1:L}|x_{1:L}) \\
= &\sum_{l=1}^{L} \sum_{v_l^X} q(v_l^X|x_{1:L}) \ln p(v_l^X) + \sum_{l=1}^{L} \sum_{v_l^Z} q(v_l^Z|x_{1:L}) \ln p(v_l^Z) \qquad \text{(A.34)} \\
&+ \sum_{v_1^Z} q(v_1^Z|x_{1:L}) \int dz_1\ q(z_1|x_{1:L}) \ln p(z_1|v_1^Z) \\
&+ \sum_{l=2}^{L} \sum_{v_l^Z} q(v_l^Z|x_{1:L}) \int dz_l dz_{l-1}\ q(z_l, z_{l-1}|x_{1:L}) \ln p(z_l|z_{l-1}, x_{l-1}, v_l^Z) \qquad \text{(A.35)} \\
&+ \sum_{v_1^X} q(v_1^X|x_{1:L}) \ln p(x_1|v_1^X) \\
&+ \sum_{l=2}^{L} \sum_{v_l^X} q(v_l^X|x_{1:L}) \int dz_{l-1} q(z_{l-1}|x_{1:L}) \ln p(x_l|x_{l-1}, z_{l-1}, v_l^X) \\
&- \sum_{v_{1:L}^X, v_{1:L}^Z} q(v_{1:L}^X, v_{1:L}^Z|x_{1:L}) \ln q(v_{1:L}^X, v_{1:L}^Z|x_{1:L}) \qquad \text{(A.36)} \\
&- \int dz_{1:L}\ q(z_{1:L}|x_{1:L}) \ln q(z_{1:L}|x_{1:L}) \qquad \text{(A.37)}
\end{aligned}$$

A.4.2. The algorithm

In the variational E step, $q(z_{1:L}|x_{1:L})$ and $q(v^X_{1:L}, v^Z_{1:L}|x_{1:L})$ are updated alternately by maximizing the variational lower bound. The update rules are as follows

$$q(v^X_{1:L}, v^Z_{1:L}|x_{1:L}) \leftarrow \frac{1}{c_{v^X v^Z}} \exp\left\langle \ln p(x_{1:L}, z_{1:L}, v^X_{1:L}, v^Z_{1:L})\right\rangle_{q(z_{1:L}|x_{1:L})}, \tag{A.38}$$

$$q(z_{1:L}|x_{1:L}) \leftarrow \frac{1}{c_z} \exp\left\langle \ln p(x_{1:L}, z_{1:L}, v^X_{1:L}, v^Z_{1:L})\right\rangle_{q(v^X_{1:L}, v^Z_{1:L}|x_{1:L})} \tag{A.39}$$

In (A.38), the expectation of the log-likelihood with respect to $q(z_{1:L}|x_{1:L})$ is calculated as

$$\begin{aligned}
&\left\langle \ln p(x_{1:L}, z_{1:L}, v^X_{1:L}, v^Z_{1:L})\right\rangle_{q(z_{1:L}|x_{1:L})} \qquad (A.40)\\
&= \sum_{l=1}^{L}\sum_{i=1}^{K_X} \ln p(v^X_{l,i}) + \sum_{l=1}^{L}\sum_{i=1}^{K} \ln p(v^Z_{l,i})\\
&- \frac{1}{2}\sum_{i=1}^{v_Z}\left(\frac{\left\langle (z_{1,i} - \mu_{i,v^Z_{1,i}})^2\right\rangle_{q(z_{1,i}|x_{1:L})}}{\sigma^2_{i,v^Z_{1,i}}} + 2\ln\sigma_{i,v^Z_{1,i}}\right)\\
&- \frac{1}{2}\sum_{l=2}^{L}\sum_{i=1}^{K_Z}\left(\frac{\left\langle \left(z_{l,i} - (Ez_{l-1})_i - (Dx_{l-1})_i - \mu_{i,v^Z_{l,i}}\right)^2\right\rangle_{q(z_l, z_{l-1}|x_{1:L})}}{\sigma^2_{i,v^Z_{l,i}}} + 2\ln\sigma_{i,v^Z_{l,i}}\right)\\
&- \frac{1}{2}\sum_{i=1}^{K_X}\left(\frac{(x_{1,i} - \mu_{i+K_Z,v^X_{1,i}})^2}{\sigma^2_{i+K_Z,v^X_{1,i}}} + 2\ln\sigma_{i+K_Z,v^X_{1,i}}\right) + \text{const}\\
&- \frac{1}{2}\sum_{l=2}^{L}\sum_{i=1}^{K_X}\left(\frac{\left\langle \left(x_{l,i} - (Cz_{l-1})_i - (Bx_{l-1})_i - \mu_{i+K_Z,v^X_{l,i}}\right)^2\right\rangle_{q(z_{l-1}|x_{1:L})}}{\sigma^2_{i+K_Z,v^X_{l,i}}} + 2\ln\sigma_{i+K_Z,v^X_{l,i}}\right). \qquad (A.41)
\end{aligned}$$

It can be seen that $q(v^X_{1:L}, v^Z_{1:L}|x_{1:L})$ further factorizes as $\left[\prod_l \prod_i q(v^X_{l,i})\right]\left[\prod_l \prod_i q(v^Z_{l,i})\right]$, which means the posterior $q(v^X_{1:L}, v^Z_{1:L}|x_{1:L})$ can be calculated separately for each channel. The computational complexity is linear in the time series length, the number of time series channels, and the number of Gaussian mixtures in each channel.

