Causal Inference with Focus on Dynamical Systems and Applications in Machine Learning

Janis Kemper

Born 27th March 1994 in Paderborn

16th March 2020

Master's Thesis Mathematics

Advisor: Prof. Dr. Jochen Garcke

Second Advisor: Prof. Dr. Martin Rumpf

Institut für Numerische Simulation

Mathematisch-Naturwissenschaftliche Fakultät der

Rheinischen Friedrich-Wilhelms-Universität Bonn

Contents

Chapter 1

Introduction

Causality is a crucial concept for understanding the interactions of objects, from macroscopic climate processes down to microscopic molecule behavior. Do anthropogenic CO_2 emissions really have a causal influence on certain environmental catastrophes, such as the European heat wave in 2003 or the Australian bush fires in 2019? Does smoking really cause lung cancer, or is the correlation driven by another common influence, such as a certain gene sequence that causes both lung cancer and an inclination towards smoking? There are many problems which require causal knowledge. This information, however, cannot always be obtained by experiments.

There are not only different mathematical definitions of causality, in fact, there is even a philosophical debate about what causality is that has been going on for centuries now. Notable contributions have been made by the famous philosophers Hume and Kant; and to this day there is a scientific debate on whether the conceptions of the two philosophers differ [PF08]. Thus, it is not surprising that even today there is no common definition.

Before the era of big data, causal insights could only be obtained via controlled experiments, e.g. to find out whether a certain treatment actually has a positive effect on the disease. However, controlled experiments are not always feasible or ethical. For example, it is not possible to perform an experiment on whether the bush fires would not have happened without human CO_2 emissions.

Apart from experiments that cannot possibly be carried out, there are some which would be unethical. For example, to ascertain whether there is a causal effect from smoking to lung cancer, one would have to do a randomized controlled trial (RCT) where participants are selected randomly for a treatment and a control group. Everybody in the treatment group would be forced to smoke, while the control group is forbidden from doing so. If the lung cancer rate in the treatment group were now significantly higher than in the control group, then one would infer a causal effect from smoking to lung cancer.
These two examples are situations where the desired knowledge can be obtained in a purely data-driven manner from just observational data.

The lack of causal knowledge in science affects us in our daily lives. Just think of

the plethora of publications in recent years dispensing contradictory advice on what is healthy for us and what is not. Some claim that even tiny bits of alcohol are unhealthy, others assert that a glass of red wine a day actually helps stave off heart disease. Fillmore et al. point out how selection bias causes this discrepancy [FKS+06].

Data is the most valuable currency of the 21st century. Huge datasets are created and analyzed. Unfortunately, 'not all data are created equal' [BP16] which presents many challenges: data is collected under different experimental conditions, with a non-random sampling procedure, or with different underlying populations. Even the famous MNIST dataset is a carefully shuffled version of the original NIST dataset, which includes data collected from various writers under different conditions [ABGLP19]. A lot of contemporary research in machine learning and data science tries to ignore and get rid of this multi-domain setting, failing to realize its potential. Causality provides the language to formalize it, allowing researchers to exploit this knowledge.

Causal knowledge is especially important as it is one of the few things that still seperates artificial from human and animal intelligence. The latter learn through manipulating, transforming, and interacting with their environment [Sch19], while the former is not yet able to use the knowledge of interventions. For example, algorithms do object recognition from pixels, while children interact with the objects and are consequently much better and faster in recognizing them.

One could divide the development of causal inference into three periods. The first has been shaped by the potential outcome framework of Rubin [Rub74], as well as the use of Wiener-Granger causality in linear settings in econometrics, and stretches until the late 1990s. During this time, the research focused on formalizing the experiments and obtaining knowledge from observational data in very restrictive cases. One of the central difficulties was the phenomenon known as interference, where the common assumption that the potential outcomes of a particular person/unit is unaffected by the treatment of others does not hold. However, algorithms have been developed that can handle difficulties such as these [HH08, AS17, AEI18, BFT19, RR83a, Imb03].
In the second period, Pearl [Pea09] developed a framework that not only defines causality via interventions, but also creates a formal language that complements the traditional concept of conditional probabilities in stochastics. Thanks to this structure, causal inference began to play a more important role in data-driven sciences, because many existing problems (e.g. selection bias and multi-domain setting) were rigorously formalized. Pearl introduced the so-called causal hierarchy, a central concept that visualizes the different aspects of causality (see Figure 1.1).
In the third period, which started only a few years ago, the connection of causality and machine learning was built and exploited by Schölkopf and Peters amongst others [PJS17]. They used the idea of independence of mechanisms as the key to causality and created the base for many applications of causality in machine learning and data science [SJPZ11].
Most of the theory in causal inference treats the i.i.d. case, where the model is not time-dependent and the samples of the data are assumed to be independent and identically distributed. In this thesis, however, we present significant advancements for both random and deterministic dynamical systems. We will consider continuous-time

Level (Symbol)	Typical Activity	Typical Questions	Examples
Association $P(y\|x)$	Seeing	What is? How would seeing X change my belief in Y?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
Intervention $P(y\|do(x), z)$	Doing	What if? What if I do X?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
Counterfactuals $P(y_x\|x', y')$	Imagining, retrospecting	Why? Was it X that caused Y? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him?

Table 1.1: Pearl's hierarchy of causality [Pea19]

dynamical systems with evolutions given by ordinary or stochastic differential equations. The dynamics of the discrete-time systems are given iteratively by a prescribed function, which is much easier to handle for standard causal inference theory.
We will see that the definition of causal relationships in the continuous-time case differs significantly from Pearl's framework; there are not many results treating continuous-time dynamical systems yet. The approach of Wiener and Granger [Wie56, Gra69] for discrete-time systems has been widely adopted in its easiest form of two variables and linear relationships, but can also be extended to the non-linear and multivariate case [PJS17]. Wiener and Granger define causality as a form of information contribution of the past of one variable to the present state of another variable. This information flow can be measured with information theoretic methods like conditional mutual information.
Contrary to Pearl's framework, the notion of local independence, which is closely related to Granger causality, works for continuous-time dynamical systems [Sch70]. Similar to the other approaches, local independence defines causality as some kind of dependency of variables. However, there are others who argue that causation in the real world does not have the form of independence of variables but of mechanisms [Daw10, PJS17, ABGLP19].

We will see that causal models can be placed in between statistical and physical models, as they contain more information than statistical ones (causation instead of correlation), but less information than physical models (which usually contain information about the dynamics of systems).

Causal inference has become a very active field of research. Many achievements have only been made during the past years, such that there is no good overview of the different notions of causality that have been developed; as well as the methods and algorithms that use them. This work does not claim to be complete in that regard, but

aims to give a comprehensive introduction to the topic of causal inference. Especially for people who have no experience in the field, it is really hard not to lose focus due to the variety of different approaches that all claim to use causality. By introducing and evaluating their definitions, frameworks, and algorithms, we hope to facilitate the entry to the field of causality for researchers of many different fields.

In Chapter 2, we provide four different definitions of causality. Based on these notions we analyze algorithms that search for causal relationships in i.i.d. data and dynamical systems (Chapters 3 and 4). Causal inference methods still struggle with time series data and only in the last three years there has been significant progress in transferring the knowledge from the non-temporal to the temporal setting. Therefore, it is especially interesting to survey the progress that has been made.
Apart from searching for the causal structure of data, there are other possible tools of causal inference, such as the estimation of the strengths of causal effects we discuss in Chapter 5. Furthermore, we analyze the connection of causality and machine learning. The goal is to cluster the work that has been done to relate the two concepts into four concise groups which can be related to different aspects of causal inference (Chapter 6).
In Chapter 7, we take a step back and review the possibilities that causal inference creates for the analysis of dynamical systems as well as non-temporal models. Furthermore, we talk about the different ways in which causality can bring progress to data science and research using dynamical systems.

Chapter 2

Defining Causality

We will discuss four of the most-applied notions of causality. As mentioned in the introduction, defining causality is a difficult topic and there is not one solution that is perfect for every situation and every kind of data. In Section 2.1, we introduce Pearl's framework of causality for i.i.d. data [Pea09], which can also be applied for discrete-time dynamical systems. However, its graphical approach does not work for uncountably many random variables, so it cannot be used for continuous-time systems (treating every time step of the dynamical system as seperate random variable). Other disadvantages are its various assumptions, which are often not satisfied in applications [ARG$^+$16, RW99, Gre10, Daw10, MN19].
The second notion we discuss in Section 2.2 is based on the same invariance assumption of causal relationships that Pearl's causality uses [PBM16]. In contrast with graphical causality, it is explicitly meant to be applied in practical situations and can be used for many machine learning tasks [ABGLP19]. We name this approach *invariance-based causality*.
Wiener-Granger causality for time-series data uses different assumptions and is less elaborate than graphical causality, but also quite practical. It has been widely adopted in the case of linear relationships. We will see that the non-linear case is far more difficult for the algorithms (Section 2.3).
Wiener and Granger assume that the dynamical systems are of stochastic nature. However, many dynamical systems that are used in practice to simulate Earth system, physical, or mechanical processes are of deterministic nature [SMY$^+$12]. Therefore, another approach has been developed for deterministic dynamical systems, which we will refer to as *topological causality*, see Section 2.4.

2.1 Graphical Causality

The notion of causality which is discussed most in the current research and considered to be the most advanced framework is Pearl's graphical causality. Pearl defines causal relationships with the help of interventions [Pea09]. To illustrate this concept, let us begin with an example. Consider the room temperature and the reading of a thermometer. If we turn on the heating or air conditioning to change the temperature (to intervene on the temperature), then we can see that the thermometer shows a different value than before. If now, we change the programming of our thermometer so that it shows a specific value, for example 60 degrees Celsius, then we will observe that

the room temperature does not automatically go up to 60 degrees Celsius. In fact, it does not change at all (if we leave everything else as it is). Pearl would conclude that there is a causal relationship from the room temperature to the value shown by the thermometer, but not the other way around.

This example might seem trivial, but it shows how intuitive Pearl's concept of causality is.

Definition 2.1.1. *Let X and Y be two random variables with a joint probability distribution P. There is a causal influence from X to Y if and only if there is an intervention on X that changes the distribution of Y.*

The easiest way to represent the causal structure of a dataset is to use a directed graph. We assume that cause precedes the effect, so that there are no cyclic effects. If we want to say that X causally influences Y and not the other way around, we simply write $X \to Y$.

We will use several notations that might not be from standard graph theory, but which lead to a better visualization. A graph $\mathcal{G}$ consists of the tuple $\mathcal{G} = (V, E)$ where V is a set (of vertices or nodes) and $E \subset V \times V$ is the set of edges. We will consider mostly directed edges that are usually denoted as (X, Y) for vertices $X, Y \in V$, but we also have undirected edges $\{X, Y\}$, as well as bidirected edges that are denoted by the two edges (X, Y) and (Y, X). To visualize the three types of edges, we will instead write $X \to Y$, $X - Y$, and $X \leftrightarrow Y$.
The vertices in V are, in general, random variables. We assume that there exists a probability space on which the random variables are defined.

2.1.1 Bayesian Networks

Directed graphs have been used in statistics, especially in machine learning and AI for a long time, particularly in Bayesian networks [Pea85]. They have been used without explicitly stating (or ensuring) that the relationships in the network are causal; in fact, this does not need to hold at all. To make sure that Bayesian networks really show intuitive causal relationships, Pearl introduced *causal Bayesian networks* [Pea09].

We assume that the reader is familiar with the basic concepts of probability theory. If not, Schum gives a good introduction to the field [Sch94]. Let us start with the joint probability distribution $P(X_1, ..., X_n)$ of an ordered set of variables $\{X_1, ..., X_n\}$. We use the chain rule to write

$$P(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i | X_1, ..., X_{i-1}). \tag{2.1}$$

Assuming that X_i depends only on a subset $\mathbf{PA}_i \subset \{X_1, ..., X_{i-1}\}$, we obtain

$$P(X_1, ..., X_n) = \prod_{i} P(X_i | \mathbf{PA}_i). \tag{2.2}$$

This form considerably simplifies the joint probability distribution. $\mathbf{PA}_i$ is the set of the Markovian parents which are defined as following.

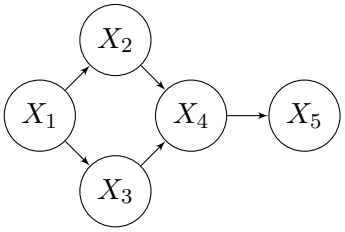


Figure 2.1: This is an example for a Bayesian network.

Definition 2.1.2. *Let $V = (X_1, ..., X_n)$ be an ordered set of variables with the joint probability distribution P. A set of variables $\mathbf{PA}_i$ is called Markovian parents of X_i if $\mathbf{PA}_i$ is a minimal set of predecessors of X_i that renders X_i independent of all its other predecessors. It holds that $P(X_i|\mathbf{PA}_i) = P(X_i|X_1, ..., X_{i-1})$.*

Note that causal parents fulfill the criteria of Markovian parents, but Markovian parents do not have to be causal. The definition depends on the variable ordering and one can find (different) Markovian parents for any ordering of the variables, whereas there is only one set of causal parents. A Bayesian network is a *directed acyclic graph* (DAG) that corresponds to a certain set of Markovian parents. It can be seen as carrier of conditional independence relations along the given ordering. For every set of Markovian parents with respect to P, we can get Equation 2.2 with the help of the chain rule. The joint probability distribution of the graph shown in Figure 2.1 can be written as

$$P(X_1)P(X_2|X_1)P(X_3|X_1)P(X_4|X_2, X_3)P(X_5|X_4).$$

Given a Bayesian network, the structure of the variables might seem interpretable, but without having any further knowledge, e.g. that the structure is temporal or causal, it is not. In fact, the directions of the edges in the graph depend solely on the order of the variables. As we can arbitrarily change this order, there is no explanation of the structure. To clarify this point, one could construct the Bayesian network that is shown in Figure 2.1 along the ordering $\{X_5, X_1, X_3, X_2, X_4\}$; it would look completely different. There is no intuitive understanding of why so different Bayesian networks should arise from the same set of variables, only by changing the order of the variables. What we need is a (unique) causal interpretation of the structure.

2.1.2 Causal Models

The reasons for making the step from normal to causal Bayesian networks are the same as for the increasing popularity of causality in statistics and machine learning: in contrast to correlations, causal relationships are explainable and unambiguous. However, causal Bayesian networks are not easy to handle in practice, as they are based on 'slippery conditional probabilities' as Pearl called them [Pea09]. The functional approach of *structural causal models* (SCMs) works much better. They have been used under the name of structural equation models (SEMs) in econometrics [Bol89] and other fields [Wri21] for decades.

Definition 2.1.3. *A structural causal model* $\mathcal{C} := (\mathbf{S}, P_N)$ *consists of a collection* $\mathbf{S}$ *of structural assignments*

$$X_j := f_j(\mathbf{PA}_j, N_j), \quad j = 1, ..., d, \tag{2.3}$$

as well as a joint probability distribution P_N *over the mutually independent random noise variables* N_j*. The functions* f_j *depend on a subset of variables, the parent sets, and a noise variable and define the value of* X_j*.*

One could also define SCMs using endogenous and exogenous variables. The X_i would correspond to the endogenous and the N_i to the exogenous variables. Note that the model is deterministic if one has given certain values for all exogenous (unmeasured) variables.

In the definition of causal Bayesian networks, we utilize Pearl's do-operator which formalizes interventions. For example, $do(X := x)$ means that we fix the variable X to the value x. In general, interventions just correspond to replacing one assignment of the SCM by something different. A proper definition will be given in Section 2.1.4.

Definition 2.1.4. *Let* P *be a probability distribution on a set* V *of variables, and let* P_x *denote the distribution resulting from the intervention* $do(X := x)$ *that sets a subset* X *of variables to constants* x*. Denote by* P_* *the set of all interventional distributions* $P_{(X,x)}, X \subset V$*, including* P*, which represents no intervention. A DAG* $\mathcal{G}$ *is said to be a causal Bayesian network compatible with* P_* *if and only if the following three conditions hold for every* $P_{(X,x)} \in P_*$*:*

1. $P_{(X,x)}$ *admits the factorization of Equation 2.2 relative to* $\mathcal{G}$ *(*P *is said to be Markov relative to* $\mathcal{G}$*);*

2. $P_{(X,x)}(v_i) = 1$ *for all* $V_i \in X$ *whenever* v_i *is consistent with* $X = x$*;*

3. $P_{(X,x)}(v_i \mid \mathbf{PA}_i) = P(v_i \mid \mathbf{PA}_i)$ *for all* $V_i \notin X$ *whenever* $\mathbf{PA}_i$ *is consistent with* $X = x$*.*

Assuming without loss of generality that $X = (V_1, ..., V_k) \subset V$, then, for $1 \leqslant i \leqslant k$, $V_i = v_i$ is consistent with $X = x$ if and only if $x_i = v_i$. In the same sense we need $\mathbf{PA}_i$ to be consistent with $X = x$ if $X \cap \mathbf{PA}_i \neq \varnothing$.
Let us explain the above. The first point is necessary for the interventional distributions to be well-defined, the second assures that the probability of a variable V_i is a point mass δ_{x_i} if we do the intervention $do(X := x)$ and $V_i \in X$. The third is a formalization of the concept of *independence of mechanisms*, which will be used several times in this thesis. The assumption is that the distributions of variables remain untouched by interventions on others.
To give an example of a causal Bayesian network, the graph in Figure 2.1 is one if and only if all edges correspond to (direct) causal relationships.

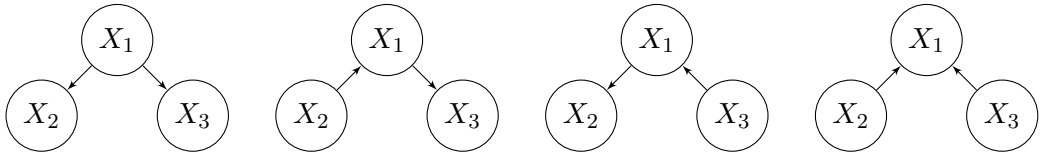


Figure 2.2: The four different possibilities of DAGs mentioned in Definition 2.1.5.

2.1.3 Connecting Graph Theory and Probability Distributions

Before continuing with causality, we need some additional theorems to connect directed acyclic graphs with probability distributions. Note that the following definitions are also valid for graphs without causal interpretation. An essential part of Pearl's theory of Bayesian networks and later causal inference is the so-called d-separation [Pea88].

Definition 2.1.5. *In a DAG $\mathcal{G} = (V, E)$, a path of nodes $(X_1, ..., X_m)$ between X_1 and X_m is blocked by a set $\mathbf{Z} \subset V$ (not containing X_1 or X_m), whenever there is a node X_k, $1 < k < m$, such that one of the following two possibilities holds:*

(i) $X_k \in \mathbf{Z}$ and either $X_{k-1} \to X_k \to X_{k+1}$, $X_{k-1} \leftarrow X_k \to X_{k+1}$, or $X_{k-1} \leftarrow X_k \leftarrow X_{k+1}$;

(ii) neither X_k nor any of its descendants is in $\mathbf{Z}$ and $X_{k-1} \to X_k \leftarrow X_{k+1}$.

Let $\mathbf{A}, \mathbf{B}$ and $\mathbf{Z}$ be three disjoint subsets of vertices. $\mathbf{A}$ and $\mathbf{B}$ are called d-separated by $\mathbf{Z}$ if every path between $\mathbf{A}$ and $\mathbf{B}$ is blocked by $\mathbf{Z}$. We write

$$\mathbf{A} \perp\!\!\!\perp_{\mathcal{G}} \mathbf{B} \mid \mathbf{Z}. \tag{2.4}$$

To build up the connection between d-separation and conditional independence, we need to assume that the distribution satisfies the Markov property and faithfulness. With these two properties we have an equivalence of the two notions [Pea09].

Definition 2.1.6. *Given a DAG $\mathcal{G}$ and a distribution P, the distribution is said to satisfy*

(i) the global Markov property with respect to $\mathcal{G}$ if for all disjoint vertex sets $\mathbf{A}, \mathbf{B}, \mathbf{Z}$

$$\mathbf{A} \perp\!\!\!\perp_{\mathcal{G}} \mathbf{B} \mid \mathbf{Z} \Rightarrow \mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{Z},$$

(ii) the local Markov property with respect to $\mathcal{G}$ if each variable is independent of non-descendants given its parents, and

(iii) the Markov factorization property with respect to $\mathcal{G}$ if the joint distribution P has a density p and

$$p(X_1, ..., X_n) = \prod_{j=1}^{n} p(X_j \mid \mathbf{PA}_j^{\mathcal{G}}). \tag{2.5}$$

It can be shown that as long as P admits a density, all three notions are equivalent [Lau96].