(A.39) can be further expressed as

$$\begin{aligned}
&\left\langle \ln p(x_{1:L}, z_{1:L}, v^X_{1:L}, v^Z_{1:L}) \right\rangle_{q(v^X_{1:L}, v^Z_{1:L}|x_{1:L})} \qquad \text{(A.42)}\\
&= -\frac{1}{2}\sum_{i=1}^{K_Z} z_{1,i}^2 \left(\sum_{v^Z_{1,i}} \frac{q(v^Z_{1,i})}{\sigma^2_{i,v^Z_{1,i}}} \right) + \sum_{i=1}^{K_Z} z_{1,i} \left(\sum_{v^Z_{1,i}} \frac{q(v^Z_{1,i})\mu_{i,v^Z_{1,i}}}{\sigma^2_{i,v^Z_{1,i}}} \right)\\
&- \frac{1}{2}\sum_{l=2}^{L}\sum_{i=1}^{K_Z} \left(z_{l,i} - (Ez_{l-1})_i\right)^2 \left(\sum_{v^Z_{l,i}} \frac{q(v^Z_{l,i})}{\sigma^2_{i,v^Z_{l,i}}} \right)\\
&+ \sum_{l=2}^{L}\sum_{i=1}^{K_Z} \left(z_{l,i} - (Ez_{l-1})_i\right) \left(\sum_{v^Z_{l,i}} \frac{q(v^Z_{l,i})\left((Dx_{l-1})_i + \mu_{i,v^Z_{l,i}}\right)}{\sigma^2_{i,v^Z_{l,i}}} \right)\\
&- \frac{1}{2}\sum_{l=2}^{L}\sum_{i=1}^{K_X} \left(x_{l,i} - (Bx_{l-1})_i - (Cz_{l-1})_i\right)^2 \left(\sum_{v^X_{l,i}} \frac{q(v^X_{l,i})}{\sigma^2_{i+K_Z,v^X_{l,i}}} \right)\\
&+ \sum_{l=2}^{L}\sum_{i=1}^{K_X} \left(x_{l,i} - (Bx_{l-1})_i - (Cz_{l-1})_i\right) \left(\sum_{v^X_{l,i}} \frac{q(v^X_{l,i})\mu_{i+K_Z,v^X_{l,i}}}{\sigma^2_{i+K_Z,v^X_{l,i}}} \right) + \text{const}, \qquad \text{(A.43)}
\end{aligned}$$

which has the form of the joint log-likelihood function of a time-varying linear dynamical system (LDS). The marginal posteriors $p(z_l|x_{1:L})$ and $p(z_l, z_{l-1}|x_{1:L})$ can be obtained by Kalman filter and smoothing algorithms.

In the M-step, we maximize the variational lower bound with respect to the parameters given the marginal posterior distributions from the E-step. The update rules for the parameters are given as follows

$$\pi_{i,c} = \begin{cases} \frac{1}{L}\sum_{l=1}^{L} q(v^Z_{l,i} = c|x_{1:L}), & i = 1, ..., K_Z, \\ \frac{1}{L}\sum_{l=1}^{L} q(v^X_{l,i-K_Z} = c|x_{1:L}), & i = K_Z+1, ..., K, \end{cases} \qquad \text{(A.44)}$$

$$\mu_{i,c} = \begin{cases} \frac{1}{\sum_{l=1}^{L} q(v_{l,i}^{Z}=c|x_{1:L})} \left(q(v_{1,i}^{Z}=c|x_{1:L}) \left(\langle z_{1,i}\rangle_{q(z_{1,i}|x_{1:L})}\right)\right. \\ \left. + \sum_{l=2}^{L} q(v_{l,i}^{Z}=c|x_{1:L}) \left(\langle z_{l,i}\rangle_{q(z_{l,i}|x_{1:L})} - \left(E\langle z_{l-1}\rangle_{q(z_{l-1}|x_{1:L})}\right)_i - (Dx_{l-1})_i\right)\right), \\ i = 1, ..., K_Z, \\ \frac{1}{\sum_{l=1}^{L} q(v_{l,i-K_Z}^{X}=c|x_{1:L})} \left(q(v_{1,i-K_Z}^{X}=c|x_{1:L})x_{1,i-K_Z}\right. \\ \left. + \sum_{l=2}^{L} q(v_{l,i-K_Z}^{X}=c|x_{1:L}) \left(x_{l,i-K_Z} - \left(C\langle z_{l-1}\rangle_{q(z_{l-1}|x_{1:L})}\right)_{i-K_Z} - (Bx_{l-1})_i\right)\right), \\ i = K_Z+1, ..., K, \end{cases} \tag{A.45}$$

$$\sigma_{i,c}^2 = \begin{cases} \frac{1}{\sum_{l=1}^{L} q(v_{l,i}^{Z}=c|x_{1:L})} \left(q(v_{1,i}^{Z}=c|x_{1:L}) \left(\left\langle z_{1,i}^2 - 2\mu_{i,c}z_{1,i}\right\rangle_{q(z_{1,i}|x_{1:L})}\right)\right. \\ \left. + \sum_{l=2}^{L} q(v_{l,i}^{Z}=c|x_{1:L}) \left\{[z_{l,i} - (Ez_{l-1})_i - (Dx_{l-1})_i]^2_{q(z_l,z_{l-1}|x_{1:L})}\right.\right. \\ \left.\left. -2\mu_{i,c}\left[\langle z_{l,i}\rangle_{q(z_{l,i}|x_{1:L})} - \left(E\langle z_{l-1}\rangle_{q(z_{l-1}|x_{1:L})}\right)_i - (Dx_{l-1})_i\right]\right\} + \mu_{i,c}^2, \\ i = 1, ..., K_Z, \\ \frac{1}{\sum_{l=1}^{L} q(v_{l,i-K_Z}^{X}=c|x_{1:L})} \left(q(v_{1,i-K_Z}^{X}=c|x_{1:L}) \left(x_{1,i-K_Z}^2 - 2\mu_{i,c}x_{1,i-K_Z}\right)\right. \\ \left. + \sum_{l=2}^{L} q(v_{l,i-K_Z}^{X}=c|x_{1:L}) \left\{\left[x_{l,i-K_Z} - (Cz_{l-1})_{i-K_Z} - (Bx_{l-1})_{i-K_Z}\right]^2_{q(z_{l-1}|x_{1:L})}\right.\right. \\ \left.\left. -2\mu_{i,c}\left[x_{l,i-K_Z} - \left(C\langle z_{l-1}\rangle_{q(z_{l-1}|x_{1:L})}\right)_{i-K_Z} - (Bx_{l-1})_{i-m}\right]\right\}\right) + \mu_{i,c}^2, \\ i = K_Z+1, ..., K, \end{cases} \tag{A.46}$$

$$E_i = \left(\sum_{l=2}^{L}\sum_{v_{l,i}^{Z}} \frac{q(v_{l,i}^{Z}|x_{1:L})}{\sigma_{i,v_{l,i}^{Z}}^2} \left\langle z_{l-1}z_{l-1}^{\top}\right\rangle_{q(z_{l-1}|x_{1:L})}\right)^{-1} \tag{A.47}$$

$$\left(\sum_{l=2}^{L}\sum_{v_{l,i}^{Z}} \frac{q(v_{l,i}^{Z}|x_{1:L})}{\sigma_{i,v_{l,i}^{Z}}^2} \left(\langle z_{l-1}z_{l,i}\rangle_{q(z_l,z_{l-1}|x_{1:L})} - \right.\right.$$

$$\left.\left. \langle z_{l-1}\rangle_{q(z_{l-1}|x_{1:L})} (Dx_{l-1})_i - \langle z_{l-1}\rangle_{q(z_{l-1}|x_{1:L})} \mu_{i,v_{l,i}^{Z}}\right)\right), \tag{A.48}$$

$$D_i = \left(\sum_{l=2}^{L} \sum_{v_{l,i}^Z} \frac{q(v_{l,i}^Z | x_{1:L})}{\sigma^2_{i,v_{l,i}^Z}} x_{l-1} x_{l-1}^\top \right)^{-1} \left(\sum_{l=2}^{L} \sum_{v_{l,i}^Z} \frac{q(v_{l,i}^Z | x_{1:L})}{\sigma^2_{i,v_{l,i}^Z}} x_{l-1} \left(\langle z_{l,i} \rangle_{q(z_{l,i}|x_{1:L})} - \right. \right.$$
$$\left. \left. (E \langle z_{l-1} \rangle_{q(z_{l-1}|x_{1:L})})_i - \mu_{i,v_{l,i}^Z} \right) \right), \tag{A.49}$$

$$C_i = \left(\sum_{l=2}^{T} \sum_{v_{l,i}^X} \frac{q(v_{l,i}^X | x_{1:L})}{\sigma^2_{i+K_Z,v_{l,i}^X}} \left\langle z_{l-1} z_{l-1}^\top \right\rangle_{q(z_{l-1}|x_{1:L})} \right)^{-1} \tag{A.50}$$
$$\left(\sum_{l=2}^{L} \sum_{v_{l,i}^X} \frac{q(v_{l,i}^X | x_{1:L})}{\sigma^2_{i+K_Z,v_{l,i}^X}} \langle z_{l-1} \rangle_{q(z_{l-1}|x_{1:L})} (x_{l,i} - \right.$$
$$\left. (B x_{l-1})_i - \mu_{i+K_Z,v_{l,i}^X} \right) \right), \tag{A.51}$$

$$B_i = \left(\sum_{l=2}^{L} \sum_{v_{l,i}^X} \frac{q(v_{l,i}^X | x_{1:L})}{\sigma^2_{i+K_Z,v_{l,i}^X}} x_{l-1} x_{l-1}^\top \right)^{-1} \left(\sum_{l=2}^{L} \sum_{v_{l,i}^X} \frac{q(v_{l,i}^X | x_{1:L})}{\sigma^2_{i+K_Z,v_{l,i}^X}} x_{l-1} (x_{l,i} - \right.$$
$$\left. \left(C \langle z_{l-1} \rangle_{q(z_{l-1}|x_{1:L})} \right)_i - \mu_{i+K_Z,v_{l,i}^X} \right) \right), \tag{A.52}$$

where E_i, D_i, C_i, and B_i denote the i-th row of E, D, C, and B respectively.

Appendix B.

Proof for Chapter 4

Most proofs for Chapter 4 were already contained in that chapter. Only the proof of Lemma 4.1 was postponed and we present it here.

B.1. Proof of Lemma 4.1

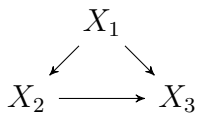


Figure B.1.: W.l.o.g. we assume this causal DAG.

Proof of Lemma 4.1. We only treat the continuous case, the discrete case is straight forward.

Recall our assumption that given a causal model M with causal DAG $G = (V, E)$, for each $X_i \in V$, the random variable $f_i(\mathrm{pa}_i, N_i)$ has a density $q_i(x_i; \mathrm{pa}_i)$ w.r.t. the Lebesgue measure. Note that this implies, that also in any post-interventional model $M_{\mathrm{do}\, X_i = x}$, the random variables $f_i^{M_{\mathrm{do}\, X_i = x}}(\mathrm{pa}_i^{M_{\mathrm{do}\, X_i = x}}, N_i)$ have densities w.r.t. the Lebesgue measure which can easily obtained from the $q_i(x_i; \mathrm{pa}_i)$. Hence, w.l.o.g., we only prove the lemma w.r.t. M.

In what follows, we will only consider the case where the causal DAG is fully connected, the other cases work similarly. W.l.o.g. we assume the DAG in Figure B.1.

Let $q(x_1, x_2, x_3) := \prod_i q_i(x_i; \mathrm{pa}_i)$.