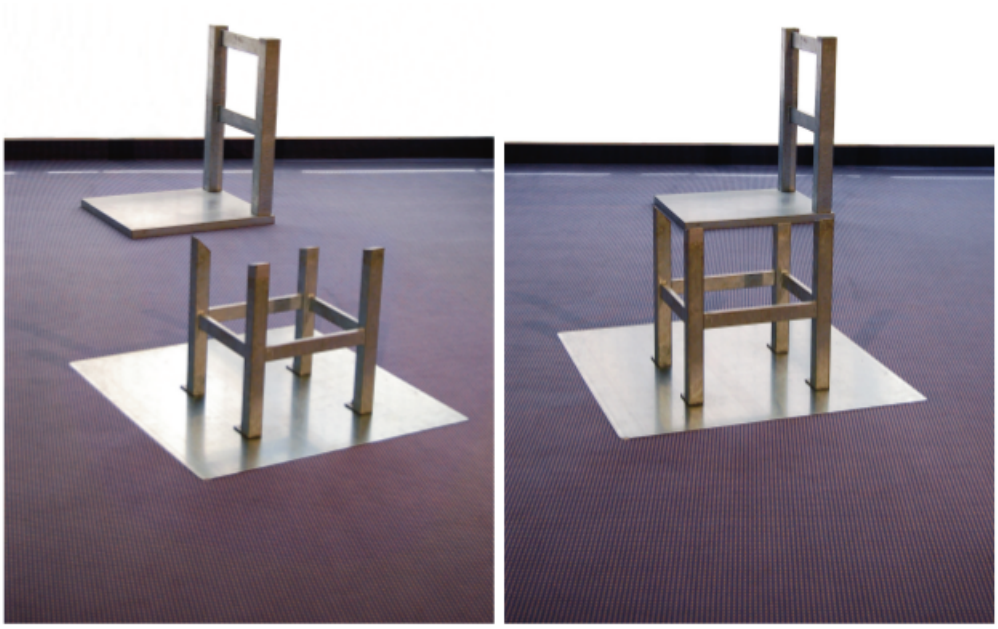


Figure 2.3: Beuchet chair. Source: Peters, Elements of Causal Inference [PJS17]

Example 2.1.7. In Figure 2.1, a distribution P_X satisfies (i) and (ii) if

$$\begin{aligned} X_1 &\perp\!\!\!\perp X_4 \mid X_2, X_3, \\ X_2 &\perp\!\!\!\perp X_3 \mid X_1, \\ X_1 &\perp\!\!\!\perp X_5 \mid X_4, \\ X_2 &\perp\!\!\!\perp X_5 \mid X_4, \\ X_3 &\perp\!\!\!\perp X_5 \mid X_4, \end{aligned}$$

and (iii) if

$$P(X_1, X_2, X_3, X_4, X_5) = P(X_1)P(X_2|X_1)P(X_3|X_1)P(X_4|X_2, X_3)P(X_5|X_4).$$

The mathematical definition of faithfulness is the following.

Definition 2.1.8. *A distribution P is called faithful to a DAG $\mathcal{G}$ if, for disjoint vertex sets $\mathbf{A}, \mathbf{B}, \mathbf{Z}$,*

$$\mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{Z} \Rightarrow \mathbf{A} \perp\!\!\!\perp_{\mathcal{G}} \mathbf{B} \mid \mathbf{Z}. \tag{2.6}$$

Note that the concept is more general in reality. Let us visualize faithfulness with an example, where it is used in its standard form, where null events, i.e. events of zero measure, are not faithful with respect to the corresponding probability distribution. The Beuchet chair (Figure 2.3) has one part that is detached from the rest and pushed a bit to the side, so that it 'flies' in the air. From almost every angle one is able to see both pieces of the chair. However, there are two very specific angles from which one might think that the chair is whole and the pieces are in reality attached to each other. These angles are 'not faithful' to reality. All other angles are faithful because we can see the true object. Mathematically speaking we would have a binary random variable X, indicating with $X = 1$ that the chair is whole and with $X = 0$ that it is

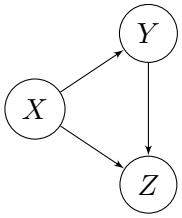


Figure 2.4: The DAG $\mathcal{G}$ of Example 2.1.9.

not. The space on which X is defined is the interval $[0, 2\pi]$ with Lebesgue measure. We have that $X(\alpha) = 0$ for almost every $\alpha \in [0, 2\pi]$, i.e. $P(X = 0) = 1$. We conclude that $X = 1$ is not faithful.

In the next example, faithfulness is used in the form that we need in this thesis.

Example 2.1.9. Let N_X, N_Y, N_Z be mutually independent, standard normal distributed variables. Let us consider the SCM

$$\begin{aligned} X &:= N_X, \\ Y &:= aX + N_Y, \\ Z &:= cX + bY + N_Z, \end{aligned} \tag{2.7}$$

which leads to the DAG shown in Figure 2.4. It is easy to see, that $X \perp\!\!\!\perp Z$ if $a \cdot b + c = 0$. Hence, in this case the distribution is not faithful to the DAG $\mathcal{G}$.

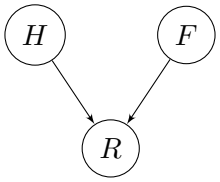
Having established the equivalence of d-separation in graphs and conditional indepenence of probability distributions [Pea09], we want to go back to Definition 2.1.5 (ii). Conditioning on an X_k that fulfills these requirements creates a dependency between two (formerly independent) variables X_1 and X_m. This effect is known under the name of Berkson's paradox, which is an example for selection bias [Ber46].

Berkson's paradox appears in reality mostly in variations of the following situation. Members of a population can have two positive characteristics, but it seems as if the traits are anticorrelated, i.e. if people have trait a, they are less likely to have b as well. The reason behind this is not that a and b are indeed dependent, but rather that the people we observe do not form a representative sample of the general population. The following example sheds some light on the question "Why are handsome men such jerks?" [Ell15, PJS17]. We will see that one reason for thinking that handsome men are less friendly (and vice versa) is that the love-seeker's dating pool consists only of men who are not in relationships.

Example 2.1.10. Assume that if a man is in a relationship ($R = 1$) is only determined by whether he is handsome ($H = 1$) and friendly ($F = 1$). Further let us assume that the SCM

$$\begin{aligned} H &:= N_H, \\ F &:= N_F, \\ R &:= \min(H, F) \oplus N_R, \end{aligned} \tag{2.8}$$

describes the setting well, where $N_H, N_F \sim Ber(0.5)$, $N_R \sim Ber(0.1)$, and $\oplus$ denotes the addition modulo two. All three noise variables are mutually independent, which implies that being handsome has nothing to do with being friendly. The corresponding DAG is the following.



Assuming that the dating pool of a woman only consists of men who are not in relationships ($R = 0$), she observes that men in her dating pool are less likely to be friendly when they are handsome and vice versa. In mathematical terms, conditioning on $R = 0$ lets the variables be anti-correlated.

Structures of the form $X \to Z \leftarrow Y$ are called *v-structures*. They do not only lead to counter-intuitive situations in statistics, but will also play a crucial role in learning causal structures. Unfortunately, we cannot find a unique DAG for every dataset, but only an equivalence class of graphs.

Definition 2.1.11. *Let $\mathcal{M}(\mathcal{G})$ denote the set of Markovian distributions with respect to $\mathcal{G}$:*

$$\mathcal{M}(\mathcal{G}) := \{P\text{: } P \text{ satisfies the global Markov property with respect to } \mathcal{G}\}$$

We say that two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent if $\mathcal{M}(\mathcal{G}_1) = \mathcal{M}(\mathcal{G}_2)$. The Markov equivalence class of a DAG $\mathcal{G}$ is the set of all DAGs that are Markov equivalent to $\mathcal{G}$.

It follows directly from the definition that two DAGs are Markov equivalent if and only if they satisfy the same set of d-separations, i.e. if they entail the same set of conditional independencies. Verma and Pearl [PV91] provide us with a characterization of Markov equivalence:

Lemma 2.1.12. *Two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent if and only if they have the same skeleton and the same v-structures.*

The skeleton of a directed graph can be obtained by taking its nodes and an undirected edge between every two nodes that are adjacent in the directed graph. The Markov equivalence class can be described by a *completed partially DAG* (CPDAG). This class of graphs additionally includes bidirected edges. A bidirected edge indicates that there exist DAGs in the Markov equivalence class that have the edge oriented in either direction.

2.1.4 Intervention Calculus

The key to finding causal relationships in Pearl's notion of causality are interventions: X causally influences Y if and only if intervening on X changes Y's distribution. To define interventions we use Pearl's do-operator.

Definition 2.1.13. *Consider an SCM $\mathcal{C} = (S, P_N)$ and its entailed distribution $P_X^{\mathcal{C}}$. There are two types of interventions on X_k that we consider in this work: stochastic and atomic interventions. Either*

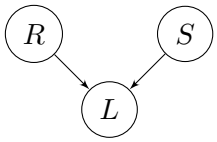
$$X_k := \tilde{N}_k,$$
$$\text{or } X_k := x,$$

where $\tilde{N}_k$ is some random variable following a certain probability distribution and $x \in \mathbb{R}$. Hence, we replace the original assignment with a new one. We call the entailed distribution of the new SCM $\tilde{\mathcal{C}}$ an intervention distribution, denoted by

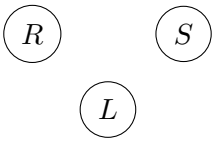
$$P_X^{\tilde{\mathcal{C}}} =: P_X^{\mathcal{C};do(X_k:=\tilde{N}_k)},$$
$$\text{or } P_X^{\tilde{\mathcal{C}}} =: P_X^{\mathcal{C};do(X_k:=x)}.$$

There are more general ways to define interventions, but this is sufficient for our needs. An intervention breaks all the links to the parent nodes as these do not appear in the structural equation any longer. Intervening on a variable makes it possible to isolate the node. Then, we can see its effect on other variables and whether there are parents that actually influence it.

Example 2.1.14. Suppose we have three binary variables where $R = 1$ means rain, $S = 1$ means the sprinkler is running, and $L = 1$ means the lawn is wet. Intervening on $L = 1$ by pouring a bucket of water over the lawn disconnects the node from the influence of its parents. It does not matter any more if it is raining or if the sprinkler is turned on, because neither can change (or influence) the status of the lawn, as it is wet anyway. The two graphs are the following.



(a) The original DAG



(b) The DAG after the intervention

There is an important difference between conditioning on a term with and without the do-operator. We continue with the example above. Let us assume that in half of the cases rain is responsible for the wet lawn. We have $P(R = 1 \mid L = 1) = 0.5$ and $P(L = 1 \mid R = 1) = 1$. If we now condition on the do-operator, the second probability does not change and $P(L = 1 \mid do(R := 1)) = 1$. Even if we actually make it rain instead of just observing it, the lawn is going to be wet. The first probability, on the other hand, is different than with normal conditioning. An intervention on L does not affect R at all, so that $P(R = 1 \mid do(L := 1)) = P(R = 1)$.

An important observation regarding interventions in SCMs is that intervening on one variable does not affect the distributions of the others. This is consistent with the definition of causal Bayesian networks, where this independence of mechanisms is even a requirement. Let $\tilde{\mathcal{C}}$ be the SCM that is constructed from $\mathcal{C}$ by intervening on some variables, but not on X_j. To facilitate the notion, write $pa(j) := \mathbf{PA}_j$. We have

$$p^{\tilde{\mathcal{C}}}(X_j \mid X_{pa(j)}) = p^{\mathcal{C}}(X_j \mid X_{pa(j)}). \tag{2.9}$$

Pearl deduced therefrom a formula that he called *truncated factorization* [Pea93]. It allows us to compute interventions without actually having interventional data. The concept of establishing whether we can identify a certain intervention with observational data is called identifiability.

Definition 2.1.15. *Let X and Y be two variables and $\mathcal{C}$ an SCM. An intervention distribution $P^{\mathcal{C};do(X:=x)}(y)$ is called identifiable if it can be computed from the observational distribution and the graph structure.*

This definition is sufficient for most of our needs. However, one can define the concept in a more general way, which has also been done by Pearl.

Definition 2.1.16. *Let M be a causal model (e.g. an SCM) and $Q(M)$ be any computable quantity. We say that Q is identifiable in a class $\mathbf{M}$ of models, if for all pairs of models $M_1, M_2 \in \mathbf{M}$*

$$P_{M_1} = P_{M_1} \Rightarrow Q(M_1) = Q(M_2). \tag{2.10}$$

Consider the structural causal model $\mathcal{C}$ and let the computable quantity $Q(\mathcal{C})$ be the intervention distribution $P^{\mathcal{C};do(X:=x)}(y)$. Let $\mathbf{M}$ be the class of models that induce the same causal graph as $\mathcal{C}$ and positive distributions on the observed variables. For $M_1, M_2 \in \mathbf{M}$ it holds that $P^{M_1;do(X:=x)}(y) = P^{M_2;do(X:=x)}(y)$. Hence, $Q(\mathcal{C})$ can be computed uniquely and the intervention distribution is identifiable.

Let us go back to the truncated factorization formula. Consider an SCM $\mathcal{C}$ with density p and assignments

$$X_j := f_j(X_{pa(j)}, N_j), \qquad j = 1, ..., d.$$

Let $\tilde{\mathcal{C}}$ be the SCM that evolves from $\mathcal{C}$ after the intervention $do(X_k := \tilde{N}_k)$. Let $\tilde{p}$ be its density. The truncated factorization formula follows from the Markov assumption and the property that an intervention on one of the variables does not affect the others.

$$\begin{aligned} p^{\mathcal{C};do(X_k:=\tilde{N}_k)}(x_1, ..., x_d) &= \prod_{j \neq k} p^{\mathcal{C};do(X_k:=\tilde{N}_k)}(x_j \mid x_{pa(j)}) \cdot p^{\mathcal{C};do(X_k:=\tilde{N}_k)}(x_k) \\ &= \prod_{j \neq k} p^{\mathcal{C}}(x_j \mid x_{pa(j)}) \tilde{p}(x_k). \end{aligned} \tag{2.11}$$

For atomic interventions, we have

$$p^{\mathcal{C};do(X_k:=a)}(x_1, ..., x_d) = \begin{cases} \prod_{j \neq k} p^{\mathcal{C}}(x_j \mid x_{pa(j)}) & \text{if } x_k = a, \\ 0 & \text{otherwise.} \end{cases} \tag{2.12}$$

The Equations (2.11) and (2.12) give us a method to calculate interventions without actually having interventional data. However, they are not easy to use and there is a more practical alternative: the do-calculus [Pea09]. It consists of three rules that are complete in a sense that all identifiable interventions can be computed with them [SP06, HV06]. Here, the term computation refers to writing out the interventional distribution in terms of standard probability theory.

Theorem 2.1.17. *Let $\mathcal{G}$ be a graph and let $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$, and $\mathbf{W}$ be disjoint subsets of vertices. The three rules of do-calculus are the following:*

(i) Consider a graph in which incoming edges in $\mathbf{X}$ are removed. If $\mathbf{Y}$ and $\mathbf{Z}$ are d-separated by $\mathbf{X}$ and $\mathbf{W}$, then

$$p^{\mathcal{C};do(\mathbf{X}:=\boldsymbol{x})}(\boldsymbol{y} \mid \boldsymbol{z}, \boldsymbol{w}) = p^{\mathcal{C};do(\mathbf{X}:=\boldsymbol{x})}(\boldsymbol{y} \mid \boldsymbol{z}).$$

(ii) Consider a graph in which incoming edges in $\mathbf{X}$ and outgoing edges from $\mathbf{Z}$ are removed. If $\mathbf{Y}$ and $\mathbf{Z}$ are d-separated by $\mathbf{X}$ and $\mathbf{W}$, then

$$p^{\mathcal{C};do(\mathbf{X}:=\boldsymbol{x},\mathbf{Z}:=\boldsymbol{z})}(\boldsymbol{y} \mid \boldsymbol{w}) = p^{\mathcal{C};do(\mathbf{X}:=\boldsymbol{x})}(\boldsymbol{y} \mid \boldsymbol{z}, \boldsymbol{w}).$$

(iii) Consider a graph in which incoming edges in $\mathbf{X}$ and in $\mathbf{Z}(\mathbf{W})$ have been removed. $\mathbf{Z}(\mathbf{W}) \subset \mathbf{Z}$ is the subset of nodes that are not ancestors of any node in $\mathbf{W}$ in a graph that is obtained from $\mathcal{G}$ after removing all edges into $\mathbf{X}$. If $\mathbf{Y}$ and $\mathbf{Z}$ are d-separated by $\mathbf{X}$ and $\mathbf{W}$, then

$$p^{\mathcal{C};do(\mathbf{X}:=\boldsymbol{x},\mathbf{Z}:=\boldsymbol{z})}(\boldsymbol{y} \mid \boldsymbol{w}) = p^{\mathcal{C};do(\mathbf{X}:=\boldsymbol{x})}(\boldsymbol{y} \mid \boldsymbol{w}).$$

With the help of these three rules, we can prove two important formulas: the back-door and the front-door adjustment.

Definition 2.1.18. *Let $\mathbf{X}, \mathbf{Y}$ and $\mathbf{Z}$ be three disjoint subsets of vertices. $\mathbf{Z}$ is said to satisfy the*

(a) back-door criterion relative to $(\mathbf{X}, \mathbf{Y})$ if

- *(i) no node in $\mathbf{Z}$ is a descendant of any node in $\mathbf{X}$; and*
- *(ii) $\mathbf{Z}$ blocks all the paths between nodes in $\mathbf{X}$ and $\mathbf{Y}$ that contain an edge into a node in $\mathbf{X}$.*

(b) the front-door criterion relative to $(\mathbf{X}, \mathbf{Y})$ if

- *(i) $\mathbf{Z}$ intercepts all directed paths from nodes in $\mathbf{X}$ to nodes in $\mathbf{Y}$;*
- *(ii) there is no back-door path from nodes in $\mathbf{X}$ to nodes in $\mathbf{Z}$; and*
- *(iii) all back-door paths from nodes in $\mathbf{Z}$ to nodes in $\mathbf{Y}$ are blocked by nodes in $\mathbf{X}$.*

We define a back-door path as a path from node X to node Y, which is not the edge (X, Y) itself and which ends with a directed edge into Y.

Theorem 2.1.19. *(a) If a set of variables* $\mathbf{Z}$ *satisfies the back-door criterion relative to* $(\mathbf{X}, \mathbf{Y})$*, then the causal effect of* $\mathbf{X}$ *on* $\mathbf{Y}$ *is identifiable and is given by the formula*

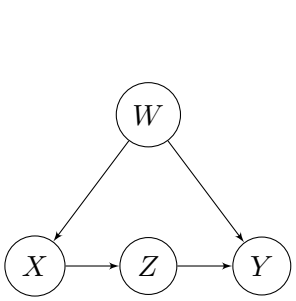
$$P(\mathbf{Y} \mid do(\mathbf{X} := x)) = \sum_{\boldsymbol{z}} P(\mathbf{Y} \mid x, \boldsymbol{z}) P(\boldsymbol{z}). \tag{2.13}$$

(b) If a set of variables $\mathbf{Z}$ *satisfies the front-door criterion relative to* $(\mathbf{X}, \mathbf{Y})$ *and if* $P(\mathbf{X}, \mathbf{Z}) > 0$*, then the causal effect of* $\mathbf{X}$ *on* $\mathbf{Y}$ *is identifiable and is given by the formula*

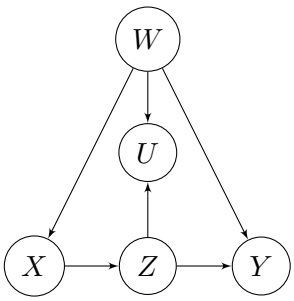
$$P(\mathbf{Y} \mid \boldsymbol{x}) = \sum_{\boldsymbol{z}} P(\boldsymbol{z} \mid \boldsymbol{x}) \sum_{\boldsymbol{x}'} P(\mathbf{Y} \mid \boldsymbol{x}', \boldsymbol{z}) P(\boldsymbol{x}'). \tag{2.14}$$

The back-door adjustment can be proved using the more general second rule in Theorem 2.1.17. First write $P(\mathbf{Y} \mid do(\mathbf{X})) = \sum_{\mathbf{z}} P(\mathbf{Y} \mid do(\mathbf{X}), \mathbf{z}) P(\mathbf{z})$. In order to apply the second rule of Pearl, we redefine $\mathbf{Z} := \mathbf{X}$ and $\mathbf{W} := \mathbf{Z}$. The proof of the front-door adjustment is not much harder and can be found in [JV18].

Example 2.1.20. Consider the pair (X, Y) in the following two graphs. We want to calculate $P(Y \mid do(X))$.