To see that $q(x_1, x_2, x_3)$ factorizes according to G, note that

$$\begin{aligned} p(x_3|x_2, x_1) &= \frac{q_3(x_3; x_1, x_2)q_2(x_2; x_1)q_1(x_1)}{\int q_3(x_3; x_1, x_2)q_2(x_2; x_1)q_1(x_1)\mathrm{d}x_3} \\ &= \frac{q_3(x_3; x_1, x_2)q_2(x_2; x_1)q_1(x_1)}{q_2(x_2; x_1)q_1(x_1)} \\ &= q_3(x_3; x_1, x_2). \end{aligned}$$

Similarly one calculates $p(x_2|x_1) = q(x_2; x_1)$ and $p(x_1) = q(x_1)$.

It remains to show that $q(x_1, x_2, x_3)$ it is a density for the joint distribution $P(X_1, X_2, X_3)$.

Keep in mind that for measurable f, Y we have [Bogachev, 2007]

$$\int Y(s)\mathrm{d}P_{f(N)}(s) = \int Y(f(r))\mathrm{d}P_N(r). \tag{B.1}$$

Let $[\cdot]$ denote the characteristic function (i.e. it equals 1 if the statement inside the brackets is true and 0 otherwise). Now we can calculate

$$\int_{-\infty}^{a}\int_{-\infty}^{b}\int_{-\infty}^{c} q_1(x_1)q_2(x_2;x_1)q_3(x_3;x_1,x_2)\mathrm{d}x_3\mathrm{d}x_2\mathrm{d}x_1 \tag{B.2}$$

$$= \int [x_1 \leq a]\int [x_2 \leq b]\int [x_3 \leq c]\mathrm{d}P_{f_3(x_1,x_2,N_3)}(x_3)\mathrm{d}P_{f_2(x_1,N_2)}(x_2)\mathrm{d}P_{f_1(N_1)}(x_1) \tag{B.3}$$

$$= \int [x_1 \leq a]\int [x_2 \leq b]\int [f_3(x_1,x_2,n_3) \leq c]\mathrm{d}P_{N_3}(n_3)\mathrm{d}P_{f_2(x_1,N_2)}(x_2)\mathrm{d}P_{f_1(N_1)}(x_1) \tag{B.4}$$

$$= \int [x_1 \leq a]\int [f_2(x_1,n_2) \leq b]\int [f_3(x_1,f_2(x_1,n_2),n_3) \leq c] \tag{B.5}$$

$$\mathrm{d}P_{N_3}(n_3)\mathrm{d}P_{N_2}(n_2)\mathrm{d}P_{f_1(N_1)}(x_1) \tag{B.6}$$

$$= \int [f_1(n_1) \leq a]\int [f_2(f_1(n_1),n_2) \leq b] \tag{B.7}$$

$$\int [f_3(f_1(n_1),f_2(f_1(n_1),n_2),n_3) \leq c]\mathrm{d}P_{N_3}(n_3)\mathrm{d}P_{N_2}(n_2)\mathrm{d}P_{N_1}(n_1) \tag{B.8}$$

$$= \int [f_1(n_1) \leq a][f_2(f_1(n_1),n_2) \leq b] \tag{B.9}$$

$$[f_3(f_1(n_1),f_2(f_1(n_1),n_2),n_3) \leq c]\mathrm{d}P_{N_1,N_2,N_3}(n_1,n_2,n_3) \tag{B.10}$$

$$= \mathbb{E}\Big[[f_1(N_1) \leq a][f_2(f_1(N_1),N_2) \leq b][f_3(f_1(N_1),f_2(f_1(N_1),N_2),N_3) \leq c]\Big] \tag{B.11}$$

$$= P(X_1 \leq a, X_2 \leq b, X_3 \leq c), \tag{B.12}$$

where equations (B.4), (B.6), (B.8) follow by applying Equation (B.1), and Equation (B.10) follow from the independence of the noise terms N_i.

This proofs that $q(x_1,x_2,x_3)$ is a density of $P(X_1,X_2,X_3)$ w.r.t. the Lebesgue measure.

□

Appendix C.

Proofs for Chapter 5

Here we present proofs for Chapter 5.

C.1. Generalized version and proof of Proposition 5.1

We start by stating and proving a generalization of Proposition 5.1.

Proposition C.1 (Generalization of Proposition 5.1)**.** *Let M_0 be a FCM over discrete variables that induces a GCM M. Let the triple (X, Y, Z) of (sets of) variables in M be such that $(Y \perp\!\!\!\perp An(Z)|Z)_M$ (i.e., are d-separated [Pearl, 2000]) and X does not influence $W := Z \setminus X$. Let E be an arbitrary set of variables in M. Let*

$$p^Z(Y_{\mathrm{do}\, X=x} = y|e) := \sum_w p(y|\mathrm{do}\, X = x, w)p(w|e). \tag{C.1}$$

Then

$$\mathrm{D}(p(Y_{\mathrm{do}\, X=x}|E)\|p^Z(Y_{\mathrm{do}\, X=x}|E)) \leq \mathrm{H}(E|Z) \tag{C.2}$$

(where $p(Y_{\mathrm{do}\, X=x}|E)$ is defined w.r.t. M_0 and $p^Z(Y_{\mathrm{do}\, X=x}|E)$ w.r.t. M).

This is a generalization of Proposition 5.1. To see this, let Z denote the set of root nodes of M. This implies

$$p^Z(Y_{\mathrm{do}\, X=x} = y|e) = \tilde{p}(Y_{\mathrm{do}\, X=x} = y|e)$$

for $p^Z(Y_{\text{do}\,X=x} = y|e)$ as defined above and $\tilde{p}(Y_{\text{do}\,X=x} = y|e)$ as defined as in Section 5.3.1. But Proposition C.1 above applied to this $p^Z(Y_{\text{do}\,X=x} = y|e)$ coincides with Proposition 5.1.

Proof. Let U_1 be the set (tuple) of background variables that influence Z and $U_0 = U \backslash U_1$. Then

$$
\begin{aligned}
& p^Z(Y_{\text{do}\,X=x} = y|e) && \text{(C.3)}\\
&= \sum_w p(y|\text{do}\,X = x, w)p(w|e) && \text{(C.4)}\\
&= \sum_{w,u_0} p(y|\text{do}\,X = x, w, u_0)p(w|e)p(u_0|\text{do}\,X = x, w) && \text{(C.5)}\\
&= \sum_{w,u_0} p(y|\text{do}\,X = x, w, u_0)p(w|e)p(u_0|w) && \text{(C.6)}\\
&= \sum_{w,u_0} p(y|\text{do}\,X = x, w, u_0)p(w|e)p(u_0), && \text{(C.7)}
\end{aligned}
$$

where Equation (C.6) is due to the fact that the distribution of U_0 is invariant and X does not influence W, so W can be written as the same function of U_0 in M_0 and $(M_0)_{\text{do}\,X=x}$; and Equation (C.7) is due to the fact that $W \subset Z$ and $Z \perp\!\!\!\perp U_0$ by definition of U_0.

On the other hand, we have

$$
\begin{aligned}
& p(Y_{\mathrm{do}\, X=x} = y|e) && \text{(C.8)}\\
&= \sum_{u:p(u,e)>0} p(y|\mathrm{do}\, X = x, u)p(u|e) && \text{(C.9)}\\
&= \sum_{u_0,u_1:p(u_0,u_1,e)>0} p(y|\mathrm{do}\, X = x, u_0, u_1)p(u_0, u_1|e) && \text{(C.10)}\\
&= \sum_{u_0,u_1,w:p(u_0,u_1,e)>0} p(y, w|\mathrm{do}\, X = x, u_0, u_1)p(u_0, u_1|e) && \text{(C.11)}\\
&= \sum_{u_0,u_1,w:p(u_0,u_1,e),p(u_0,u_1,w)>0} p(y, w|\mathrm{do}\, X = x, u_0, u_1)p(u_0, u_1|e) && \text{(C.12)}\\
&= \sum_{u_0,u_1,w:p(u_0,u_1,e),p(u_0,u_1,w)>0} p(y|\mathrm{do}\, X = x, u_0, u_1, w)p(w|\mathrm{do}\, X = x, u_1, u_0)p(u_0, u_1|e) && \text{(C.13)}\\
&= \sum_{u_0,u_1,w:p(u_0,u_1,e),p(u_0,u_1,w)>0} p(y|\mathrm{do}\, X = x, u_0, w)p(w|\mathrm{do}\, X = x, u_1, u_0)p(u_0, u_1|e) && \text{(C.14)}\\
&= \sum_{u_0,u_1,w:p(u_0,u_1,e),p(u_0,u_1,w)>0} p(y|\mathrm{do}\, X = x, u_0, w)p(w|u_1, u_0)p(u_0, u_1|e) && \text{(C.15)}\\
&= \sum_{u_0,u_1,w:p(u_0,u_1,e),p(u_0,u_1,w)>0} p(y|\mathrm{do}\, X = x, u_0, w)p(w|u_0, u_1, e)p(u_0, u_1|e) && \text{(C.16)}\\
&= \sum_{u_0,u_1,w:p(u_0,u_1,e),p(u_0,u_1,w)>0} p(y|\mathrm{do}\, X = x, u_0, w)p(w, u_0, u_1|e) && \text{(C.17)}\\
&= \sum_{u_0,u_1,w:p(u_0,u_1,e,w)>0} p(y|\mathrm{do}\, X = x, u_0, w)p(w, u_0, u_1|e) && \text{(C.18)}\\
&= \sum_{u_0,w:p(u_0,e,w)>0} p(y|\mathrm{do}\, X = x, u_0, w) \sum_{u_1:p(u_0,u_1,e,w)>0} p(w, u_0, u_1|e) && \text{(C.19)}\\
&= \sum_{u_0,w:p(u_0,e,w)>0} p(y|\mathrm{do}\, X = x, u_0, w)p(w, u_0|e) && \text{(C.20)}\\
&= \sum_{u_0,w} p(y|\mathrm{do}\, X = x, u_0, w)p(w, u_0|e), && \text{(C.21)}
\end{aligned}
$$

where Equation (C.14) is due to Markovianity and $(Y \perp\!\!\!\perp An(Z)|Z)_M$, which implies $(Y \perp\!\!\!\perp U_1|Z)_M$, and thus $(Y \perp\!\!\!\perp U_1|W)_{M_{\mathrm{do}\, X=x}}$, Equation (C.15) follows from the fact that X does not influence W, Equation (C.16) follows from the fact that U_1 already determines W.

Note that $p(w|e)p(u_0) = 0$ implies $p(w, u_0|e) = 0$ and therefore

$$\mathrm{D}[p(Y_{\mathrm{do}\,X=x}|E)\|p^W(Y_{\mathrm{do}\,X=x}|D)]$$

is defined.

Now we can calculate

$$\begin{aligned}
&\mathrm{D}[p(Y_{\mathrm{do}\,X=x}|E)\|p^Z(Y_{\mathrm{do}\,X=x}|E)] && \text{(C.22)}\\
&= \sum_e p(e)\mathrm{D}[p(Y_{\mathrm{do}\,X=x}|e)\|p^Z(Y_{\mathrm{do}\,X=x}|e)] && \text{(C.23)}\\
&\leq \sum_e p(e)\mathrm{D}[p(W, U_0|e)\|p(W|e)p(U_0)] && \text{(C.24)}\\
&= \sum_{e,w,u_0} p(w, u_0, e)\log\frac{p(w, u_0|e)}{p(w|e)p(u_0)} && \text{(C.25)}\\
&= \sum_{e,w,u_0} p(w, u_0, e)\log\frac{p(w, u_0, e)}{p(w, e)p(u_0)} && \text{(C.26)}\\
&= \mathrm{I}(W, E : U_0) && \text{(C.27)}\\
&\leq \mathrm{I}(Z, E : U_0) && \text{(C.28)}\\
&= \mathrm{I}(Z : U_0) + \mathrm{I}(E : U_0|Z) && \text{(C.29)}\\
&= 0 + \mathrm{H}(E|Z) - \mathrm{H}(E|U_0, Z), && \text{(C.30)}
\end{aligned}$$

where inequality (C.24) follows from the monotonicity (which follows from the chain rule) of the Kullback-Leibler divergence [Cover and Thomas, 1991] together with equations (C.21) and (C.7), Equation (C.29) is the chain rule for mutual information, and $\mathrm{I}(W : U_0) = 0$ is due to U_0 not influencing W and Markovianity.