(a) DAG $\mathcal{G}_1$



(b) DAG $\mathcal{G}_1$

There are actually two possibilities: we can use the back-door and the front-door adjustment. The node W satisfies the back-door criterion in both $\mathcal{G}_1$ and $\mathcal{G}_2$ as it blocks all back-door paths from X to Y. The front-door criterion is slightly more complicated, but it is fairly easy to see that in $\mathcal{G}_1$, it is satisfied by Z. Z obviously intercepts all directed paths from X to Y, there are no back-door paths from X to Z, and the back-door path from Z to Y is blocked by X.

Trying to use the front-door criterion in $\mathcal{G}_2$ shows us that Pearl's conditions in Definition 2.1.18 (b) are too restrictive. There are some back-door paths from Z to Y which do not have to be blocked by nodes in X because they are blocked anyway by colliders. This is actually the case in $\mathcal{G}_2$. Therefore, even though it does not, strictly speaking, fulfill the conditions, Z is still enough to apply the front-door adjustment.

We will now talk about algorithms that learn causal structures. Unfortunately, the algorithms are only able to learn Markov equivalence classes and they thus cannot distinguish two graphs of the same Markov equivalence class. Given a probability distribution P which is Markovian and faithful with respect to a DAG $\mathcal{G}$, the Markov equivalence class of $\mathcal{G}$ is identifiable, while $\mathcal{G}$ itself is not. Only additional domain knowledge might help with finding the true underlying graph.

The good news is that there are several types of models for which it is easier to find identifiability statements for individual graphs. However, these model classes are quite restrictive, such as additive noise models (ANMs) of the form $X_j := f_j(PA_j) + N_j$ with nonlinear f_j and linear Gaussian models with equal error variance [PMJS14, PJS17].

2.1.5 Critique

Confounding variables are one of the main difficulties in causal inference applications. Cole et al. give an example of a medical study where confounders lead to false conclusions of whether a treatment effect is mediated or not [CH02]. However, the problem with confounders goes beyond the practical problem of finding causal relationships in models with (hidden) confounding variables. In fact, there is a quite philosophical debate on whether it makes sense to make causal conclusions in a way that Pearl proposes, knowing that there are quite likely millions of (possibly very small) potential confounders in real-world applications, which can never be included in the analysis [RW99].
Apart from that, there is a lively ongoing discussion between followers of Rubin's potential outcome framework and Pearl's graphical causality on whether Pearl's approach is useful or not. All arguments are wrapped up in the recent paper of Imbens [Imb19]. He also argues in favor of the potential outcome framework, as it has been widely adopted in fields like econometrics. According to him, graphical causality still needs to find its way into the mainstream of statistics and data science.
Maclaren et al. criticize the definition of identifiability of Pearl which is used as 'can be estimated from data' [MN19]. Instead, they define identifiability in a more general way and analyze it with methods of algebra and category theory. Essentially, identifiability is equivalent to the injectivity of a certain function (see Definition 2.1.16) and can be seen as an inverse problem. They argue that the problem is ill-posed because it does not fulfill the conditions of Hadamard [Had02] that are commonly thought of as the basis for well-posed inverse problems in statistics. Maclaren et al. claim that what they call estimability is equivalent to the continuity of the inverse of the function, so that it is a different concept. They conclude that identifiability is widely misused in practice. Note that in this work we also used identifiability in the sense of Pearl.
Greenland argues that more realism is needed when thinking about the possibilities of causal inference [Gre10]. He stresses the amount of hypotheses needed to do inference and to find causal structures. There will always be many confounders, measurement errors, and few actual independences. Dawid criticizes Pearl's assumptions as too strong [Daw10]. He concentrates on the connection of conditional independence and what he calls probabilistic causality and argues that it is unlikely to be fully true. Furthermore, he says that invariance of mechanisms, a central assumption in graphical causality, will not hold across all regimes. Instead, he proposes another notion of causality. It is not as elaborate, but uses fewer assumptions.

2.2 Invariance-Based Causality

We will follow the approach that was first developed by Schölkopf et al. [SJPZ11] and extended by Peters et al. [PBM16] for the linear case as well as by Heinze-Deml et al. [HDMP18] for non-linear models. Another line of work has been started by Zhang et

al. [ZHZ+17] who further generalize the approach of Peters et al. We will see whether or not it is able to give new insights to causal inference.

Let $X = (X_1, .., X_p)$ be the multivariate predictor, Y the target variable, E the environmental variables, and let us assume that we are given an SCM over (Y, X, E). Environmental variables are the variables that are neither descendants nor parents of Y in the causal graph and are allowed to be non-random. Note that conditional independence relations can be generalized to non-random variables. Let $S^* \subset \{1, ..., p\}$ be the indices of X that correspond to the causal parents $\mathbf{PA}_Y$ of Y. We can write the assignment of the SCM as

$$Y := f(X_{S^*}) + \epsilon. \tag{2.15}$$

The definition of environmental variables and the Markov property imply that

$$Y \perp\!\!\!\perp E \mid X_{S^*}. \tag{2.16}$$

Our goal is to find the set S^*, which we will achieve by exploiting the above relation. Assuming to be able to test for the null hypothesis

$$H_{0,S} : \qquad Y \perp\!\!\!\perp E \mid X_S, \tag{2.17}$$

for all sets $S \subset \{1, ..., p\}$, Peters et al. [PBM16] propose to define an estimate $\hat{S}$ for the parental set S^* by setting

$$\hat{S} := \bigcap_{S: H_{0,S} \text{ not rejected}} S, \tag{2.18}$$

where the intersection runs over all sets S with $X_S \cap E = \varnothing$. By definition $\hat{S} \subset S^*$. In the linear case, the hypothesis can be tested via linear regression. In the non-linear case, it can be tested with non-linear conditional independence tests.

A similar approach to test for invariance across environments has been proposed recently by Arjovsky et al. [ABGLP19]. They consider a very similar setup and assume to have training environments $e \in \mathcal{E}_{tr}$ and related but unseen environments $\mathcal{E}_{all} \supset \mathcal{E}_{tr}$. The idea is to minimize

$$R^{OOD}(f) = \max_{e \in \mathcal{E}_{all}} R^e(f), \tag{2.19}$$

where $R^E(f) := \mathbb{E}_{X^E, Y^E}[l(f(X^E, Y^E)]$ is the risk under environment E. Formally, this is done with the help of a constrained optimization problem which can be solved using gradient descent techniques. The notion of causality that is used here differs slightly from the invariance-based causality of Peters et al., as the goal is not to get information about causal structures in data that can be represented in form of DAGs. Instead, it tries to find invariant mechanisms which do not have to have the form of a causal relation from X to Y.

An easy example of Arjovsky et al. is an image classification problem, where the goal is to distinguish cows from camels. Of course, most pictures of cows are taken in green pastures and the pictures of camels are taken in the desert. Hence, there is a selection

bias in the data. Assume that we have data from different environments, say different countries, where the percentages of pictures of cows with green background differ. Then, the proposed method realizes that the mechanism 'green pasture means cow' is not invariant and thus cannot be generalized. Instead, the output is a conditional distribution which is able to generalize across the environments, e.g. some features of the animals' bodies that differentiate them. This task cannot be solved with a graphical approach, as pixels do not qualify as vertices of a causal graphical model. One pixel does not causally relate to the image label, at least not in a robust model.

In her dissertation, Li [Li18] considered the case of invariance-based causality for time series. She developed a method, MINT-T, to estimate causal effects. Unfortunately, there is no implementation of this algorithm yet, and it remains open whether it can actually enter the mainstream of time series analyses. Thus, we are only able to rate algorithms for the i.i.d. case that use the invariance-based approach.

2.3 Wiener-Granger Causality

Wiener-Granger causality is a notion of causality for dynamical systems. We will only be able to treat the case of discrete-time dynamical systems extensively. The continuous-time case is much harder to treat and has not been adopted for practical purposes yet. We will give a continuous-time version of the definition of Granger causality without discussing it further. Dynamical systems can be divided into two types: stochastic and deterministic. In this section, we will treat the stochastic case. Discrete-time dynamical systems that are random can be seen as stochastic processes, which justifies the usage of terms like stationarity from probability theory.

Even if it sounds promising, we cannot just transfer all of Pearl's theory from the i.i.d. case to time series. Even though the time structure gives us the causal ordering, there are other difficulties that occur. For example, we usually have only one repetition of the time series. This is fundamentally different from the i.i.d. case, where we usually have many samples of one variable. Therefore, it is necessary to assume stationarity of the dynamical system.
Another problem might be the sampling rate, which can be too low to capture the true causal relationships. Especially when it is reasonable to assume that the underlying process has contiuous time, the sampling rate causes trouble [ARG+16].

An additional nuance is that there are different ways to define causality for dynamical systems. An overview, which also includes some less known approaches of defining causality for dynamical systems, can be found in [Eic12]. We focus mainly on Granger causality, which is applied most in practice as it has an empirical counterpart. Information theory is predominantly used to test for Granger causality, as we will see later.

2.3.1 Preliminaries

Let $(\mathbf{X}_t)_{t\in\mathbb{N}}$, $\mathbf{X}_t = (X_t^1, ..., X_t^d)$ be a d-variate time series. Assume that the X_t^i are stationary stochastic processes. The graph $\mathcal{G} := (V \times \mathbb{Z}, E)$, where $V = \{X^1, ..., X^d\}$

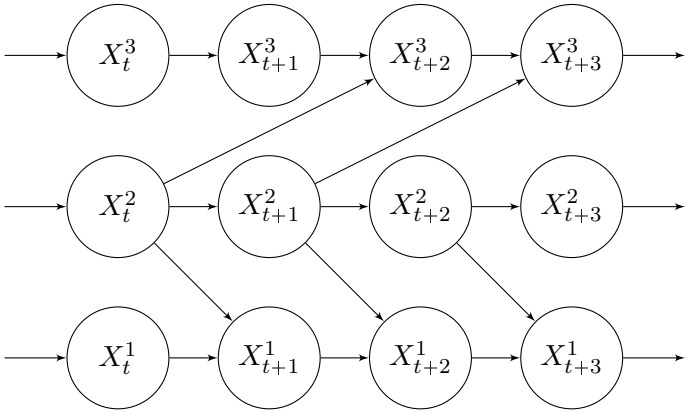


Figure 2.7: This is an example for a full time graph of a time series.

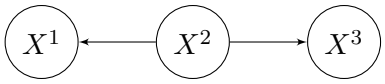


Figure 2.8: This is the summary graph of the full time graph in Figure 2.7.

and $E \subset (V \times \mathbb{Z}) \times (V \times \mathbb{Z})$ is the set of edges, is called *full time graph* [PJS17]. It consists of the nodes X_t^j for $(j,t) \in \{1,...,d\} \times \mathbb{Z}$ and of edges that cannot go backward in time. If there exists an edge from X_t^j to X_t^k for some $t \in \mathbb{N}$ and $1 \leqslant k, j \leqslant d$, then we say that there are instantaneous effects. Not all algorithms can cope with instantaneous effects, and most of the ones that can, are not able to orient these edges. The reason is that they usually test for undirected measures, so that they have to exploit the time structure to infer (directed) causation.

Following Schölkopf [PJS17], we also define the *summary graph*. It is the directed graph $\mathcal{G} := (V, E)$ where $V = \{X^1, ..., X^d\}$ is the set of nodes. It holds that $e = (X^j, X^k) \in E$ if and only if there exists an edge from X_t^j to X_s^k for some $t \leqslant s \in \mathbb{Z}$. Note that if there is a causal relation from X_t^j to X_s^k, then stationarity implies that there is also one from $X_{t+t'}^j$ to $X_{s+t'}^k$ for all $t' \in \mathbb{Z}$, see Figure 2.7. We assume that the full time graph is acyclic, but the summary graph may contain cycles.

2.3.2 Granger and Other Notions of Causality

Eichler [Eic13] used interventions to define causal effects, just as Pearl did in the i.i.d. case. Let $\sigma = \{\sigma_t^j, t \in \tau \subset \mathbb{N}, 1 \leqslant j \leqslant d\}$ be a set of indicators denoting interventions in $\mathbf{X}_t$ at time points $t \in \tau$. If $\sigma = \varnothing$, then no intervention is performed and we call the corresponding probability distribution $P := P_{\sigma=\varnothing}$ the observational regime. Just like in the i.i.d. case, we consider atomic interventions where $\sigma_t^j = x_0$, and stochastic interventions where $\sigma_t^j = p$ for some probability distribution p. For atomic interventions, we obtain

$$P_{\sigma_t^j = x_0}(X_t^j = x \mid \mathbf{X}_{t-1}) = \delta_{x_0}(x). \tag{2.20}$$

For stochastic interventions the distribution $P_{\sigma_t^j = p}(X_t^j \mid \mathbf{X}_{t-1})$ is the same as p. Note that both types of interventions break all the links to the parent nodes. There are

some independence assumptions to assure that the distribution of other nodes is not affected by the intervention on one node, see [Eic13] for details. Similar to the static case, we say that there is a causal effect from node X_t^j to X_s^k if the distribution of X_s^k is not the same under the observational regime P as it is under at least one interventional regime $P_{\sigma_s^k=p}$ or $P_{\sigma_s^k=x_0}$ for some p or x_0 respectively.

The second possibility to define causal effects is to use structural equations of the form

$$X_t^i = f_i(X_1^1, X_1^2, ..., X_{t-1}^{d-1}, X_{t-1}^d), \tag{2.21}$$

where X_t is a function of the whole past of the stochastic process. We assume that discrete-time dynamical processes are given by these structural equations, while continuous-time dynamical processes are defined via ODEs. According to White and Lu there is a causal effect from X_t^j to X_s^k for $t \leqslant s$ if and only if the function of X_s^k is constant in X_t^j [WL10]. Alternatively, one could also define causal effects using interventions on the structural equations, similar to Pearl's graphical causality in the i.i.d. case.

Granger [Gra69] and Sims causality [Sim72] are two probabilistic approaches to define causality that originate from econometrics. In contrast to Granger causality, Sims causality takes into account not only direct but also indirect causal relations. However, they are quite similar and Granger is the one usually used in practice [Eic13]. As mentionend in [HSPVB07], the inspiration that the Nobel prize winner Clive W.J. Granger needed for his work [Gra69] came from Norbert Wiener [Wie56]. Based on his assertion that 'for two simultaneously measured signals, if we can predict the first signal better by using the past information from the second one than by using the information without it, then we call the second signal causal to the first one', Granger identified two fundamental principles for his definition of causality:

- The cause occurs before the effect; and
- the cause contains unique information about the effect that is not available otherwise.

The first principle is commonly accepted, the second one is more delicate. Let $(X_t, Y_t, \mathbf{Z}_t)_{t\in\mathbb{N}}$ be a d-dimensional time series where we assume that $\mathbf{Z}$ contains all observed variables apart from X and Y. Granger defines two information sets:

- $\mathcal{I}(t)$ is the set of all information in the universe up to time t;
- $\mathcal{I}_{-X}(t)$ is the set of all information up to time t except for the process X.

If X causes Y, we expect the conditional distributions $P(Y_{t+1} \mid \mathcal{I}(t))$ and $P(Y_{t+1} \mid \mathcal{I}_{-X}(t))$ to differ from each other. Granger used this observation to define his notion of causality.

Definition 2.3.1. *The series X does not cause series Y if for all $t \in \mathbb{N}$*

$$Y_{t+1} \perp\!\!\!\perp \mathcal{I}(t) \mid \mathcal{I}_{-X}(t); \tag{2.22}$$

otherwise, X is said to cause Y.

The definition states that there is no causal influence from X to Y if the past of X does not provide additional information for predicting Y. As the set $\mathcal{I}(t)$ is of a very abstract nature and there are measure-theoretic subtleties, we cannot use this definition in practice. For example, it is unclear whether $\mathcal{I}_{-X}(t)$ contains truly less information than $\mathcal{I}(t)$, since this would imply that we can discretize the universe in time and space. To avoid this problem (to which we will return in Section 2.4), we substitute the information sets with the σ-algebras $\sigma(\{X_{\leqslant t}, Y_{\leqslant t}, \mathbf{Z}_{\leqslant t}\})$ and $\sigma(\{Y_{\leqslant t}, \mathbf{Z}_{\leqslant t}\})$ generated by the observed stochastic processes, where $X_{\leqslant t} = (X_s, s \leqslant t)$ denotes the past of variable X up to time t. We obtain the following modified version of Granger's definition.

Definition 2.3.2. *Let $(X, Y, \mathbf{Z})$. The process X is Granger-noncausal for the process Y with respect to $(X, Y, \mathbf{Z})$ if*

$$Y_{t+1} \perp\!\!\!\perp X_{\leqslant t} \mid Y_{\leqslant t}, \mathbf{Z}_{\leqslant t}. \tag{2.23}$$

Otherwise we say that X Granger-causes Y.

Using the same independence assumptions as before, we can connect the concept of Granger causality to interventions [ED10].

Corollary 2.3.3. *Consider a multivariate time series $(X, Y, \mathbf{Z})$ and an intervention $\sigma_t^x = s$ on the process X satisfying some independence assumptions. If X is Granger non-causal for Y, then there is no causal effect on Y_{t+1} of intervening in X_t.*

Granger's concept of causality is non-parametric. However, he applied it himself only on the class of linear models, as this is the easiest one to treat. In fact, many researchers in fields like econometrics only know Granger causality for this reduced set of models. For example, they use linear regression to test for Granger causality and compare

$$\begin{aligned} Y_n &= \sum_{i=1}^{k} a_i Y_{n-i} + N_n, \\ \text{and} \quad Y_n &= \sum_{i=1}^{k} a_i Y_{n-i} + \sum_{i=1}^{k} b_i X_{n-i} + \tilde{N}_n, \end{aligned}$$

where $(N_i)_{i\in\mathbb{N}}$ and $(\tilde{N}_i)_{i\in\mathbb{N}}$ are assumed to be i.i.d. time series. X Granger-causes Y if the noise terms $\tilde{N}_i$ have significantly smaller variance than N_i. There are several other test statistics for the linear case [HSPVB07], but to use the concept for non-parametric model classes, one needs different methods to test for Granger causality.

2.3.3 Testing for Granger Causality - Information Theory

Information theoretical methods are mostly used as a non-parametric counterpart for linear Granger causality. Let X and Y be two absolutely continuous random variables with the joint distribution $p_{(X,Y)}$, marginal densities p_X and p_Y, and the conditional distribution function $p_{Y|X}$ which is defined for all x. Note that if $p_X > 0$, we can write $p_{Y|X}(y|x) = \frac{p_{(X,Y)}(x,y)}{p_X(x)}$. All definitions will be given for the case of absolutely

continuous variables. The discrete case follows by integration with respect to the counting measure. We set

$$H(X) := -\int p_X(x)\log(x)dx \tag{2.24}$$

to be the Shannon entropy of X. The conditional entropy Y given X can be defined as

$$H(Y \mid X) := -\int p_{(X,Y)}(x,y)\log p_{(Y|X)}(y|x)d(x,y). \tag{2.25}$$

Define the mutual information $I(X;Y)$ of X and Y via

$$\begin{aligned} I(X;Y) &:= H(X) - H(Y \mid X) \\ &= D_{KL}(p_{(X,Y)} \parallel p_X p_Y) \end{aligned} \tag{2.26}$$

where $D_{KL}(P \parallel Q) = \int p(x)\log(\frac{p(x)}{q(x)})dx$ for $p << q$ denotes the Kullback-Leibler divergence. We assume here that $0 \cdot \log(0) = 0$. Although the Kullback-Leibler divergence is inherently asymmetric, the mutual information is not. Let Z be another random variable and $p_{(X,Y)|Z}$ the conditional joint probability density. We can define the conditional mutual information

$$\begin{aligned} I(X;Y \mid Z) &:= H(X \mid Z) - H(X \mid Y,Z) \\ &= \int D_{KL}(P_{(X,Y)|Z} \parallel P_{X|Z} \otimes P_{Y|Z})dP_Z \\ &= \int \left(\int \log\left(\frac{p_{(X,Y)|Z}(x,y|z)}{p_{X|Z}(x|z)p_{Y|Z}(y|z)} \right) p_{(X,Y)|Z}(x,y|z)d(x,y) \right) p_Z(z)dz. \end{aligned} \tag{2.27}$$

It is easy to see that $I(X;Y|Z) = 0$ if and only if $X \perp\!\!\!\perp Y \mid Z$. Therefore, we can use it as a measure to test for Granger causality [CT06]. As the CMI is symmetric, it does not help us with finding any directionality. Thus, the information about the time structure of our stochastic processes is absolutely vital to find causal relations.