□

Note that, if we chose the set Z in the above proposition such that it is as "close" (in the causal diagram) to Y as possible, this could yield better approximations $p^Z(Y_{\mathrm{do}\,X=x} = y|e)$ than simply letting Z be the root nodes, as done in $\bar{p}(Y_{\mathrm{do}\,X=x} = y|e)$. We leave this as a question for future work.

C.2. Proof of Proposition 5.2

Here we give a proof for Proposition 5.2.

Proof. We calculate

$$
\begin{aligned}
&\mathrm{D}(p(Z)\|\bar{p}(Z)) && \text{(C.31)}\\
&\leq \mathrm{D}(p(X_0,\ldots,X_K,C)\|p(C)\Pi_k p(X_k|C))\\
&= \mathrm{D}(p(C)p(X_0|C)p(X_1|X_0,C)\cdots p(X_K|X_0,\ldots,X_{K-1},C)\|p(X)\Pi_k p(X_k|C))\\
&= \sum_{x_0,\ldots,x_K,c} p(x_0,\ldots,x_K,c)\log\frac{p(c)}{p(c)}\frac{p(x_0|c)}{p(x_0|c)}\frac{p(x_1|x_0,c)}{p(x_1|c)}\frac{p(x_2|x_0,x_1,c)}{p(x_2|c)}\cdots\\
&\quad\cdot\frac{p(x_K|x_0,\ldots,x_{K-1},c)}{p(x_K|c)}\\
&= \sum_{x_0,\ldots,x_K,c} p(x_0,\ldots,x_K,c)\log\frac{p(c)}{p(c)}\frac{p(x_0|c)}{p(x_0|c)}\frac{p(x_1,x_0|c)}{p(x_1|c)p(x_0|c)}\frac{p(x_2,x_0,x_1|c)}{p(x_2|c)p(x_0,x_1|c)}\cdots\\
&\quad\cdot\frac{p(x_K,x_0,\ldots,x_{K-1}|c)}{p(x_K|c)p(x_0,\ldots,x_{K-1}|c)}\\
&= \mathrm{I}(X_1:X_0|C)+\mathrm{I}(X_2:X_0,X_1|C)+\ldots+\mathrm{I}(X_K:X_0,\ldots,X_{K-1}|C)\\
&\leq \mathrm{H}(X_1|C)+\mathrm{H}(X_2|C)+\ldots+\mathrm{H}(X_K|C), && \text{(C.32)}
\end{aligned}
$$

where inequality (C.31) follows from the monotonicity (which follows from the chain rule) of the Kullback-Leibler divergence [Cover and Thomas, 1991]. □

Bibliography

S. Angel, H. Ballani, T. Karagiannis, G. O'Shea, and E. Thereska. End-to-end performance isolation through virtual datacenters. In *11th USENIX Symposium on Operating Systems Design and Implementation (OSDI 14)*, pages 233–248, 2014.

J. Angrist, G. Imbens, and D. Rubin. Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996.

M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, et al. A view of cloud computing. *Communications of the ACM*, 53(4):50–58, 2010.

K. J. Aström and R. M. Murray. *Feedback systems: an introduction for scientists and engineers*. Princeton university press, 2010.

C. Avin, I. Shpitser, and J. Pearl. Identifiability of path-specific effects. In *Proceedings of the International Joint Conference in Artificial Intelligence*, pages 357–363, Edinburgh, Scotland, 2005.

N. Ay and D. Krakauer. Geometric robustness and biological networks. *Theory in Biosciences*, 125:93–121, 2007.

N. Ay and D. Polani. Information flows in causal networks. *Advances in Complex Systems*, 11(1):17–41, 2008.

A. Balke and J. Pearl. Counterfactual probabilities: Computational methods, bounds and applications. In *Proceedings of the Tenth international conference on Uncertainty in artificial intelligence*, pages 46–54. Morgan Kaufmann Publishers Inc., 1994.

D. Barber. *Bayesian reasoning and machine learning*. Cambridge University Press, 2012.

E. Bareinboim and J. Pearl. Transportability of causal effects: Completeness results. In *Proceedings of the 26th National Conference on Artificial Intelligence (AAAI)*, pages 698–704. AAAI Press, Menlo Park, CA., 2012.

E. Bareinboim and J. Pearl. Transportability from multiple environments with limited experiments: Completeness results. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 280–288. Curran Associates, Inc., 2014.

E. Bareinboim, A. Forney, and J. Pearl. Bandits with unobserved confounders: A causal approach. In *Advances in Neural Information Processing Systems*, pages 1342–1350, 2015.

V. Bogachev. *Measure Theory*. Springer, 2007.

L. Bottou, J. Peters, J. Quinonero-Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *The Journal of Machine Learning Research*, 14(1):3207–3260, 2013.

X. Boyen, N. Friedman, and D. Koller. Discovering the hidden structure of complex dynamic systems. In *Proceedings of the 15 th Conference on Uncertainty in Artificial Intelligence*, pages 91–100. Morgan Kaufmann, San Francisco, 1999.

K. H. Brodersen, F. Gallusser, J. Koehler, N. Remy, S. L. Scott, et al. Inferring causal impact using bayesian structural time-series models. *The Annals of Applied Statistics*, 9(1):247–274, 2015.