We therefore go back to time series and assume to have given a multivariate time series $(X, Y, \mathbf{Z})$. If we find that the conditional mutual information of, say X_{t-1} and Y_t, is high, we can infer that there is a causal influence from the one variable to the other. Note that additional precautions have to be taken in order to avoid false conclusions in data with confounding variables, i.e. where there is another variable, say Z_{t-2}, which causes both X_{t-1} and Y_t. We will see later how the algorithms handle the multivariate case where this can be a problem.

Conditional mutual information is not the only measure that can be used to find causal structures. Next to be discussed is the so-called transfer entropy [Sch00], which follows the two principles of Granger and hence is closely related to his definition of causality. It is based on mutual information, but is additionally able to measure dynamical and directional information. Instead of using static probabilities, transfer entropy uses transition probabilities. Therefore, it is supposed to be better in quantifying the information flow from one variable to another.
Assume that we have two discrete-time stochastic processes X and Y with joint probability distribution P that are absolutely continuous, such that P admits a density p

and all conditional densities exist. Further, assume that X is of order τ_X and Y of order τ_Y, i.e. the processes depend only on the τ_X or τ_Y preceding time steps. We define

$$\begin{aligned} T_{X\to Y} &:= H(Y_{n+1} \mid Y_n^{(\tau_Y)}) - H(Y_{n+1} \mid X_n^{(\tau_X)}, Y_n^{(\tau_Y)}) \\ &= \int\int\int p(y_{n+1}, y_n^{(\tau_Y)}, x_n^{(\tau_X)}) \log\left(\frac{p(y_{n+1}|y_n^{(\tau_Y)}, x_n^{(\tau_X)})}{p(y_{n+1}|x_n^{(\tau_X)})}\right) dx_{n+1} dx_n^{(\tau_X)} dy_n^{(\tau_Y)}, \end{aligned} \tag{2.28}$$

where $X_n^{(\tau_X)} = (X_n, X_{n-1}, ..., X_{n-\tau_X+1})$. Note that transfer entropy does nothing else than calculating the mutual information of Y_{n+1} and $X_n^{(\tau_X)}$ given $Y_n^{(\tau_Y)}$. In other words, we measure the mutual information of the present state of Y and the past of X given the past of Y. Since information flow cannot go back in time, it is obvious that, if the mutual information is greater than zero, there is an information flow from X to Y. In practice, the definition is often used with $\tau_X = \tau_Y = 1$, which makes it considerably easier to handle.

Transfer entropy works perfectly in the bivariate case, but as soon as we are in the multivariate case, things get more difficult. If we have a causal chain $X \to Z \to Y$ for stochastic processes X, Y and Z, then transfer entropy does not consider the intermediate variable Z, so that we cannot distinguish between direct and indirect effects. Causation entropy [SB14] gives us the solution to this problem with generalizing transfer entropy:

$$C_{X\to Y|(X,Z)} := H(Y_{n+1} \mid Y_n, Z_n) - H(Y_{n+1} \mid X_n, Y_n, Z_n) \tag{2.29}$$

With this measure, one is able to not only condition on the past of Y, but also on the past of as many other measured processes as needed. In the case of the causal chain described above, we would have $C_{X\to Y|(X,Z)} = 0$, but $T_{X\to Y} > 0$. Transfer entropy, on the other hand, does not help us to find out whether the information flow from X to Y is direct, or whether all the information flows through the intermediate variable Z.

2.3.4 Limitations of Granger Causality

It is important to know the situations where it makes sense to apply Granger causality and where it does not. Peters et al. [PJS17] give examples where Granger causality fails to detect the right causal relationships. However, the examples are quite specific, e.g. purely deterministic relations. Janzig et al. [JBGWS13] argue that information theoretical measures, such as transfer entropy, sometimes fail to quantify the amount of information flow correctly. They show that the measures are able to detect without quantifying them correctly. One reason is described in the following.
Assume that we want to calculate the information flow from time series X to time series Y. It has been shown that transfer entropy adds up the information that comes from the past of X with the information that comes from X and Y together, pretending that all this information comes solely from X [JBC16]. In this way, the real flow of information from X to Y can be overestimated, because some of the information only comes from both time series combined, e.g. when $Y_t = \min(X_{t-1}, Y_{t-1})$. If there are

joint influences of two variables on a third, transfer entropy can also underestimate the information flow of each variable to the target.

A problem emphasized by Granger himself is that his notion of causality cannot handle hidden variables. In Definition 2.3.1, we assume to have all information in the universe; in Definition 2.3.2, all information about the other variables. If we do not have the necessary information, then Granger causality cannot be applied.
Of course, there are other aspects that can be criticized. Friston et al. [FBO+14] point out several of them in their evaluation of Granger causality and show its limitations on biological time series data.

2.4 Topological Causality

Topological causality is an answer to some of the problematic assumptions that Granger causality is built on, such as the separability of the space. Deterministic dynamical systems are inherently non-separable as a direct consequence of Takens' theorem [SBDH97]. For other dynamical systems that describe real-world physical, mechanical, or Earth system processes it is not known whether seperability holds [HLSP17, RBB+19, SMY+12].

There is no notion of causality yet that does not assume separability and can be used for any kind of dynamical system. Using Takens' theorem, however, other methods that are based on delay embeddings have been developed for (non-linear) deterministic systems. We call this approach topological causality. As it is not related to Pearl's graphical approach, a full discussion would be outside the scope of this thesis. We will give a brief summary and present an algorithm that uses the ideas of topological causality, but we will not go deeper into the topological details.

According to Stark et al. [SBDH97], Takens' theorem can be informally described as follows. Let ϕ be a scalar observable of a state x of a deterministic dynamical system. Then, we can typically reconstruct a copy of the original system by considering blocks $(\phi(x_t), \phi(x_{t+\tau}), \phi(x_{t+2\tau}), ..., , \phi(x_{t+(m-1)\tau})$ of m successive observations of ϕ, for m sufficiently large and a sampling interval $\tau > 0$. Note that in practice, x_t is unknown while $\phi(x_t)$ is measured.
What follows from this theorem is that we might be able to use a delay embedding of one of the variables, say X^0, to reconstruct the system's dynamics entirely. In other words, all the information of the dynamical system might be stored in just one variable, which makes a separation of the information impossible.
With topological causality one can create an asymmetric measure of dependency for systems having the following setup. Assume that there are two system parts X^1 and X^2 and that the dynamics are governed by

$$\begin{aligned} \dot{X^1} &= f_1(X^1, w_{12}\mu(X^2)), \\ \dot{X^2} &= f_2(X^2, w_{21}\mu(X^1)), \end{aligned} \tag{2.30}$$

where $\mu_i(X^i)$ denote fixed scalar functions and w_{ij} coupling constants, for $i = 1, 2$ and $j = 3 - i$. We assume that the trajectories (X_t^1, X_t^2) form an invariant manifold in the

phase space of the joint dynamical system. The two manifolds of the state space of the delay embedding $r_t^i = (\phi(X_t^i), ..., \phi(X_{t+(m-1)\tau}^i))$ are both topologically equivalent to the manifold of the trajectories as long as $w_{ij} \neq 0$ and $w_{ji} \neq 0$, respectively. By transitivity, a unique mapping $M_{i \to j}$ from r^i to r^j exists if and only if $w_{ij} \neq 0$ [CGS15].

2.5 Difficulties with Defining Causality for Dynamical Systems

There are several difficulties that occur when defining causality for dynamical systems, but even more when applying these definitions in practice. We will point out a few that may have been shortly mentioned already and others that have not been talked about yet.
The first problem encountered by many researchers is that the given data does not fulfill the assumptions that are usually made in causal inference. Let $(X_t^i)_{i \in I, t \in T}$ be the repetitions $i \in I$ of stochastic processes that are measured in time points $t \in T$. A sample (or a repetition) of a stochastic process has the form of a time series. Assuming we have only a small number of repetitions, as is usually the case in practice, we cannot compute conditional independence or conditional mutual information, since this would require many samples. Thus, there is only one possibility to get enough data points: we have to assume that the time series is stationary, or at least that the effects from X_t to $Y_{t+\tau}$ (e.g. in form of a functional relationship) do not change for different t. Then, we can obtain the required data points not by taking different samples X_t^i for a fixed t, but rather by sampling over time. We use the ordered sets $\{X_t \mid t \in T, t \geqslant \tau\}$ and $\{Y_{t+\tau} \mid t \in T\}$ to compute conditional independence of processes X and Y with a time lag of τ. All of the algorithms that work with conditional independence testing use this idea.
This approach makes sense if the stochastic process consists in reality of i.i.d. random variables, so that there is no time structure. If the stochastic process is non-stationary, however, so that the distribution of the process in one time point t and in another time point t' differs fundamentally, then the data in the two sets will be very chaotic and no algorithm will be able to find the true causal structure. To avoid this issue, we have to assume stationarity, conceding that this will greatly reduce the number of possible applications.

The second problem that is considered in research is called subsampling. If we sample from a discrete-time stochastic process, then it is not clear whether the sampling rate is small enough to capture the causal relationships. One can think of many easy examples where the resulting causal graph can be completely wrong if, say, every second time step is not measured. There has been research on how to handle subsampling [HPJ+16, GZS+15], but no general solution has been presented yet.

The biggest difficulty of the algorithms and the theory of causal inference in general is that most of the time, it is assumed to have a discrete-time process, even though the world works in continuous time. For example, Aalen et al. criticize the ignorance of this problem and argue that the use of directed acyclic graphs does not make sense in a real-world setting [ARG+16]. They argue for a different notion of causality that works for continuous-time systems: the concept of local independence [Sch70]. It has

been developed by Schweder during the same time where Granger has been working on his approach, and the two concepts are closely related. Local independence requires stochastic independence of two variables on an infinitesimal level, i.e. a condition imposed on the generator of the stochastic process.
Unfortunately, there has not been much research following the concept of local independence, even Aalen et al. merely criticize existing research. Didelez is one of the few who extended the work of Schweder, applying it to multivariate marked point processes [Did08]. However, her work is rather theoretic, so that there is no algorithm yet which uses local independence to find causal relationships. One other publication should be mentioned here, as it uses local independence to generate causal knowledge, even though it is not about finding causal relationships, but to get counterfactual knowledge: Roysland used local independence graphs to identify consistent estimators for counterfactual parameters, connecting tools of stochastic analysis and causal inference [Roy12].

We have only defined Granger causality for discrete-time processes, but the definition can be extended to continuous time. The following is not supposed to be a precise definition and we will not think about the measure theoretic details, but it gives an intuition of how this can be done. Let $(\mathcal{F}(X)_t)_{t\in\mathbb{R}_{\geqslant 0}}$ be the filtration of the process X which contains all its information and let $\mathcal{F}(X)_{t^-}$ be the σ-algebra that contains all information of X before t. The difference between $\mathcal{F}(X)_{t^-}$ and $\mathcal{F}(X)_t$ is that the latter contains the information of X_t, while the former does not. One could say that X is Granger non-causal to Y if

$$Y_t \perp\!\!\!\perp \mathcal{F}(X)_{t^-} \mid \mathcal{F}(Y)_{t^-}, \mathcal{F}(Z)_{t^-}; \tag{2.31}$$

i.e. if the past of process X contributes information to Y_t, given all information about the past of Y and Z.
Publications that apply Granger causality to continuous-time dynamical systems treat rather special cases. Barnett et al. [BS17], for example, analyzed neuro-physiological problems in continuous time with Granger causality using analytic solutions of stochastic models.

Chapter 3

Finding Causal Relationships - IID Case

Two of the notions of causality from Chapter 2 were developed for i.i.d. data: Pearl's graphical causality and invariance-based causality of Peters, Schölkopf, and others. The latter approach is quite new and there are not many algorithms yet. The one that we analyze here is called nonlinearICP (nonlinear invariant causal prediction). On the other hand, many algorithms have been developed which use graphical causality. They can be divided into two groups: the constraint-based and the score-based algorithms.

In this chapter, we start with the algorithms based on Pearl's causality (Sections 3.1 and 3.2), then discuss some difficulties that usually appear in practice (Section 3.3), and finish with analyzing the methods in an empirical way using different experiments (Section 3.4). As a contrast to the other algorithms, nonlinearICP is discussed at the end of this chapter in Section 3.5.

3.1 Constraint-Based Algorithms

The most basic algorithm to find causal structures is Pearl's inductive causation (IC) algorithm [Pea09]. It belongs to the class of the *constraint-based methods*. They assume that the distribution is Markovian and faithful to the underlying DAG, so that the Markov equivalence class is identifiable. Constraint-based algorithms search for d-separation statements which can be tested via conditional independence tests. The following lemma simplifies the search process [PV91]:

Lemma 3.1.1. *Let X, Y be two vertices in a DAG $\mathcal{G} = (V, E)$. The following two statements hold.*

(i) X and Y are adjacent if and only if they cannot be d-separated by any disjoint subset $\mathbf{W} \subset V$.

(ii) If X and Y are not adjacent, then they are d-separated by either $\mathbf{PA}_X$ or $\mathbf{PA}_Y$.

Let us, for now, assume that we have an oracle giving us the right answers to all questions concerning conditional independence. This will allow us to focus on the

algorithms themselves. Later, we will talk about the difficulties to test for conditional independence.

The first step of the IC and SGS (Spirtes, Glymour, and Scheines) [SGS00] algorithm is to find the skeleton of the underlying DAG. To that end, they use Lemma 3.1.1 (i) and search through all possible subsets of nodes $\mathbf{W} \subset V$ to check whether nodes X and Y are d-separated given $\mathbf{W}$. The nodes are adjacent if and only if no such set can be found. The PC algorithm [SGS00], named after Peter Spirtes and Clark Glymour, works like the IC algorithm but has a more efficient way of searching for the conditioning set using Lemma 3.1.1 (ii). The PC algorithm starts with an empty set and increases the size of the set in each iteration. It makes use of the fact that it is sufficient to iterate over the subsets of neighbors of X or Y.

The second step consists of orienting the edges. Assume that the skeleton contains the structure $X - Z - Y$ where X and Y are not adjacent, such that there is a set $\mathbf{W}$ that d-separates X and Y. Looking again at Figure 2.2 and all possible orientations of the undirected structure, we observe that in all but one case Z has to be in $\mathbf{W}$ and $\mathbf{W}$ automatically blocks the path between X and Y. The one case where Z is not in $\mathbf{W}$, the three nodes form a v-structure and Z is is called a *collider*. As conditioning on a collider renders the two independent nodes dependent, the collider cannot be in the set $\mathbf{W}$. Thus, $Z \notin \mathbf{W}$ implies that the orientation of the above structure has to be $X \rightarrow Z \leftarrow Y$. If $Z \subset \mathbf{W}$, we cannot orient the edges, as there are several possible orientations we cannot distinguish.

Using this observation we can start orienting the edges of the skeleton. Having oriented all edges using v-structures in the data, we might be able to additionally orient some edges, e.g. to avoid directed cycles. In fact, there is a set of orientation rules proven to be complete, known as Meek's orientation rules [Mee95]. The aforementioned algorithms use Meek's rules to find the Markov equivalence class.

To summarize, the IC algorithm has three steps. First, we compute the skeleton using various conditional independence tests, second, we search for v-structures in all triples $X - Z - Y$, and third, we use Meek's orientation rules to further orient edges of the graph. The outcome of the algorithm is the Markov equivalence class of the true underlying graph. The PC algorithm applies the same steps, but has a more efficient method to compute the skeleton with fewer conditional independence tests.

Latent variables pose a particularly common and difficult problem for causal inference algorithms. If there are hidden variables in a dataset, we want to be able to distinguish between an association of two variables that is caused by a latent confounder and one that is due to a direct causal relationship. The Fast Causal Inference (FCI) algorithm [SGS00] has a third type of edge in its output to mark confounded variables. For a triple $X \leftarrow U \rightarrow Y$, where U is a latent confounder, FCI draws a bidirected edge $X \leftrightarrow Y$ indicating that there is no direct causal influence from X to Y or the other way around, but another hidden variable that influences both of them. Thus, there are three different possibilities for every undirected edge: it can be oriented in one of the two directions, or it can be bidirected.

The PC and FCI algorithms use different types of graphs. Recall that the PC algorithm uses DAGs and outputs a CPDAG (defined in Section 2.1.3), describing the Markov equivalence class of the true underlying structure. The FCI algorithm, on the other hand, requires different types of edges to indicate hidden variables and uses the so-called *maximal ancestral graphs* (MAGs), which include bidirected edges. The Markov equivalence classes are described by *partial ancestral graphs* (PAGs). PAGs can contain edges with circles on one or both ends. A circle means that there has to be at least one graph in the Markov equivalence class where the edgemark is an arrowhead and one graph where it is a tail.

We assumed above to have an oracle answering all questions of the form "Are nodes X and Y conditionally independent given $\mathbf{W}$?". These questions, however, are not at all easy to answer in most cases. Statistical significance tests have to be used to find out how likely an independence is, but finite sample sizes and statistical errors may lead to wrong and contradictory results. There are non-parametric, kernel-based tests like the Kernel Conditional Independence Test (KCIT) [ZPJS11] and approximations of kernel-based tests such as the Randomized Conditional Independence Test (RCIT) and Randomized Conditional Correlation Test (RCoT) [SVZ19, Str20]. However, additional domain knowledge makes the results less prone to error and it makes sense to restrict ourselves to a subclass of possible causal models, e.g. for (joint) Gaussian distributions. For these subclasses, easier and more reliable statistical testing is possible; e.g. for Gaussian distributions it is sufficient to test for vanishing partial correlations.

3.2 Score-Based Algorithms

The second class of algorithms consists of *score-based methods*. The idea is to test the ability of different graph structures to fit the data. Given data $\mathcal{D}$ containing i.i.d. samples of a set of variables V, the space $\mathcal{B}$ of possible DAGs $\mathcal{G} = (V, E)$ for some edge set E, and the scoring function $S(\mathcal{D}, \mathcal{G})$, we search for the DAG with the highest score.

$$\hat{\mathcal{G}} := \underset{\mathcal{G} \in \mathcal{B}}{\operatorname{argmax}}\ S(\mathcal{D}, \mathcal{G}) \tag{3.1}$$

There are many different possibilities for chosing the scoring function, the space of possible DAGs, as well as the method for searching the space for the DAG with the highest score. As the number of DAGs scales exponentially in the number of vertices, one has to find an efficient way of searching through the space [HGC94, HMC99]. A known score-based algorithm is the Greedy Equivalence Search (GES) [Chi02]. However, the score function of GES only works for Gaussian data.

Recently, Huang et al. [HZL+18] proposed a method for generalized score functions. It is based on defining suitable scores for a particular regression problem in the Reproducing Kernel Hilbert Space (RKHS). The framework can be used for nonlinear causal relationships as well as linear ones and for both continuous and discrete data.

3.3 Hidden Variables and Other Difficulties

There are many improvements of the aforementioned constraint-based and score-based algorithms. Some are faster in practice or theory, some are more accurate, and some only work for a restricted class of functions. The basic algorithms such as the PC algorithm have exponential run time in the worst-case, but work reasonably fast for smaller graphs in practice.

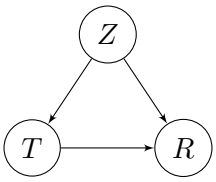
Learning causal structures is a very difficult task due to the many potential pitfalls: conditional independence tests can give false results, or the method for searching the space of possible DAGs might not be efficient for the true underlying DAG. Perhaps the biggest challenge for causal inference is to not observe all variables. If a variable is hidden, a true causal relation can actually be reversed. This is what happens in the famous Simpson's paradox, which we will look at in the following example.

Example 3.3.1. The dataset from kidney stone recovery [CWPW86] shows the success of two possible treatments of kidney stones a and b. There are two categories of patients: group A with small and group B with large kidney stones. In the following table, we see that treatment b has the higher overall success rate, even though treatment a works better for both categories of patients.

	Overall	Group A	Group B
Treatment a	78% (273/350)	**93%** (81/87)	**73%** (192/263)
Treatment b	**83%** (289/350)	87% (234/270)	69% (55/80)

This is the so-called Simpson's paradox. The two groups of patients have very different sizes and both treatments perform better on one group than on the other. Treatment b seems to be, overall, the better one, even though it is neither for group A, nor for group B. Here, the importance of observing all variables is evident. Without the information about the two patient groups, it would not be possible to see that treatment b is worse, in fact.

We can wrap this experiment up in the language of causal inference. Let $R = 1$ be the event of full recovery, T the treatment and Z the patient group. The true underlying DAG is the following.



The computations in [PJS17] show that

$$\begin{aligned} &P^{\mathcal{C}}(R = 1 \mid T = a) < P^{\mathcal{C}}(R = 1 \mid T = b), \text{ but} \\ &P^{\mathcal{C};do(T:=a)}(R = 1) > P^{\mathcal{C};do(T:=b)}(R = 1), \end{aligned} \tag{3.2}$$

as both

$$P^{\mathcal{C};do(T:=a)}(R=1 \mid Z=A) > P^{\mathcal{C};do(T:=b)}(R=1 \mid Z=A), \text{ and}$$
$$P^{\mathcal{C};do(T:=a)}(R=1 \mid Z=B) > P^{\mathcal{C};do(T:=b)}(R=1 \mid Z=B). \tag{3.3}$$

Let us for the purpose of demonstration ignore that there are different patient groups and assume that recovery solely depends on the treatment (and not on the size of the kidney stones), i.e. we assume $T \to R$ is the correct graph. Denote the causal model that can be built on this false assumption by $\tilde{\mathcal{C}}$. We can rewrite Equation 3.2 as

$$P^{\tilde{\mathcal{C}};do(T:=a)}(R=1) < P^{\tilde{\mathcal{C}};do(T:=b)}(R=1), \text{ and}$$
$$P^{\mathcal{C};do(T:=a)}(R=1) > P^{\mathcal{C};do(T:=b)}(R=1), \tag{3.4}$$

using the fact that in a model of the form $X \to Y$ conditioning and intervening on X have the same effect on the distribution of Y. We see, that the causal statement gets reversed because of model misspecification. We ignored the relevance of patient groups for the recovery and came to a wrong conclusion.

When using causal inference algorithms, Simpson's paradox has to be kept in mind, as the algorithms would also draw false conclusions if not all relevant variables are measured. However, there are algorithms that can handle situations with hidden (unobserved) variables, e.g. the FCI algorithm [SGS00]. It can not only decide whether an influence between two variables is directed in some way, but also whether it stems from a hidden confounding variable.
Assume that we have $X \leftarrow U \to Y$ and U is not observed. Then, U is called a hidden variable and works as a confounder for X and Y. The FCI algorithm is able to discover that the association between X and Y does not come from a causal relationship but from a hidden confounder. The PC algorithm, on the other hand, cannot handle hidden variables. As it is not able to find a set that renders X and Y conditionally independent, it will connect them in the graph and conclude that there is a causal relationship.

Constraint-based algorithms using a similar approach as the IC algorithm tend to amplify mistakes they made, such that a single falsely oriented edge causes multiple mistakes in the output. The orientation procedure with Meek's orientation rules iteratively uses the edges that have been oriented before, so one mistake may cause several other edges to be oriented in the wrong direction.

Finding a balance between taking decisions and averting making wrong ones is very delicate. The data scientist's knowledge of the data continues to play an important role in discovering wrong edges and orientations and most algorithms allow the use of some prior knowledge, e.g. that a certain relation needs to appear in every possible output.

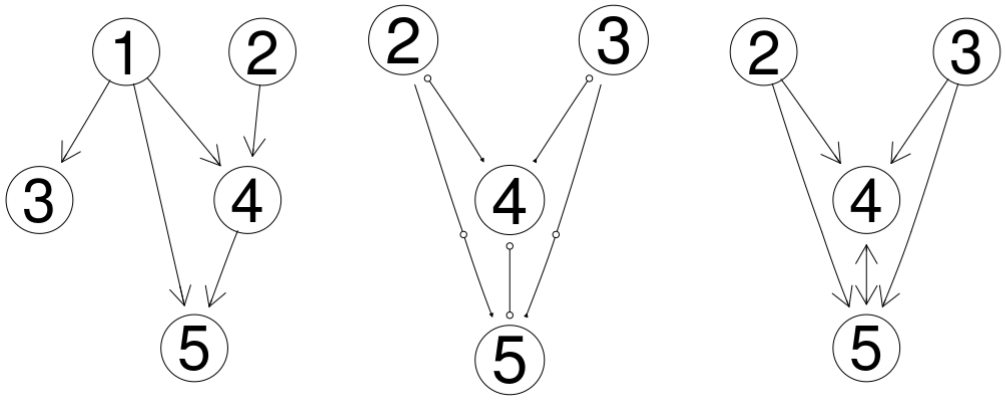
3.4 Comparison of the Algorithms

In the following, we will show some results of experiments that should clarify the opportunities and limitations of Pearl's causal inference in practice. We use different datasets generated randomly with the R package *pcalg* [KHM+20]. To generate the datasets, one only has to define the structure of the underlying DAG, the number of samples, as well as the distribution of the data. Given these three parameters, the function generates the dataset following the causal structure given by the respective graph. The different graphs that we use are shown in the respective figures. The datasets have the number of samples $T = 5000$. The data follows either a multivariate Gaussian (with or without latent variables), Cauchy, or t-distribution. As conditional independence tests, we use Fisher's z-transformation [Fis15] for Gaussian data and RCIT or RCoT [SVZ19] for non-parametric datasets.
All algorithms are implemented in pcalg. The package includes not only the original versions of the PC, FCI and GES algorithms, but also a few variations and advancements. Recall that the PC is based on the IC algorithm, the most basic constraint-based algorithm of Pearl. FCI is also constraint-based but works differently and can handle latent variables, while GES is a score-based algorithm.
Even though it is not the best choice for outputting the graphs, pcalg has different meanings for bidirected edges in the output of the PC and the output of the FCI algorithm. Another difficulty with interpreting the output is that it can happen that edges remain undirected through the whole algorithm. In the output of the PC and the GES algorithms, undirected edges are shown as bidirected, in the output of the FCI algorithm this is not the case.

Variations of the PC algorithm include conservative PC [RSZ06] and PC with majority rule [CM14], which both try to avoid making mistakes during the process of orienting edges. After the skeleton is computed, they take every triple $A - B - C$ in the undirected graph and check all subsets of the adjacent nodes to see whether A and C are independent given the subset. If B is in some sets that render A and C independent but not in others, it is a contradiction to the theory we have seen above. B should be either in none of these sets (so that the triple forms a v-structure) or in all of them. Conservative PC marks all triples with contradicting test results as ambiguous. As mistakes happen easily since we rely on statistical testing, this rule is very strict.
Majority rule handles ambiguous edges in a different, less strict manner, by simply following the opinion of the majority. For example, if B is in less than half of all subsets of neighbors of A and C that render them independent, the majority rule concludes that the triple is a v-structure. There are similar versions of FCI (conservative and with majority rule), as well as FCI+ [CMH13] and RFCI [CMKR12], which use different techniques to make the original algorithm much faster.

3.4.1 Data with Latent Confounders

Hidden variables make the task of orienting the edges harder because of two different reasons. The first one lies at hand: there is one more possibility to orient an edge. The second reason is that the algorithm is more likely to make mistakes with the skeleton when the correlations in the data are less obvious. A wrongly detected edge in the skeleton can lead to subsequent mistakes. As between the basic algorithms,



(a) True DAG (b) Estimated PAG by FCI (c) Estimated CPDAG by GES

Figure 3.1: A model with five variables and multivariate Gaussian data where the first variable has been deleted, so that it is hidden for FCI and GES.

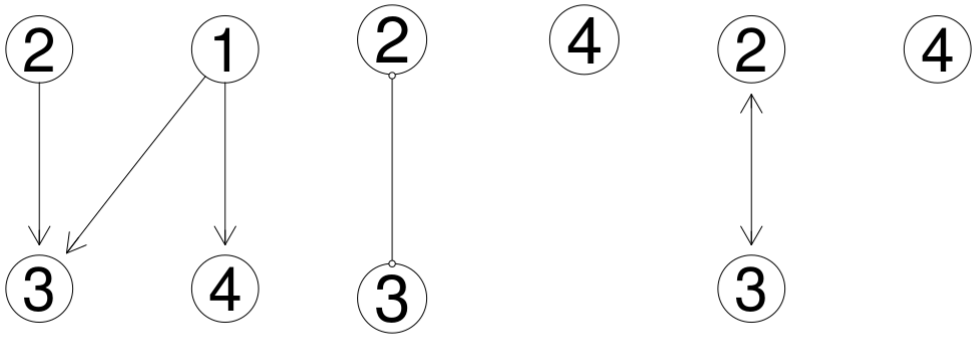
FCI is the only one that can handle latent variables in theory, we are going to use it in the following experiments. As a comparison, we try the GES algorithm, although it is not meant to be used on data like this, as it assumes that all variables are measured.

Figure 3.1 shows the result of one of the experiments, where the data of the true DAG has been manipulated in such a way, that the algorithms did not get the data of the first node. Hence, there is a latent confounder for nodes 3 and 4, for nodes 4 and 5 and nodes 3 and 5. These nodes should be connected in the output of FCI with a bidirected edge. We can see that they are indeed connected, but not with bidirected edges. Apparently, there are some MAGs with directed edges so that the edge of the PAG has circles. Not even the skeleton is correct, as there is an edge from node 2 to node 5. The output of GES has the same skeleton as the one of FCI algorithm and the latter is only slightly better.

The second set of experiments we carried out confirms the doubts about the reliability of the FCI algorithm, see Figure 3.2. The overall structure can be mostly trusted, but not the individual edgemarks. GES algorithm appears to perform reasonably well if the goal is simply to find a skeleton.

3.4.2 Data with Different Distributions

To find out whether different distributions have an impact on the results of the conditional independence tests, we will evaluate the tests on Gaussian, Cauchy, and t-distributed data. For Gaussian data there exists a parametric test using Fisher's z-transformation, for the other two distribution only non-parametric tests can be applied. Note that all datasets in this chapter have 5000 samples.



(a) True DAG (b) Estimated PAG by FCI (c) Estimated CPDAG by GES

Figure 3.2: A model with four variables and multivariate Gaussian data where the first variable has been deleted, so that it is hidden for FCI and GES.

Experiments with Gaussian Distribution

In Figure 3.3, one can see that GES works better than FCI with Fisher's z-transformation on Gaussian data, as it orients some of the edges which are not oriented by FCI algorithm. The skeleton is correct in both cases. In Figure 3.4, we used FCI with the non-parametric tests RCIT and RCoT. The skeletons are both correct and the edge-marks differ only slightly. The experiments shown in Figures A.2 - A.5 indicate as well that it is preferable to use Fisher's z-transformation instead of non-parametric tests for the FCI algorithm.

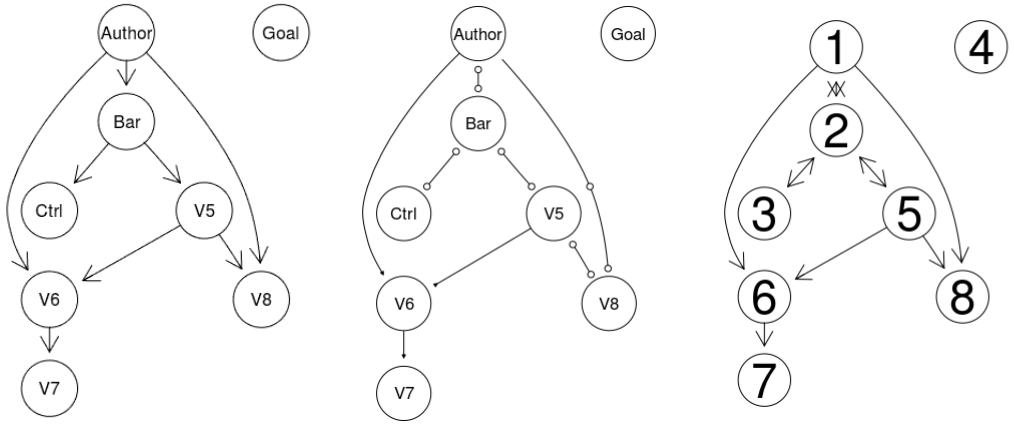
Experiments with Cauchy Distribution

Figure 3.5 shows that FCI with the conditional independence test RCoT infers the right skeleton, while FCI with RCIT makes two mistakes. Also, the other two experiments, see Figures A.6 and A.7, indicate that RCoT works (slightly) better than RCIT, both for the task of finding the skeleton and for orienting the edges.

Experiments with t-Distribution

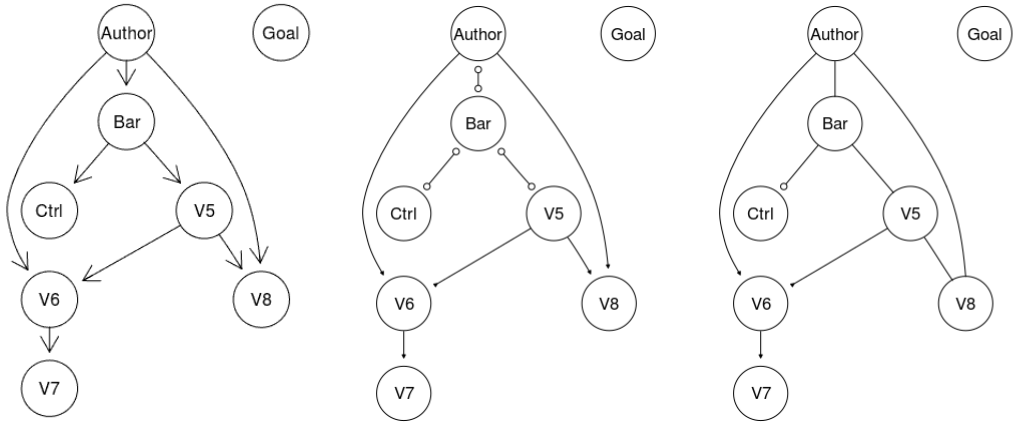
The t-distribution apparently causes more difficulties than the Gaussian or Cauchy distribution, as can be seen in Figure 3.6. FCI was not able to infer the right skeleton, neither with RCIT nor with RCoT. However, Figures A.8 and fig:exdiffdistrt3 show that with other data the skeletons can be estimated correctly.

To summarize, the algorithms are mostly able to infer the true skeleton of the causal causal. Depending on the distribution of the data, there are more mistakes (t-distribution) or less (Gaussian distribution). Both RCIT and RCoT work well lead to similar results. However, the algorithms only orient few edges, so that one can differentiate between direct and indirect relations, but the directions are mostly unknown.



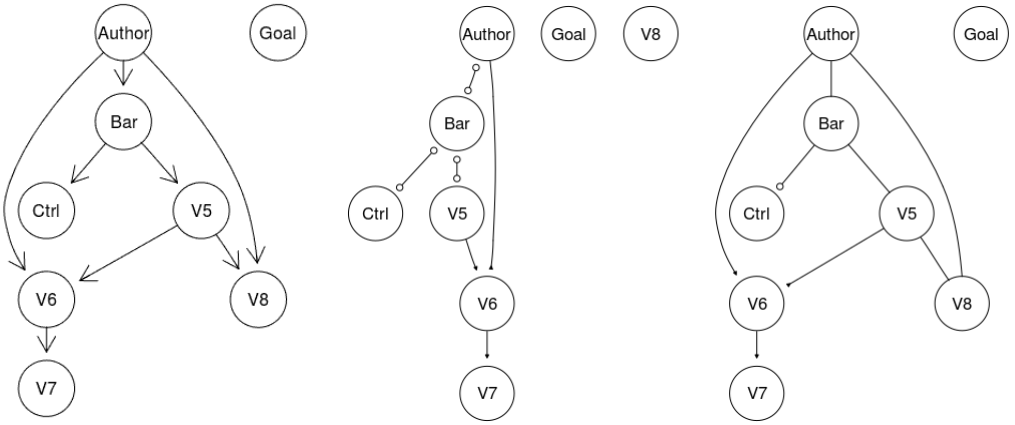
(a) True DAG (b) Estimated PAG by FCI (c) Estimated CPDAG by GES

Figure 3.3: Multivariate Gaussian data, the FCI algorithm uses Fisher's z-transformation as conditional independence test. Note that bidirected edges (e.g. between the nodes 1 and 2) in the CPDAG can be undirected in reality.



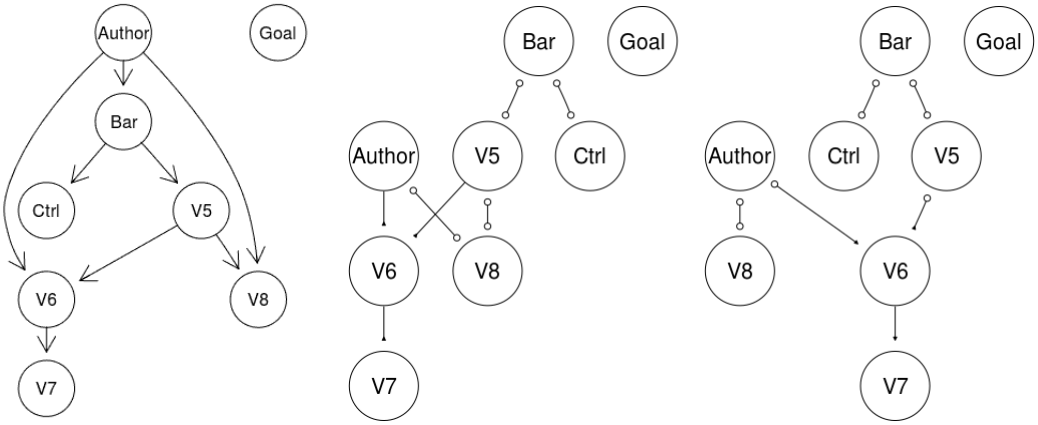
(a) True DAG (b) Estimated PAG using RCIT (c) Estimated PAG using RCoT

Figure 3.4: Multivariate Gaussian data. FCI uses the conditional independence tests RCIT and RCoT.



(a) True DAG (b) Estimated PAG using RCIT (c) Estimated PAG using RCoT

Figure 3.5: Cauchy distributed data. FCI outputs different skeletons using RCIT and RCoT.



(a) True DAG (b) Estimated PAG using RCIT (c) Estimated PAG using RCoT

Figure 3.6: t(df=4)-distributed data. FCI with RCIT and RCoT both made mistakes with the skeleton.

3.4.3 Complexity Analysis

The complexity of these algorithms is quite high. For independence-based algorithms, it is given by the complexity of conditional independence test as well as the numbers of tests that have to be performed. For score-based methods, it depends on the size of the space of possible graphs and the method used to search the space. In this thesis, we will only do an exemplary complexity analysis for the PC algorithm.

The first driver of complexity is the number of conditional independence tests that have to be carried out in the worst case. It depends on the graph structure, namely the number of nodes and adjacent edges. The second step is to consider the complexity of the conditional independence test itself, which depends on the number of samples. If n is the number of nodes and k the largest degree of a node (the degree is the number of adjacent nodes), then the PC algorithm uses at most $2\binom{n}{2}\sum_{i=0}^{k}\binom{n-1}{i}$ conditional independence tests, which is bounded by $\frac{n^2(n-1)^{k-1}}{(k-1)!}$. In practice, we usually work with sparse graphs, whose degree k is bounded by a constant, and therefore achieve a polynomial run time bounded by $\mathcal{O}(n^{k+1}A)$, where A is the complexity of the conditional independence test. FCI algorithm, although faster in practice, also needs exponentially (with respect to the number of nodes) many independence tests in the worst case. RFCI can be used to get a better run time with polynomially many conditional independence tests.
The complexity of the conditional independence tests varies a lot. Fisher's z-transformation is used to test for the partial correlations in multivariate Gaussian data. It only needs basic mathematical operations such as summation, division, as well as calculating the logarithm, and thus scales linearly in sample size. The chi-squared test is a known method to test for conditional independence in the discrete setting and has a linear run time as well. There are also different methods to test for independence in the case of linear models.
General models, on the other hand, are naturally much harder to treat. One approach could be to discretize the space, but this suffers strongly from the curse of dimensionality, as we need small bins to get good results [Hua10]. Reproducing kernel-based methods are known to work well in high-dimensional settings, but they usually scale at least quadratically in sample size, since the computation of the kernel matrix itself scales quadratically. KCIT even has cubic scaling [ZPJS11]. To find a balance between accuracy and functionality, RCIT and RCoT try to approximate kernel-based tests using random Fourier features [SVZ19]. The tests scale roughly linear in sample size, but are much slower in practice than a chi-squared test, for example.

Empirical Results

As the usability of algorithms not only depends on their performance, but also on their run time, we will analyze the latter for the causal inference algorithms considered. We perform the experiments for one constraint-based algorithm (FCI) and one score-based algorithm (GES). We used the same three different conditional independence tests for FCI as before: the non-parametric tests RCIT and RCoT, and Fisher's z-transform for Gaussian data. Because of the difficulty of developing a general score function, the implementation of GES we use works only for Gaussian data. The run time of both algorithms depends on the sample size of the data and on the number of variables.