L. Carata. *Provenance-based computing*. PhD dissertation, University of Cambridge, 2016.

L. Carata, O. Chick, J. Snee, R. Sohan, A. Rice, and A. Hopper. Resourceful: fine-grained resource accounting for explaining service variability. Technical report, University of Cambridge, Computer Laboratory, 2014.

R. C. Chiang, J. Hwang, H. H. Huang, and T. Wood. Matrix: Achieving predictable virtual machine performance in the clouds. In *11th International Conference on Autonomic Computing (ICAC 14)*, pages 45–56, 2014.

T. Chu and C. Glymour. Search for additive nonlinear time series causal models. *JMLR*, (9):967–991, 2008.

W. Conover. *Practical Nonparametric Statistics*. John Wiley & Sons, 1971.

T. Cover and J. Thomas. *Elements of Information Theory*. Wileys Series in Telecommunications, New York, 1991.

A. Dickenstein and I. Z. Emiris. *Solving Polynomial Equations: Foundations, Algorithms, and Applications*. Springer Publishing Company, Incorporated, 1st edition, 2010.

F. Eberhardt and R. Scheines. Interventions and causal inference. *Philosophy of Science*, 74(5):981–995, 2007.

B. Efron and D. Feldman. Compliance as an Explanatory Variable in Clinical Trials. *Journal of the American Statistical Association*, 86(413):9–17, 1991.

M. Eichler. Causal inference in time series analysis. In C. Berzuini, A. Dawid, and L. Bernardinelli, editors, *Causality*, pages 327–354. John Wiley and Sons, Ltd, 2012.

D. Entner and P. O. Hoyer. On causal discovery from time series data using fci. *Probabilistic graphical models*, 2010.

J. Etesami, N. Kiyavash, and T. Coleman. Learning minimal latent directed information trees. In *IEEE International Symposium on Information Theory (ISIT)*, pages 2726–2730, 2012.

A. Falcon. Aristotle on causality. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2015 edition, 2015.

P. Geiger, D. Janzing, and B. Schölkopf. Estimating causal effects by bounding confounding. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*, pages 240–249, 2014.

P. Geiger, K. Zhang, M. Gong, D. Janzing, and B. Schölkopf. Causal inference by identification of vector autoregressive processes with hidden components. *arXiv preprint arXiv:1411.3972*, 2015a.

P. Geiger, K. Zhang, M. Gong, D. Janzing, and B. Schölkopf. Causal inference by identification of vector autoregressive processes with hidden components. In *Proceedings of 32th International Conference on Machine Learning (ICML 2015)*, 2015b.

P. Geiger, L. Carata, and B. Schoelkopf. Causal models for debugging and control in cloud computing. *arXiv preprint arXiv:1603.01581*, 2016a.

P. Geiger, L. Carata, and B. Schölkopf. Causal inference for cloud computing. *arXiv preprint arXiv:1603.01581*, 2016b.

M. Gong, K. Zhang, B. Schoelkopf, D. Tao, and P. Geiger. Discovering temporal causal relations from subsampled data. In *Proceedings of 32th International Conference on Machine Learning (ICML 2015)*, pages 1898–1906, 2015.

C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):pp. 424–438, 1969. ISSN 00129682.

P. W. Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

HowStuffWorks.com. What causes high tide and low tide? why are there two tides each day?, 2009. HowStuffWorks.com. URL: `http://science.howstuffworks.com/environmental/earth/geophysics/tide-cause.htm`. Accessed: 2016-12-06.

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(2):362 – 378, 2008. doi: http://dx.doi.org/10.1016/j.ijar.2008.02.006.

M. Hulswit. A short history of causation. *SEED Journal (Semiotics, Evolution, Energy, and Development)*, 4(3):16–42, 2004.

D. Hume and C. W. Hendel. *An inquiry concerning human understanding*, volume 49. Bobbs-Merrill Indianapolis, 1955.

A. Hyvaerinen, K. Zhang, S. Shimizu, and P. O. Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, (11):1709–1731, 2010.

G. Imbens and T. Lemieux. Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142:615–635, 2008.

G. W. Imbens and D. B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press, 2015.

J. J. E. Dennis, J. F. Traub, and R. P. Weber. The algebraic theory of matrix polynomials. *SIAM Journal on Numerical Analysis*, 13(6):831–845, 1976.

A. Jalali and S. Sanghavi. Learning the dependence graph of time series with latent factors. In *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012.

D. Janzing, E. Sgouritsa, O. Stegle, P. Peters, and B. Schölkopf. Detecting low-complexity unobserved causes. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, 2011.

D. Janzing, J. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniušis, B. Steudel, and B. Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182:1–31, 2012.

D. Janzing, D. Balduzzi, M. Grosse-Wentrup, and B. Schölkopf. Quantifying causal influences. *Annals of Statistics*, 41(5):2324–2358, 2013.

D. Janzing, R. Chaves, and B. Schölkopf. Algorithmic independence of initial condition and dynamical law in thermodynamics and causal inference. *New Journal of Physics*, 18(9):093052, 2016.

R. C. Jeffrey. *The logic of decision.* University of Chicago Press, 1990.

A. M. Kagan, Y. V. Linnik, and C. R. Rao. *Characterization Problems in Mathematical Statistics.* Wiley, New York, 1973.

I. Kant. *Kritik der reinen Vernunft.* Meiner Verlag, Hamburg„ 1998.

I. Kant and P. Guyer. *Critique of pure reason.* Cambridge University Press, 1998.

I. Kant, J. B. Schneewind, M. Baron, and S. Kagan. *Groundwork for the Metaphysics of Morals.* Yale University Press, 2002.

A. Klenke. *Probability theory: a comprehensive course*. Springer Science & Business Media, 2013.

N. Kolodny and J. Brunero. Instrumental rationality. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition, 2016.

S. Lauritzen. *Graphical Models*. Clarendon Press, Oxford, New York, Oxford Statistical Science Series edition, 1996.

D. A. Lawlor, G. Davey Smith, and S. Ebrahim. Socioeconomic position and hormone replacement therapy use: explaining the discrepancy in evidence from observational and randomized controlled trials. *American journal of public health*, 94(12):2149–2154, 2004.