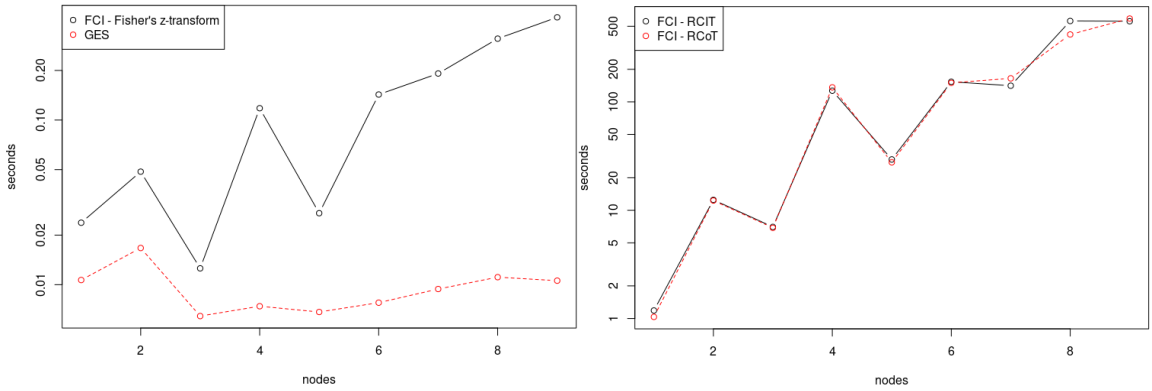


Figure 3.7: The run time of GES algorithm and FCI with respect to the number of nodes. Both graphs use a log-scale.

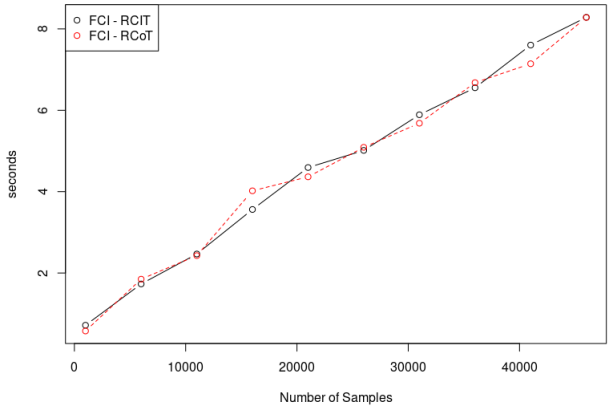


Figure 3.8: The run time of FCI with respect to the sample size.

To compare the algorithms, we generate different graphs and datasets with a multivariate Gaussian distribution using pcalg. To test the run time with respect to the number of nodes, we generate a random DAG where edges are included with a probability of 0.3 in every step. For the second line of experiments that have the goal to evaluate the run time with respect to sample size, we use a random graph with 5 nodes.

Figure 3.7 shows that FCI, using the non-parametric tests RCIT and RCoT, is significantly slower than with using Fisher's z-transform. The GES algorithm does not perform conditional independence tests and is faster than any version of FCI. The log-scale shows that FCI scales exponentially in the number of variables. This makes sense, as the number of conditional independence tests scales exponentially as well. Strobl et al. [SVZ19] claimed that both RCIT and RCoT are approximately linear in sample size. Our experiments verify this, see Figure 3.8.

3.5 NonlinearICP

The setup of nonlinearICP is a bit different to that of the FCI and GES algorithms, as it uses invariance-based causality. For our experiments, we define the environment

variable $E \in \{1, 2\}$. The different experimental setups either use E as an additive or multiplicative factor. Our models consist of two source variables X^0 and X^1, as well as a target variable Y. The source variables of experiment one are additive

$$\begin{aligned} X^0 &= E + 0.1 \cdot N_0, \\ X^1 &= E + X^0 + 0.1 \cdot N_1, \end{aligned} \tag{3.5}$$

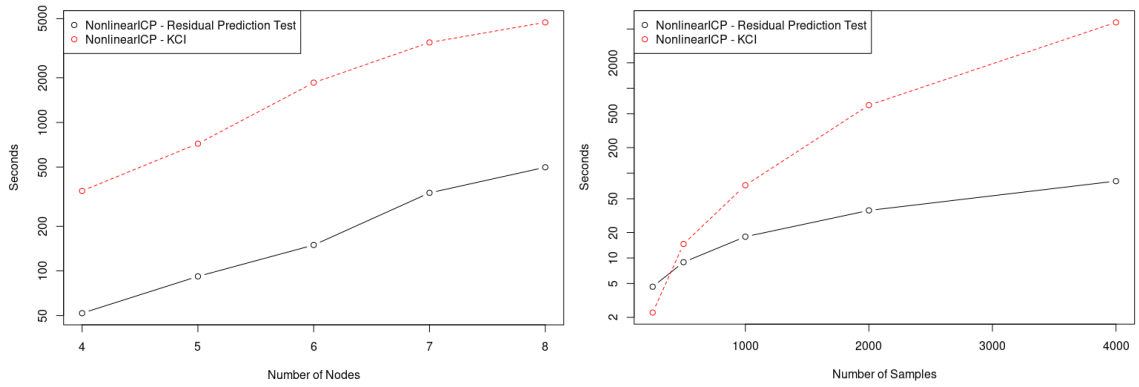
and the ones of experiment two have a multiplicative term

$$\begin{aligned} X^0 &= 0.1 \cdot E \cdot N_0, \\ X^1 &= E \cdot X^0 + 0.1 \cdot N_1. \end{aligned} \tag{3.6}$$

In both cases, Y is of the form $Y = f_0(X^0) \cdot f_1(X^1) + 0.1 N_Y$. The noise variables N_0, N_1, and N_Y are Gaussian distributed. We try different f^0 and f^1 from the set of functions $\{sin(x), sin(x^2), 1/x, x^2, x^3\}$. While nonlinearICP has different options for conditional independence tests, most of them gave similar results so that we only took the kernel-based test KCI and the default Residual Prediction Test. The datasets consist of $T = 1000$ samples.

The experiments showed that KCI led to the best results. In fact, they were correct every single time. This is remarkable, as it is the only algorithm that we tried with such good results. The results of nonlinearICP with the default option were also quite good, but not all non-linearities, namely $sin(x^2)$ and $1/x$, were found.

The reason for KCI not being the best option in every application is that its run time is much slower than that of the other conditional independence tests. In our run time experiments, we used a similar setup as in Equations 3.5 and 3.6. Figure 3.9 shows that the algorithm scales exponentially in the number of variables and less than exponentially in the number of samples. According to Heinze-Diemel et al. has cubic scaling in sample size [HDMP18].



(a) Run time wrt. the number of variables (b) Run time wrt. the number of samples

Figure 3.9: The run time of nonlinearICP with the kernel-based test KCI and the Residual Prediction Test. Both graphs use a log-scale.

Chapter 4

Finding Causal Relationships - Dynamical Systems

We will consider four algorithms for multivariate dynamical systems following the principles of the different notions of causality, as well as algorithms that treat simpler cases such as bivariate systems. The algorithm PCMCI [RNK$^+$19, Run20] uses graphical causality and combines it with an information theoretic measure to test for associations (Section 4.2). Multivariate transfer entropy applies the idea of Granger to a multivariate and non-linear setting, see Section 4.3. In Section 4.4, we discuss CausalKinetiX which exploits data from a multi-domain setting to find invariant models. Lastly, convergent-cross mapping uses a topological approach to find causal relationships in deterministic systems (Section 4.5).

The starting point of causal analyses of dynamical systems was Granger's linear regression technique for vector autoregressive models. Apart from econometrics (Granger himself was an economist), it has been widely adopted in various domains, such as environmental science [Ste16], neuroscience [BS16], political science [Fre83], as well as all other fields where time series are analyzed. A popular approach to test for causality proposed by Granger himself is to use linear regression [Gra69], but there are many other methods that can be employed [HSPVB07]. The implementations of linear Granger causality work well and reliably, and there are many options available for bivariate and multivariate data, e.g. in the R package *MTS* [TW20]). We shortly discuss the bivariate and non-linear case in Section 4.1, before coming to the four algorithms which can be used on multivariate data. In Section 4.6, we will test them on different datasets.

4.1 Bivariate Dynamical Systems

The bivariate case is easier than the multivariate as there are no indirect effects. We test the implementation *RTransferEntropy* [BZDP20] which calculates transfer entropy in bivariate and non-linear data. Continuous data gets discretized with a finite number of bins on which an estimator for the Shannon entropy is used. Transfer entropy is then approximated by an unbiased estimator using the empirical probability densities that have been calculated using the bins.

We use different time series $(X_t, Y_t)_{t\in\mathbb{N}}$ given by a structural causal model of the form

$$\begin{aligned} X_t &= N_t^X, \\ Y_t &= f(X_{t-\tau}) + N_t^Y, \end{aligned} \tag{4.1}$$

where N_t^X and N_t^Y are i.i.d. noise variables with standard Gaussian distribution. The algorithm is tested with a time lag $\tau \in \{1, 2, 3\}$ and non-linear functions $f(x) \in \{sin(x), 1/x, x^2, 1/sin(x), 1/sin(x)^2\}$.

The experiments show that this implementation of transfer entropy is only able to output good results for $\tau = 1$, where the causal direction is found for all functions f. Somehow, it does not find any relation having a time lag of $\tau > 1$. The p-values of the tests, which indicate how confident the algorithm is with its decision, is zero for all experiments, implying absolute certainty. Hence, the results are promising to a certain degree, but questions remain why the implementation is not able to handle a time lag greater than one. It is advisable to take great care when using bivariate transfer entropy in practice.

4.2 PCMCI

PCMCI [RNK+19, Run18b, Run18a, RPD+15, Run15] is an algorithm which is implemented in the *Tigramite* package for Python, published by Jakob Runge [Run20]. It uses two main steps. In the first one, a version of the PC algorithm is used that has been slightly adapted for time series to compute estimates of the parent sets of each variable.
In the second step we compute the information theoretical measure *momentary information contribution* (MCI) for each pair of variables and condition on the estimated parent sets from step one. In this way, only the true direct effects are computed. Note that the definition of MCI is given in Equation 4.4.

PCMCI uses the following setup. Assume that we have a stationary, multivariate time series $(\mathbf{X}_t)_{t\in\mathbb{N}}$, $\mathbf{X}_t = (X_t^1, ..., X_t^d)$, with a maximum time lag $\tau_{\max} \in \mathbb{N}$. We will refer to X^i as a (stochastic) process and X_t^i as a (random) variable. Recall that stationarity implies that if $X_t^i \to X_s^j$ for some $t < s$, then also $X_{t+t'}^i \to X_{s+t'}^j$ for all $t' > 0$. Instead of searching for causal relations between processes, PCMCI tries to find for causal relations between random variables. The difference to the i.i.d. case is that it is enough to find a relation between X_t^i and $X_{t+\tau}^j$ for some time lag $\tau > 0$ and $t \in \mathbb{N}$, to get causal information about the whole processes.

One can transfer a lot of Pearl's graphical causality to the time series case, such as the connection between edges in the graph and conditional dependencies in the distribution of the data.

$$(X_{t-\tau}^i \to X_t^j) \notin E \Leftrightarrow (X_{t-\tau}^i \perp\!\!\!\perp X_t^j) \mid \mathbf{X}_t^- \backslash \{X_{t-\tau}^i\} \tag{4.2}$$

Here, $\mathbf{X}_t^- = \{\mathbf{X}_{t-1}, ..., \mathbf{X}_{t-\tau_{\max}}\}$ is the (relevant) past of the multivariate process. The parent set of node X_t^j is defined as

$$\mathcal{P}(X_t^j) := \{X_{t-\tau}^i \mid 1 \leqslant i \leqslant d, 1 \leqslant \tau \leqslant \tau_{\max}, X_{t-\tau}^i \to X_t^j\}. \tag{4.3}$$

As MCI is a symmetric measure, it is not possible to find the direction of a causation between two variables from the same time step, i.e. for instantaneous effects with $\tau = 0$. We use a minimum time lag of $\tau = 1$ and will not treat instantaneous effects.

The process of finding causal parents has to be adapted to time series. In the i.i.d. case, we said that for $X, Y \in V$, if we assume faithfulness and the Markov property, an edge $X - Y$ exists if and only if there is no $\mathbf{Z} \subset V \backslash \{X, Y\}$ such that $X \perp\!\!\!\perp Y \mid \mathbf{Z}$. In Equation 4.2, we use $\mathbf{Z}$ as the set of all other possible variables. In the i.i.d. case, it would simply be wrong to condition on a set which is too large, as we never know if this could introduce additional dependencies (see v-structures in Section 2.1.3). However, in the time series case, this does not pose a problem, as the only nodes that can render $X^i_{t-\tau}$ and X^j_t dependent are common children. As such, they come after time point t and we can condition on everything that comes before t without having the problem of creating additional dependencies.

Algorithm 1 First step of PCMCI: condition selection

Require: Time series dataset $\mathbf{X} = (X^1, ..., X^d)$, selected variable X^j, maximum time lag $\tau_{\max} \in \mathbb{N}$, significance threshold $\alpha \in (0, 1)$, and maximum condition dimension $p_{\max} \in \mathbb{N}$.

Initialize preliminary set of parents $\widehat{\mathcal{P}}(X^j_t) := \{X^i_{t-\tau} : i = 1, ..., d, \tau = 1, ..., \tau_{\max}\}$
Initialize dictionary of test statistic values $I^{\min}(X^i_{t-\tau} \to X^j_t) := \infty \; \forall X^i_{t-\tau} \in \widehat{\mathcal{P}}(X^j_t)$
function ComputeParentEstimate($\mathbf{X}, X^j, \tau_{\max}, \alpha, p_{\max}$)
    for $p = 0, ..., p_{\max}$ **do**
        if $|\widehat{\mathcal{P}}(X^j_t)| - 1 < p$ **then**
            Break for-loop
        end if
        for all $X^i_{t-\tau} \in \widehat{\mathcal{P}}(X^j_t)$ **do**
            Define the set $\mathcal{S} \subset \widehat{\mathcal{P}}(X^j_t) \backslash \{X^i_{t-\tau}\}$ of the p elements with strongest association
            (p-value, test statistic value I) $\leftarrow$ ConIndep($X^i_{t-\tau}, X^j_t, \mathcal{S}$)
            if $|I| < I^{\min}(X^i_{t-\tau} \to X^j_t)$ **then**
                $I^{\min}(X^i_{t-\tau} \to X^j_t) := |I|$
            end if
            if p-value $> \alpha$ **then**
                Mark $X^i_{t-\tau}$ for removal from $\widehat{\mathcal{P}}(X^j_t)$
                Break from inner for-loop
            end if
        end for
        Remove non-signicant parents from $\widehat{\mathcal{P}}(X^j_t)$
        Sort parents in $\widehat{\mathcal{P}}(X^j_t)$ by $I^{\min}(X^i_{t-\tau} \to X^j_t)$ from largest to smallest.
    end for
    return $\widehat{\mathcal{P}}(X^j_t)$
end function

Algorithm 1 is the first step of PCMCI. It is similar to the PC algorithm, but adapted to the time series case. The default value for the maximum size of conditioning sets is

Algorithm 2 Second step of PCMCI: causal discovery

Require: Time series dataset $\mathbf{X} = (X^1, .., X^d)$, for all variables X^j the sets $\widehat{\mathcal{P}}(X_t^j)$ computed in the first step as well as the maximum number of parents $p_j \in \mathbb{N}$, maximum time lag $\tau_{\max} \in \mathbb{N}$.

1: **function** TestMCI($\mathbf{X}, \{\widehat{\mathcal{P}}(X_t^j) : j = 1, ..., d\}, \tau_{\max}, \{p_j : j = 1, ..., d\}$)
2: **for all** $(X_{t-\tau}^i, X_t^j)$ with $i = 1, ..., d$ and $\tau = 0, ..., \tau_{\max}$, excluding (X_t^j, X_t^j) **do**
3: Remove $X_{t-\tau}^i$ from $\widehat{\mathcal{P}}(X_t^j)$ if it is in the set
4: Define $\widehat{\mathcal{P}}_{p_i}(X_{t-\tau}^j)$ as the first p_i parents from $\widehat{\mathcal{P}}(X_{t-\tau}^i)$
5: Run MCI test to obtain (p-value, I) $\leftarrow$ ConIndep($X_{t-\tau}^i, X_t^j, \mathbf{Z} = \{\widehat{\mathcal{P}}(X_t^j), \widehat{\mathcal{P}}_{p_i}(X_{t-\tau}^j)\}$)
6: **end for**
7: **return** All p-values and MCI test statistic values
8: **end function**

$p_{\max} = d\tau_{\max}$. The algorithm uses PC_1 to reduce the computation time, where only the set with the p elements with strongest association is chosen, instead of all sets of size p. Otherwise there would be another for-loop in the ninth line of Algorithm 1.
The goal is to find supersets of the parent sets that can be used as conditioning sets in Algorithm 2. There we test again for all pairs $(X_t^j, X_{t-\tau}^i)$ whether $X_{t-\tau}^i \to X_t^j \in E$. For this we use momentary conditional independence (MCI) defined by

$$\text{MCI: } X_{t-\tau}^i \perp\!\!\!\perp X_t^j \mid \widehat{\mathcal{P}}(X_t^j) \backslash \{X_{t-\tau}^i\}, \widehat{\mathcal{P}}_{p_i}(X_{t-\tau}^i), \tag{4.4}$$

where $\widehat{\mathcal{P}}(X_t^j)$ denotes the estimate of the parent set and $\widehat{\mathcal{P}}_{p_i}(X_{t-\tau}^i) \subset \widehat{\mathcal{P}}(X_{t-\tau}^i)$ is the set of the p_i parents with the biggest information contribution. Note that indeed all combinations of variables are checked again, even if the PC algorithm in the first step indicated that some variables are conditionally independent. This increases the run time, but leads to more stable results in theory.

PCMCI can be used with different conditional independence tests. There are options for a multivariate Gaussian (partial correlation test) as well as for non-parametric data. We concentrate on the latter case and focus on the k-nearest-neighbor-based test CMIknn [Run18b, Run20]. As an alternative, the test RCIT [SVZ19, Str20], which we have seen in Section 3.1, is also available.

Our experiments show that the algorithm's output strongly depends on the parameters. The user has to define the minimum and maximum time lag, the conditional independence tests, and the significance level α. This requires knowledge about the structure of the data, especially the choice of the maximum time lag is delicate and usually even domain knowledge is not sufficient for estimating it well. If it is chosen too large, the algorithm can find all correct relationships, but the risk of making mistakes is much higher, the result less interpretable, and the run time slower. On the other hand, with a small maximum time lag it might not be possible to capture all causal relations.
The significance level can also lead to very different results. If it is chosen too large, we include too many or even all potential nodes in the conditioning set (see line 14

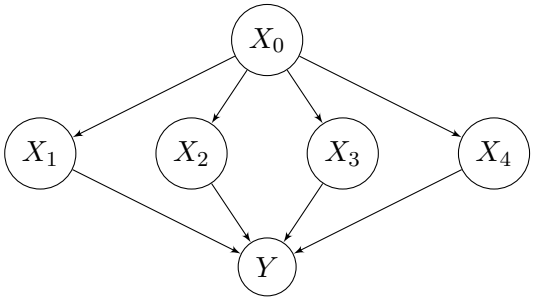


Figure 4.1: This graph shows a summary graph of six-dimensional dynamical system with $n = 4$ intermediate processes. In this graph the effect of X_0 to Y gets mediated through all other variables. For the algorithms the indirect effect might seem stronger than the direct ones.

Process	Time Lag	p-value	Strength of causal relation
X_0	2	0.0	0.279
X_1	1	0.0	0.150
X_2	1	0.0	0.161
X_3	1	0.0	0.218

Table 4.1: Parents of variable Y

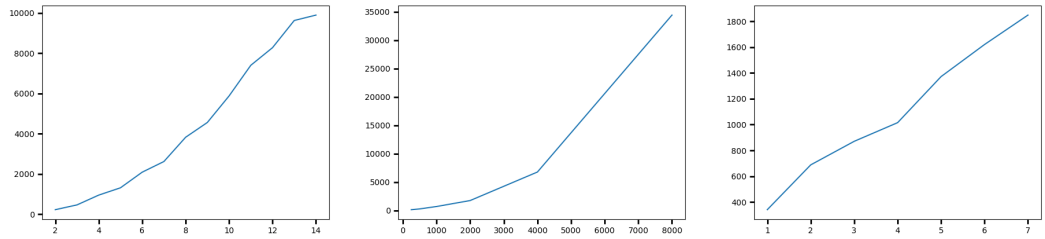
in Algorithm 1). Then, Algorithm 2 gets really slow as the conditional independence tests scale with the number of conditioning variables. The significance level can also be too small, so that we might not include all parents in the conditioning set. If the conditioning sets are empty, then Algorithm 2 reduces to a standard test of MCI.

The two-step procedure of PCMCI aims to minimize the chance of mistakes. However, if a mistake does happen in the first step and one of the sets found is not a superset of the respective parent set, PCMCI runs into trouble with distinguishing direct from indirect effects. To show this, we use the setup shown in Figure 4.1 with different numbers n of intermediate processes and artificially limit the size of the conditioning set of Y to only one element.

The result for $n = 3$ can be seen in Table 4.1. Recall that a p-value of zero implies absolute certainty of the decision. The information flow from X_0 to Y is the strongest, even though it is mediated through other variables, so that it should not even appear in this list. If the conditioning set of Y consisted of all intermediate processes, then no mistake with confusing direct and indirect effects would have happened.

Of course, in this example PCMCI is purposely deceived to conclude that X_0 is a direct parent of Y. But in practice, too, the algorithm might not always make correct choices during the first step. Thus, we argue that one has to be careful with the output of PCMCI and keep in mind that some of the effects shown might be indirect.

The computational complexity of PCMCI can be calculated with multiplying the



(a) Run time in seconds wrt. the number of variables

(b) Run time in seconds wrt. the number of samples

(c) Run time in seconds wrt. the maximum time lag

Figure 4.2: The run time of PCMCI with respect to the three parameters number of variables, number of samples and maximum time lag.

complexity of the conditional independence test (used in line ten of Algorithm 1 and line five in Algorithm 2) with the number of tests that have to be carried out in the worst case, i.e. when the network is fully connected. The number of tests of step one amounts to

$$d \sum_{p=0}^{d\tau_{\max}-1} d\tau_{\max} = d^3\tau_{\max}^2.$$

The second step involves $d^2\tau_{\max}$ conditional independence tests with a dimensionality of $2+|\widehat{\mathcal{P}}(X_t^j)|+|\widehat{\mathcal{P}}(X_{t-\tau}^i)|$. The complexity of conditional independence tests typically scales in dimensionality and the length T of the time series, amongst others. As there are several tests that can be used for PCMCI, we will not give a detailed run time analysis of every test that is mentioned in this thesis and refer to the respective papers.

The experiments we carried out were done with 500 samples, four variables, and a maximum time lag of $\tau_{\max} = 2$ in the minimal setting. In Figure 4.2, one can see that the algorithm scales not much slower than linear in all three cases. It is possible that conducting experiments with longer time series and more variables would lead to more interesting results. However, such experiments were not feasible, due to limited computational power.

The version of PCMCI that was used for the experiments did not yet support parallelization. However, this is a planned feature, so given sufficient computing resources, it may be able to treat systems with many variables better in the future. Without parallelization all target variables are treated one after another to compute the respective parent sets, so that one can expect a significant speed-up with this feature.

4.3 Multivariate Transfer Entropy

The algorithm multivariate transfer entropy (multivariate TE), implemented in the R package *IDTxl* [Wol18, Wol20a], is based on the ideas of Wiener and Granger and uses conditional mutual information, an information theoretic measure defined in Section 2.3, to test for Granger causality [NWM+19]. Multivariate TE is a heuristic approach,

defining those variables as parents that contribute a significant amount of information to the target. A true causal parent that contributes only little information is not considered as such. However, as most of the algorithms that try to infer a causal structure from data have to rely on some kind of statistical testing, this heuristic does not mean that there is a big difference between multivariate TE and other algorithms.

The algorithm gets a target process $Y = (Y_t)_{t\in\mathbb{N}}$ as input and tries to find the source processes in a set $\mathbf{X} = \{X^1, ..., X^d\}$ where $X^i = (X^i_t)_{t\in\mathbb{N}}$. Multivariate TE quantifies the amount of information that flows from the past of X^i to Y_t when taking into account the information that is provided by the past of Y and of $\mathbf{X}\backslash X^i$. As conditioning on the entire past of a process is intractable, the set we condition on in practice needs to be significantly smaller.

Algorithm 3 Multivariate transfer entropy

Require: Time series source processes $\mathbf{X} = (X^1, ..., X^d)$, target process Y, maximum time lags $\tau_Y \in \mathbb{N}$ and $\tau_X \in \mathbb{N}$ for target and source processes, and significance level $\alpha \in (0, 1)$.

Define $Y^C_{<t} := \{Y_{t-1}, ..., Y_{t-\tau_Y}\}, X^C_{<t} := \{X_{t-1}, ..., X_{t-\tau_X}\}$
Initialize $X^S_{<t}, Y^S_{<t} := \varnothing$.

repeat ▷ Step 1
    Compute CMI contribution $I(C; Y_t | Y^S_{<t})$ for all $C \in Y^C_{<t}$
    Select C^* maximizing the CMI contribution. Use maximum statistic to test for significance. If it is, add C^* to $Y^S_{<t}$ and remove it from $Y^C_{<t}$.
until maximum CMI contribution is not significant or $Y^C_{<t}$ is empty

repeat ▷ Step 2
    Compute CMI contribution $I(C; Y_t | Y^S_{<t}, X^S_{<t})$ for all $C \in X^C_{<t}$
    Select C^* maximizing the CMI contribution. Use maximum statistic to test for significance. If it is, add C^* to $X^S_{<t}$ and remove it from $X^C_{<t}$.
until maximum CMI contribution is not significant or $X^C_{<t}$ is empty

repeat ▷ Step 3
    Compute CMI contribution $I(C; Y_t | Y^S_{<t}, X^S_{<t}\backslash\{C\})$ for all $C \in X^S_{<t}$
    Select C^* minimizing the CMI contribution. Use minimum statistic to test for significance. If it is, remove C^* from $X^S_{<t}$.
until minimum CMI contribution is not significant or $X^S_{<t}$ is empty

Perform omnibus test to test whether the CMI contribution of $X^S_{<t}$ is significant. If not, set $X^S_{<t} = \varnothing$. ▷ Step 4

With an abuse of notation we define $X^C_{<t} \subset \mathbf{X}$ and $Y^C_{<t} \subset Y$ as the sets of possible candidates. Here, $\mathbf{X}$ and Y are used as collections of random variables, i.e. $Y = \{Y_n : n \in \mathbb{N}\}$ and $\mathbf{X} = \{X^i_n : 1 \leqslant i \leqslant k, n \in \mathbb{N}\}$. The goal is to find the sets of relevant sources $X^S_{<t} \subset X^C_{<t}$ and $Y^S_{<t} \cup Y^C_{<t}$.

In order to better handle auto-correlation, we compute the conditional mutual inform-

ation $I(C; Y_t|\mathbf{Z})$ for all $C \in Y^C_{<t}$ in step one (recall the definition of CMI in Section 2.3.3). For the C^* with the maximum information contribution, we perform a significance test against a distribution estimated from surrogate data. This surrogate data is obtained by permuting the time series. The p-value denotes the fraction of surrogate estimates where the test statistic has a more extreme value than the original estimate of information contribution [Wol18]. If the p-value is below a significance threshold, we say that the result is significant, i.e. a high enough information contribution in our case.

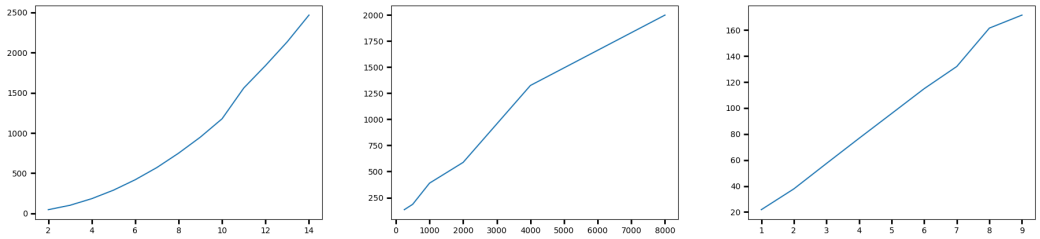
The test statistic used here is the maximum statistic. We compute the information contribution of all surrogates of $C \in Y^C_{<t}$ and take the maximum. This is repeated many times to get the distribution of the maximum values. Then, we can calculate the p-value with taking the fraction of maximum values bigger than $I(C^*; Y_t|\mathbf{Z})$ and compare it to the significance level α. If the information contribution of C^* is indeed significant, we add C^* to $Y^S_{<t}$ and remove it from the candidate set. This procedure is repeated until $Y^C_{<t}$ is empty or C^* is not significant. In the second step the same process is done for $X^C_{<t}$.

After adding all variables with significant information contribution to $X^S_{<t}$, we perform a pruning step to see whether some variables are no longer relevant (step three). This can easily happen as we condition on a small $X^S_{<t}$ in the beginning. We compute $I(C; Y_t \mid (X^S_{<t} \cup Y^S_{<t}) \backslash C)$ for all $C \in X^S_{<t}$ and choose C^* with the minimal information contribution. The minimum statistic is computed analogously to the maximum statistic and is used to test whether the information contribution of C^* is still significant. If not, we remove it and iterate until the information contribution of every single variable in $X^S_{<t}$ is significant.

The remaining task is to test whether the total information transfer from all source variables $X^S_{<t}$ is significant. In the fourth step, we permute the realizations of Y_t to obtain Y'_t and calculate $I(X^S_{<t}; Y'_t|Y^S_{<t})$. Repeating this procedure many times, we can calculate the p-value and decide whether the source variables $X^S_{<t}$ contribute a significant amount of information. If not, all variables are removed and the set of relevant sources is returned empty. As a last step, the multivariate transfer entropy from X^i to Y can be calculated by taking from $X^S_{<t}$ all variables of X^i's past, A^i, and calculating $I(A^i; Y_t \mid (X^S_{<t} \cup Y^S_{<t}) \backslash A^i)$.

There is one big disadvantage of multivariate TE in the context of finding causal relationships: it does not distinguish between direct and indirect effects, only between stronger and weaker effects. Consider as example the target variable Y and two source processes X^1 and X^2 with $X^1_{t-2} \to X^2_{t-1} \to Y_t$, then the effect of X^1 on Y gets mediated through X^2. If, for some reason, the direct effect is weaker than the indirect, then it will choose the latter and probably ignore the former. With just three variables, this is quite unlikely to happen. In a setting with more variables, however, this is a valid concern.

We again use the setup of Figure 4.1 to show how the algorithm can be misled. For an increasing number n of intermediate processes, the information contribution of X^0



(a) Run time wrt. the number of variables

(b) Run time wrt. the number of samples

(c) Run time wrt. the maximum time lag

Figure 4.3: The run time of multivariate TE with respect to the three parameters number of variables, number of samples and the maximum time lag.

to the target becomes stronger and the contributions of X^1 to X^n get weaker. Our experiments show that for $n < 3$, X_0 is not found as a parent. For $3 \leqslant n \leqslant 8$, it is considered as a parent in addition to the true parents, and for $n > 8$, it is found as the only parent of Y. In other words, only for few intermediate variables the algorithm outputs the true result. For many, it even refers to X_0 as the only parent, even though the effect is indirect and the intermediate variables are the true causal parents.

According to Wiener's principle, we want to know what information other time series provide additionally to the information of the target's own past [WPP+13]. The implementation of multivariate TE follows this principle. Unlike PCMCI, which treats the past time steps of all processes equally, multivariate TE first measures how the target process is affected by auto-correlation. Other potential sources are only considered in the second step and have to provide strictly more information than the past of the target variable.

The number of CMI calculations scales, in the worst case of a fully connected network, with $O(k^3 \tau_{\max} S)$ where S is the number of surrogate calculations. Conditional mutual information can be calculated with the non-parametric Kraskov estimator [KSG04], which scales with $O(KT \log T)$, where T is the length of the time series and K denotes the number of nearest neighbors considered during the estimation. On a CPU, the algorithm can be parallelized over targets; and computation on a GPU even allows for parallelization over targets and surrogates, significantly cutting down on run time [Wol20b].

Our experiments show that multivariate TE with the Kraskov estimator scales almost linearly with respect to number of samples and maximum time lag, see Figure 4.3. It scales worse than linear in the number of variables, but more extensive experiments would be necessary to establish the exact relation.

4.4 CausalKinetiX

The problem of finding the causal structure of continuous-time systems is closely related to finding the structure of ODEs. The Picard-Lindelöf theorem states that there

is at least locally a solution to an ODE $\dot{X} = f(X)$ if the function f is Lipschitz [CL55]. This implies that it is possible to predict the immediate future given the past values, so that ODEs get causal interpretation [Sch19]. This connection is exploited by Pfister et al. [PBP19], who propose an algorithm to find the causal structure of a kinetic system via estimating the structure of the underlying ODEs. They use invariance-based causality to incorporate knowledge from different environments (e.g. stemming from interventions) and want to find an invariant model that predicts the data well across the environments.

Pfister et al. highlight the connection of predictability and invariance explicitly. The former is the standard goal of machine learning techniques, while the latter is the key to causality. By taking into account the different environments they want to improve the model, so that it generalizes better to unseen data. CausalKinetiX [PBP20] transfers the methods of i.i.d. data [PBM16, HDMP18, ABGLP19] to dynamical systems. The approach of CausalKinetiX is similar to the one of score-based algorithms in the i.i.d. case: there is a space of possible models and the algorithm searches for the one that fits the data best.

It is assumed that the dynamical system is described by ODEs and measured at discrete time points. As input data of CausalKinetiX, observational and/or interventional data of a so-called causal kinetic model are taken.

Definition 4.4.1. *A causal kinetic model over processes* $(\mathbf{X}_t)_{t\in\mathbb{R}_{\geqslant 0}} := (X_t^1, ..., X_t^d)_{t\in\mathbb{R}_{\geqslant 0}}$ *is a finite collection of* d *ODEs*

$$\begin{aligned}
\dot{X}_t^1 &:= f^1(X_t^{\mathbf{PA}_1}, X_t^1),\ X_0^1 = x_0^1\\
\dot{X}_t^2 &:= f^2(X_t^{\mathbf{PA}_2}, X_t^2),\ X_0^2 = x_0^2\\
&\vdots\\
\dot{X}_t^d &:= f^d(X_t^{\mathbf{PA}_d}, X_t^d),\ X_0^d = x_0^d
\end{aligned}$$

where $\dot{X}_t^j$ *denotes the time derivative of the component* X^j *at time* t *and* $\mathbf{PA}_j \subset \{1, ..., d\}\backslash\{j\}$ *is the set of direct parents. The system of ODEs needs to be solvable. Interventions on the process correspond to replacing the* j*-th initial condition or the* j*-th ODE with*

$$X_0^j := x_0 \text{ or } \dot{X}_t^j := g(X_t^{\mathbf{PA}_j}, X_t^j),$$

for some $x_0 \in \mathbb{R}$ *and function* g*, where* $\mathbf{PA} \subset \{1, .., d\}\backslash\{j\}$ *is the set of new parents. The system of ODEs is still required to be solvable after the intervention.*

Note that the definition of the causal kinetic model is closely related to structural causal models. Furthermore, we want to stress that the possibilities for creating different environments as inputs for CausalKinetiX are numerous. It is assumed that the function is a version of the mass-action kinetic law.

$$\dot{Y}_t = f_\theta(\mathbf{X}_t) = \sum_{j=1}^{d} \theta_{0,j} Y_t^j + \sum_{i=1}^{d}\sum_{j=i}^{d} \theta_{i,j} X_t^i X_t^j \tag{4.5}$$

Here, $\theta \in \mathbb{R}^{d(d+1)/2+d}$ is the parameter vector. Let v be the sparsity pattern indicating the zero entries of θ. Define

$$G^v := \left\{ f_\theta : \mathbb{R}^d \to \mathbb{R} \mid \forall x \in \mathbb{R}^d : f_\theta(x) = \sum_{i,j} \theta_{i,j} x^i x^j, v \times \theta = \theta \right\}, \tag{4.6}$$

where $\times$ denotes the element-wise product. Let

$$\mathcal{M} := \{G^v, v \in \{0,1\}^{d(d+1)/2+d}\} \tag{4.7}$$

be a model space. As this space would become too large, CausalKinetiX tries to reduce its size. The default model space of the implementation is

$$\mathcal{M}_p := \{M \in \mathcal{M} : v \text{ has at most } p \text{ non-zeros}\}. \tag{4.8}$$

However, model spaces can be defined in any way that is best adapted to the task at hand. If the number of variables is too large, one can use screening or variable selection techniques that are based on predictability, such as Lasso [Tib96].

The algorithm does the following to compute the score of the models. We fit two cubic splines computed with the help of a convex quadratic program on the data. For the details we refer to [PBP19]. The first is calculated without any constraints, the second is required to go through the predictions of the current model. If the predictions are good, the difference of the two splines is small. If, on the other hand, they are not, then the second spline will have difficulties with properly fitting the data. The stability score is computed based on the difference of the two splines.

Apart from ranking models, it is possible to rank the importance of individual variables for the prediction of the target. In order to do this, we have to guess the number K of invariant models (i.e. models that are able to correctly describe the dynamics of the system in every experimental environment). Then, the algorithm just computes the fraction of invariant models that depend on a certain variable. For example, if variable X_1 appears L times in the K best-ranked models, then the score would be L/K.

To wrap up the algorithm, we can divide it into four steps: For a collection of models $\mathcal{M} = \{M^1, ..., M^m\}$, repetitions $i \in \{1, ..., n\}$, and (noisy) data $\tilde{Y}_{t_1}^{(i)}, ..., \tilde{Y}_{t_L}^{(i)}$ we need to do the first step once and repeat $(2) - (4)$ for every model M^j.

(1) Smooth target trajectories directly from the data using smoothing splines for each repetition.

(2) Fit for each repetition j the candidate target model across all repetitions of the other experiments and obtain fitted values for the repetition j (without actually using the information to obtain invariance).

(3) Smooth target trajectories with the new (estimates) of the data from the step before.

(4) Compute the score by comparing the two smoothed trajectories.

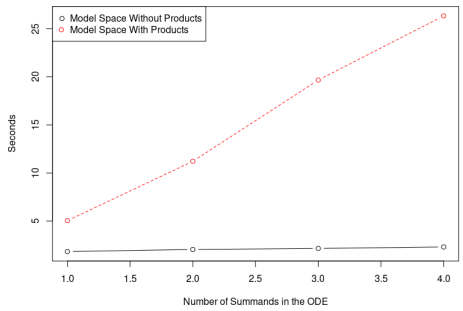


Figure 4.4: The run time of CausalKinetiX with respect the maximum number of summands in the ODEs.

According to the experiments of Pfister et al., the algorithm scales in the worst case cubicly in the number of variables and sample size, although in practice it is supposed to be faster [PBP19]. We tested the run time for different model spaces. We tested the run time with respect to the maximum number of summands of the ODEs. In one space, we allowed for products of the form X^2 (no products like XY are included yet), in the other one we did not. Figure 4.4 shows that the algorithm scales linearly, even though it is much slower for the space including the products.

4.5 Topological Causality and Convergent Cross Mapping

We use the same setup as in Section 2.4. The mappings $M_{i\to j}$ from the delay embedding r^i to the delay embedding r^j exist if and only if $w_{ij} \neq 0$, i.e. if and only if the system part X^i does not depend on the system part X^j where $i = 1, 2$ and $j = 3 - i$. This is exploited by both Harnack et al. [HLSP17] and Sugihara et al. [SMY+12] to create the algorithms topological causality (TC) and convergent cross-mapping (CCM). Both use a nearest-neighbor approach. They project the neighborhood of r_t^i to the other delay embedding of X^j and measure how well the projection performs. If it works well, then there is a dependency.
Note that the direction of the prediction and of the causal effects are 'reversed': there is a (causal) dependency from e.g. X^1 to X^2 if and only if X^2 helps to predict X^1, i.e. $w_{21} \neq 0$.

There are two ways to evaluate the quality of the projection, which leads to the two different algorithms. Let $\{t_1^i, ..., t_k^i\}$ be the k-nearest neighbors around r_t^i. Topological causality uses a local linearization $M_{i\to j}^t$ of $M_{i\to j}$, which is either estimated from data, or analytically computed as the Jacobian of a differentiable function. Then, singular values $\sigma_t^k(M_{i\to j}^t)$ are used to define the expansion $e_{i\to j}^t$ of $M_{i\to j}^t$.

$$e_{i\to j}^t := \prod_k \max(1, \sigma_t^k(M_{i\to j}^t)) \tag{4.9}$$

The expansion measures the quality of the mapping $M_{i\to j}$: the smaller it is, the

better the approximation of $\{r^j(t_1^j), ..., r^j(t_k^j)\}$, which is given by the projection of $\{r^i(t_1^i), ..., r^i(t_k^i)\}$ to $\{r^j(t_1^i), ..., r^j(t_k^i)\}$. Topological causality is defined as

$$C_{i\to j}^t := \frac{1}{1+\log(e_{i\to j}^t)},$$
$$C_{j\to i}^t := \frac{1}{1+\log(e_{j\to i}^t)}. \tag{4.10}$$

Convergent cross-mapping, on the other hand, uses a slightly different idea. It focuses not only on the measure itself, but rather on its convergence. In practice, convergence of the prediction is limited by observational error, process noise, and time series length T. Hence, increasing the time series length should improve predictions. If not, then no prediction is possible in the first place, and there is no causal relationship in the corresponding direction. Thus, CCM uses the predictability that increases with T and is a necessary condition for causation. The algorithm is implemented in the R package *rEDM* [SYC+20].