D. Lee and T. Lemieux. Regression Discontinuity Designs in Economics. *Journal of Economic Literature*, 48:281–355, 2010.

J. Lemeire, E. Dirkx, and F. Verbist. Causal analysis for performance modeling of computer programs. *Scientific Programming*, 15(3):121–136, 2007.

D. Lewis. Causal decision theory. *Australasian Journal of Philosophy*, 59(1):5–30, 1981.

H. Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer, Berlin, Heidelberg, New York, oxford statistical science series edition, 2006.

C. Macleod. John stuart mill. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Fall 2016 edition, 2016.

K. Marx. Das kapital: kritik der politischen ökonomie. *Verlag von Otto Meisner, Germany*, 1885:1894, 1867.

K. Marx. *Zur Kritik der politischen ökonomie*. BoD–Books on Demand, 2014.

F. McSherry and K. Talwar. Mechanism design via differential privacy. In *Foundations of Computer Science, 2007. FOCS'07. 48th Annual IEEE Symposium on*, pages 94–103. IEEE, 2007.

B. D. Meyer. Natural and quasi-experiments in economics. *Journal of business & economic statistics*, 13(2):151–161, 1995.

J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *arXiv preprint arXiv:1412.3773*, 2014.

J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *Journal of Machine Learning Research*, 17(32):1–102, 2016.

S. M. Oh, A. Ranganathan, J. M. Rehg, and F. Dellaert. A variational inference method for switching linear dynamic systems. Technical report, 2005.

K. Ostrowski, G. Mann, and M. Sandler. Diagnosing latency in multi-tier black-box services. In *5th Workshop on Large Scale Distributed Systems and Middleware (LADIS 2011)*, volume 3, page 14, 2011.

P. Padala, K.-Y. Hou, K. G. Shin, X. Zhu, M. Uysal, Z. Wang, S. Singhal, and A. Merchant. Automated control of multiple virtualized resources. In *Proceedings of the 4th ACM European conference on Computer systems*, pages 13–26. ACM, 2009.

J. Pearl. *Causality*. Cambridge University Press, 2000.

J. Pearl. Direct and indirect effects. In *Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 411–420, San Francisco, CA, 2001. Morgan Kaufmann.

J. Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 2009.

J. Pearl and E. Bareinboim. Transportability of causal and statistical relations: A formal approach. In *In Proceedings of the Twenty-Fifth National Conference on Artificial Intelligence. AAAI Press, Menlo Park, CA*, pages 247–254, 2011a.

J. Pearl and E. Bareinboim. Transportability of causal and statistical relations: A formal approach. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, pages 540–547. IEEE, 2011b.

E. Pereira. On solvents of matrix polynomials. *Applied Numerical Mathematics*, 47 (2):197 – 208, 2003. Second International Workshop on Numerical Linear Algebra - Numerical Methods for Partial Differential Equations and Optimization.

J. Peters, J. M. Mooij, D. Janzing, B. Schölkopf, et al. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(1):2009–2053, 2014.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. 2017. to appear at MIT press.

B. Ramachandran. *Advanced theory of characteristic functions*. Series in probability and statistics. Statistical Pub. Society, 1967.

R. C. Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.*, 37:81–91, 1945. ISSN 0008-0659.

H. Reichenbach. *The direction of time*. University of California Press, Berkeley, 1956.

J. Robins and S. Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–155, 1992.

A. Roebroeck, E. Formisano, and R. Goebel. Mapping directed influence over the brain using Granger causality and fMRI. *Neuroimage*, 25:230–242, 2005.

T. Schreiber. Measuring information transfer. *Physical Review Letters*, 85:461–464, 2000a.

T. Schreiber. Measuring information transfer. *Phys. Rev. Lett.*, 85:461–464, Jul 2000b.

W. R. Shadish, T. D. Cook, and D. T. Campbell. *Experimental and quasi-experimental designs for generalized causal inference.* Houghton, Mifflin and Company, 2002.

S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.

A. Smith and H. C. Recktenwald. *An Inquiry into the Nature and Causes of the Wealth of Nations*. Verlag Wirtschaft und Finanzen, 1986.

J. Snee, L. Carata, O. R. Chick, R. Sohan, R. M. Faragher, A. Rice, and A. Hopper. Soroban: attributing latency in virtualized environments. In *7th USENIX Workshop on Hot Topics in Cloud Computing (HotCloud 15)*, 2015.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, prediction, and search*. MIT, Cambridge, MA, 2nd edition, 2000.

K. Steele and H. O. Stefánsson. Decision theory. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition, 2016.

R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 1998.

D. Thistlewaite and D. Campbell. Regression-discontinuity analysis: an alternative to the ex-post facto experiment. *Journal of Educational Psychology*, 51:309–317, 1960.

I. Tsamardinos, S. Triantafillou, and V. Lagani. Towards integrative causal analysis of heterogeneous data sets and studies. *Journal of Machine Learning Research*, 13(Apr): 1097–1157, 2012.

J. Von Neumann and O. Morgenstern. *Theory of games and economic behavior*. Princeton university press, 2007.

P. Weirich. Causal decision theory. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition, 2016.

E. Weisstein. Polynomial discriminant, 2015. MathWorld - A Wolfram Web Resource. URL: `http://mathworld.wolfram.com/PolynomialDiscriminant.html`. Accessed: 2015-05-21.

N. Wiener. The theory of prediction. *Modern mathematics for engineers*, 1:125–139, 1956.

J. Woodward. *Making things happen: A theory of causal explanation*. Oxford University Press, 2005.

W. Zheng, R. Bianchini, G. J. Janakiraman, J. R. Santos, and Y. Turner. Justrunit: Experiment-based management of virtualized data centers. In *Proc. USENIX Annual technical conference*, 2009.