4.6 Comparison of the Algorithms

In the following, we want to evaluate the algorithms in different situations and for different kinds of data. In Section 4.6.2, we consider data from discrete-time dynamical systems where functions define the evolution of the process. In Section 4.6.3, the algorithms are tested on data from continuous-time systems where the dynamics are given by an ODE. In Section 4.6.4, we look at data from chemical reaction networks that can be either sampled using fixed time steps, or by sampling every reaction separately via Gillespie algorithm [Gil76, Gil77].
Not all algorithms are meant to be used on all data. PCMCI, for example, assumes that the underlying process has a discrete-time evolution. However, as PCMCI was explicitly tested for climatological data [RNK+19, RBB+19], we will try it out on continuous-time data as well. The same holds for multivariate TE. CausalKinetiX works only for dynamical systems based on ODEs and convergent-cross mapping only for deterministic systems.

4.6.1 Selection of Hyperparameters

PCMCI: The maximum time lag $\tau_{\max} = 5$, the minimum time lag $\tau_{\min} = 1$, the conditional independence test based CMIknn, and the significance level $\alpha = 0.05$.

Multivariate TE: The JIDT KSG estimator for CMI (the default non-parametric option for continuous data), the significance level $\alpha = 0.05$, the maximum time lag $\tau_{\max} = 5$, and the minimum time lag $\tau_{\min} = 1$.

CausalKinetiX: The maximum of three summands per ODE, products of the form X^2, additional models if they include summands of the form XY, 15 repetitions of the same experiment (one is declared to belong to another environment if all the data stems from the same environment), 100 samples per time series.

Convergent Cross Mapping: The maximum forecast horizon $tp = 0$ in the continuous-time case (i.e. a prediction horizon of 0 time steps in the future) and $tp = 1$ in the functional case, a minimum of 10 and a maximum of 150 data points (recall that CCM considers the convergence of the respective measure).

The time series that we sample have a length of $T = 1000$.

4.6.2 Functional Data

We consider data of the form

$$f(X_t^j) := f_j(\mathbf{PA}_j) + N_t^j \tag{4.11}$$

where $(N_t^j)_{t\in\mathbb{N}}^{0\leqslant j\leqslant d}$ are i.i.d. noise variables and

$$\mathbf{PA}_j \subset \{X_{t-\tau_{\max}}^0, X_{t-\tau_{\max}+1}^0, ..., X_{t-\tau_{\min}+1}^d, X_{t-\tau_{\min}}^d\} \tag{4.12}$$

is the parent set of X_t^j. We will perform two experiments: one where f_j is linear and one where it is non-linear. Both experiments are carried out with Gaussian, Cauchy and t-distributed noise, i.e. we use (1) $N_t^i \sim \mathcal{N}(0,1)$, (2) $N_t^i \sim \text{Cauchy}(0,1)$, and (3) $N_t^i \sim \text{t}(2)$.

The linear model is the following. Let $(\mathbf{X}_t)_t$ be a time series where $\mathbf{X}_t = (X_t^0, ..., X_t^4)$ and

$$\begin{aligned}
X_t^0 &:= 0.8X_{t-1}^0 + N_t^0,\\
X_t^1 &:= 0.5X_{t-1}^0 + 0.8X_{t-1}^1 + N_t^1,\\
X_t^2 &:= 0.5X_{t-1}^1 + 0.8X_{t-1}^2 + 0.5X_{t-1}^4 + N_t^2,\\
X_t^3 &:= 0.8X_{t-1}^3 + 0.5X_{t-2}^4 + N_t^3,\\
X_t^4 &:= 0.8X_{t-1}^4 + N_t^4,
\end{aligned} \tag{4.13}$$

where the initial values are given by the noise variables.

In Figure 4.5, as well as in Figures B.2 and B.1, it can be seen that PCMCI did a perfect job for the models with Gaussian and t-distributed noise, but did not find the link $X_{t-1}^4 \to X_t^2$ for the one with Cauchy-distributed noise. Overall, the results are good and one can consider the algorithm reliable.

Multivariate TE does not include auto-correlation in its output and uses a summary graph instead of a time series graph. An example can be found in Figure 4.6. All correct links are included in the output, but the links between 4 and 1 as well as 3 and 2 are incorrect, even though these effects are not even indirect. Apparently, the algorithm finds an information flow that goes backward in time at one point. For example, to get information from X^4 to X^1, it has to go one step backward from X^2 to X^1 (see Figure 4.5). However, a causation cannot go backward in time and a weakness of multivariate TE becomes apparent. Unlike PCMCI, it does not condition

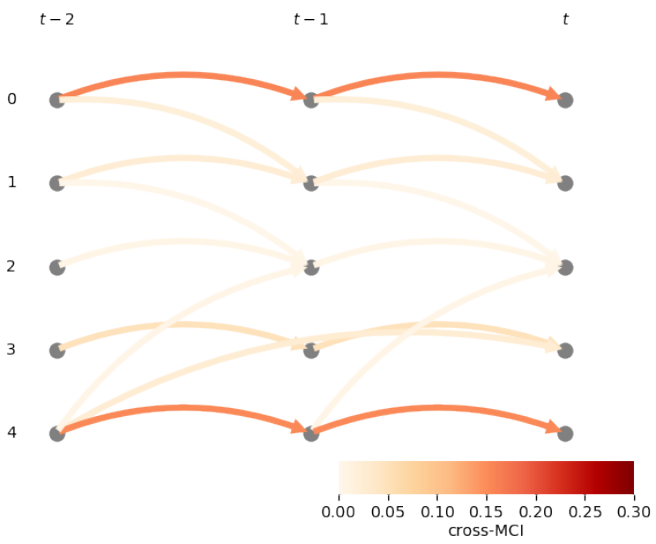


Figure 4.5: This is the output of PCMCI for the dataset described by the functions in Equation (4.13) with Gaussian noise. Recall that the data is stationary, so that all causal relations can be shifted in time.

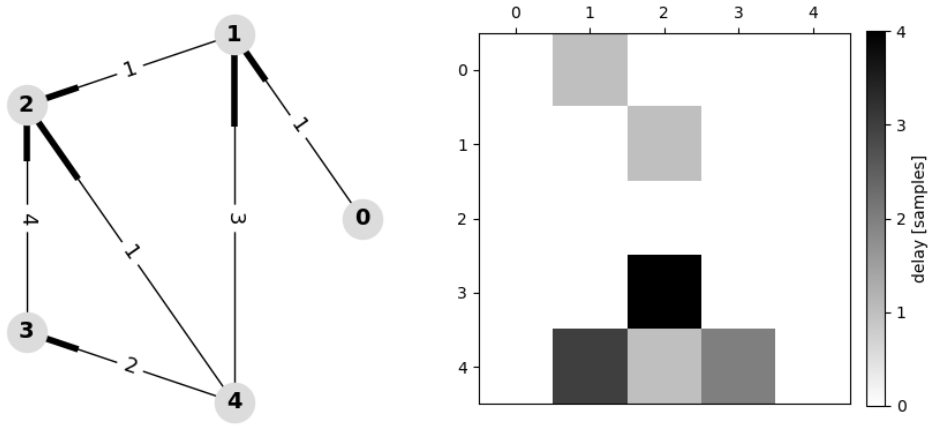


Figure 4.6: This is the output of multivariate TE for the dataset described by the functions in Equation 4.13 with Gaussian noise. The thick ends of the edges stand for edgemarks, i.e. they give the orientation of the edges. The numbers refer to the time lags of the causal relations. On the right-hand side there is a matrix which can be read in the following way: if the square (i, j) is non-white, then there is a causal relationship from X^i to X^j.

on the parent sets. Otherwise, it would take the information of X^1_{t-1} into account, making any information flow from X^4 to X^1 impossible.
Figure B.3 shows the result of the experiment with Cauchy-distributed noise; one obtains a similar picture. All true relationships were found, but also two additional dependencies that are different from the ones before. The time lag of the relationship between X^4 and X^3 is in reality $\tau = 2$ and not $\tau = 1$ as the output of multivariate TE suggests. Additionally, it found the false link $X^0_{t-1} \to X^2_t$. A relationship between the processes exists, but with time lag $\tau = 2$ and it is indirect not direct. Probably the information flow from X^0_{t-1} goes first the step back, and then over X^1_{t-1} to X^2_t.

Let us now consider the non-linear model $(\mathbf{X}_t)_t$ where $\mathbf{X}_t = (X^0_t, X^1_t, X^2_t)$ and

$$\begin{aligned} X^0_t &:= X^0_{t-2} + N^0_t, \\ X^1_t &:= 1/X^0_{t-1} + N^1_t, \\ X^2_t &:= \sin(X^1_{t-1}) + N^2_t. \end{aligned} \tag{4.14}$$

The results of the experiments with PCMCI (Figures B.5 - B.7) show that almost all non-linear relationships were found. Only the relationship $X^0 \to X^1$ has not been found in the case of Cauchy-distributed noise. Hence, even in the non-linear case, PCMCI can be trusted as long as the data is of the required form.

In Figures B.8 - B.10, we see that multivariate TE found the link $X^0 \to X^1$ only in the data with Gaussian noise. Since multivariate TE does not show auto-correlation in its outputs, the only link that is found in the Cauchy- and t-distributed data is $X^1_{t-1} \to X^2_t$. This time, the algorithm did not output more relationships than there actually are in the data.

We conclude that the results of PCMCI are a better than the ones of multivariate TE for data with both linear and non-linear relationships. PCMCI is able to reliably estimate most relationships in data where the dynamics are given by the functional approach from Equation 4.11.

4.6.3 ODE-Based Data

The algorithms were tested on datasets with two different underlying mechanisms. To integrate the ODEs we used the function *odeint* from the Python package *scipy*[1] with a sampling rate of 0.01 and the time interval $[0.0, 5.0]$. First, we simulated the Lorenz attractor. The system is described by

$$\begin{aligned} \dot{X}^0 &= \sigma(X^1 - X^0),\ X^0_0 := 1, \\ \dot{X}^1 &= X^0(\rho - X^2) - X^1,\ X^1_0 := 1, \\ \dot{X}^2 &= X^0 X^1 - \beta X^2,\ X^2_0 := 1, \end{aligned} \tag{4.15}$$

where $\sigma = 10, \rho = 28$, and $\beta = 8/3$. The second model is a variation of the Lotka-Volterra model, where we added a third variable. The dynamics are described by

[1] https://docs.scipy.org/doc/scipy/reference/generated/scipy.integrate.odeint.html, last accessed 2020-03-12

target	process	time lag	p-value	strength of causal relation
X_0	X_0	1	0.0	0.215
	X_1	4	0.009	0.002
	X_2	0	0.0	0.009
	X_2	2	0.001	0.011
	X_2	3	0.0	0.011
	X_2	4	0.0	0.01
	X_2	5	0.001	0.011
X_1	X_0	0	0.0	0.001
	X_0	1	0.0	0.001
	X_0	2	0.0	0.001
	X_0	3	0.0	0.001
	X_0	4	0.0	0.002
	X_0	5	0.004	0.01
	X_1	1	0.0	0.225
	X_1	5	0.004	0.002
X_2	X_1	0	0.0	0.007
	X_1	1	0.0	0.01
	X_1	2	0.006	0.011
	X_1	4	0.009	0.008
	X_1	5	0.006	0.011
	X_2	1	0.0	0.043

Table 4.2: The output of PCMCI for the dataset simulated from the Lorentz system.

$$\begin{aligned} \dot{X}^0 &= 0.1X^0 - 0.2X^0X^1,\ X_0^0 := 1, \\ \dot{X}^1 &= 0.3X^0X^1 - 0.1X^1,\ X_0^1 := 1, \\ \dot{X}^2 &= X^0 - X^2,\ X_0^2 := 1. \end{aligned} \tag{4.16}$$

Results of PCMCI and Multivariate Transfer Entropy

Although PCMCI and multivariate TE assume a discrete-time underlying process, we will try to find out how they behave on this kind of data as well. The results do not show a clear picture. PCMCI's output for the Lorenz system, shown in Table 4.2, tells us that all processes are interconnected, which is true. However, as expected, the notion of time lag loses its meaning and the algorithm shows many causal relationships from X^i to X^j for all $i, j \in \{0, 1, 2\}$. Both the summary graph of PCMCI's output and the one of multivariate TE (see Figure 4.7) are correct.
The results for the modified Lotka-Volterra system seem to be quite different. Both algorithms do not manage to find the true summary graph. As can be seen in Table 4.3, PCMCI finds many links again, but most of them are either due to auto-correlation or simply wrong. In reality, there is neither a causal relationship from X^2 to X^0,

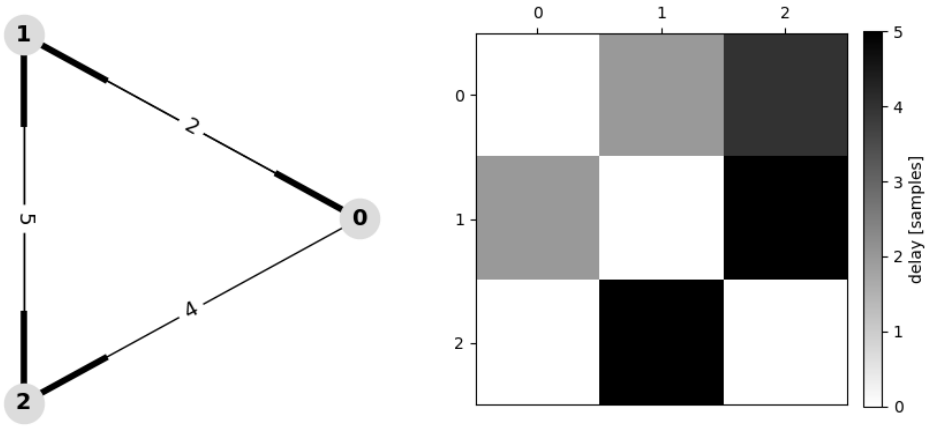


Figure 4.7: The output of multivariate transfer entropy of the experiment with the dataset of simulations of the Lorenz system.

nor from X^2 to X^1. The causal relationships from X^1 to X^2 and to X^0 are missing. Multivariate TE does not find any wrong links, but is missing two out of three correct ones in the summary graph, see Figure 4.8.

As a maximum time lag of $\tau_{\max} = 5$ is too much for a dataset where, in theory, the effects are instantaneous, we also tried PCMCI with $\tau_{\max} = 1$. We observe in Table 4.4 that, in this case, the algorithm found two of the correct causal relations and only missed the one from X^1 to X^0. Apparently, the results with different maximum time lags are not consistent, in the sense that one cannot just add all relationships with a certain time lag to a result with smaller maximum time lag. From a theoretical perspective, this comes from the fact that the parent sets we condition on depend on the maximum time lag.

We conclude for the two algorithms, that they are not really suited for continuous-time dynamical systems. One can probably find parameters so that the result is reasonable, but there is much doubt over where this is possible in actual applications with an unknown underlying structure.

Results of CausalKinetiX

First, we evaluate the algorithm on data stemming from the same environment. The algorithm is meant to discover the structure of the ODE which describes the dynamics of the target process. For example, if we have the ODE $\dot{X} = 2Y + 3Z - XY$, then we write the output as $\{\{Y\}, \{Z\}, \{X, Y\}\}$.
The true output for the Lorenz system is $\{\{X^0\}, \{X^1\}\}$ for the equation of X^0, $\{\{X^0, X^2\}, \{X^0\}, \{X^1\}\}$ for the equation of X^1, and $\{\{X^0, X^1\}, \{X^2\}\}$ for the equation of X^2.

target	process	time lag	p-value	strength of causal relation
X_0	X_0	1	0.0	0.079
	X_0	2	0.0	0.000
	X_0	3	0.0	0.000
	X_0	4	0.0	0.000
	X_0	5	0.0	0.000
	X_2	1	0.0	0.000
	X_2	2	0.0	0.000
	X_2	3	0.0	0.000
	X_2	4	0.0	0.000
	X_2	5	0.0	0.000
X_1	X_1	1	0.0	0.649
	X_2	1	0.0	0.000
	X_2	2	0.0	0.000
	X_2	3	0.0	0.000
	X_2	4	0.0	0.000
	X_2	5	0.0	0.000
X_2	X_2	1	0.0	0.001

Table 4.3: The output of PCMCI for the dataset simulated from the Lotka-Volterra system. Note that the strength of causal relations can be 0.000 due to rounding.

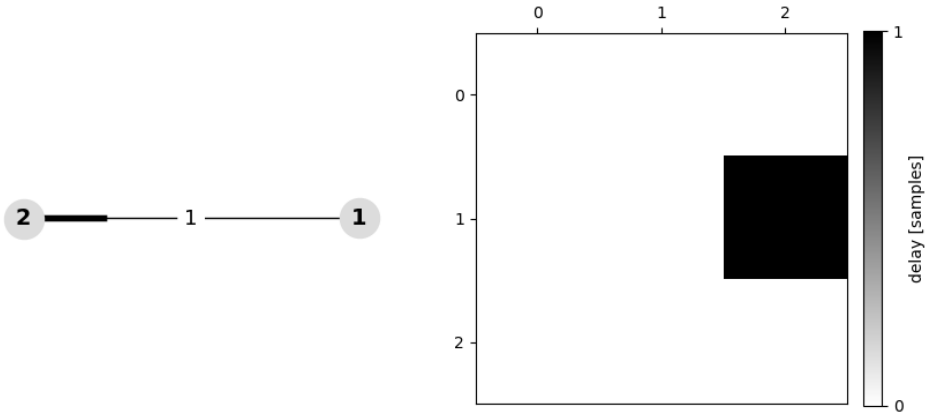


Figure 4.8: The output of multivariate TE for the modified Lotka-Volterra system.

target	process	time lag	p-value	strength of causal relation
X_0	X_0	1	0.0	0.093
X_1	X_0	1	0.022	0.016
X_2	X_1	1	0.0	0.098

Table 4.4: The output of PCMCI for the dataset simulated from the Lotka-Volterra system with $\tau_{\max} = 1$.

For the target X^0, CausalKinetiX gives five models a score below 0.03, while the remaining ones have scores larger than 200 (recall that the best score is zero). The true model is only ranked fifth. In several experiments, we observed that CausalKinetiX ranks models with fewer than the maximum number of summands (i.e. with one or two summands in our case) worse. The four models with a lower score are probably a perfect fit for the data as well, as they all have the same two summands with a varying third term.
The result of the algorithm for the targets X^1 and X^2 is perfect, as the true model has best score of almost 0 in both cases and all other models have scores of more than 1300 and 200 respectively.

Next, we look at the modified version of the Lotka-Volterra model with a third variable that is dependent on the first one. The true model is set to $\{\{X^0\}, \{X^0, X^1\}\}$, $\{\{X^1\}, \{X^0, X^1\}\}$, and $\{\{X^0\}, \{X^2\}\}$ for the targets X^0, X^1, and X^2 respectively.

The results are similar to the experiment above. For X^0 and X^1, the true model is found and ranked highest. For the target variable X^2, the algorithm put four models before the true model in the ranking that have both true summands plus a varying third.

To find out how CausalKinetiX profits from the multi-domain setting, we performed the same experiments with a varying number of interventional environments. For one experiment, we created up to five of them. The interventions we used have either the form $\dot{X} = 0$ to fix one variable to a certain value, or they reduced the number of summands that appear in the ODE, e.g. from $\dot{X} = f(X, Z) + g(Y)$ to $\dot{X} = f(X, Z)$.
One would expect the algorithm to perform better having this additional information, but the results were not as good as in the experiments with just one environment. The algorithm was not able to find the true structure of the ODE for any of the variables that we intervened on. Additionally, the algorithm had difficulties with finding the structure of the variables that were not intervened on. These results are probably caused by the alpha state of the implementation.
To summarize, CausalKinetiX can discover the model structure well in the case of purely observational data, although the best-ranked model is not unique in some cases. However, it does not yet appear to be suitable for a multi-domain setting.

Results of Convergent Cross-Mapping

We tried out convergent cross-mapping on all the datasets that we mentioned previously, but the algorithm gave the wrong answer almost every time. Either the effects were found in both directions and rated equally strong, or the wrong direction had a higher score. Only in its original setup described in Section 2.4, with equations of the form $\dot{X^1} = f_1(X^1, w_{12}\mu(X^2))$, the algorithm gave correct answers. We did not find a reason for this, even though it is likely that there is a structural error in the implementation, as the links were reversed in many experiments.

4.6.4 Data from Chemical Reaction Networks

Chemical reaction networks (CRNs) can be simulated in two different ways. Either, one takes fixed time steps and samples them like ODEs, or one uses no fixed time steps and simulates every single reaction that is happening separately. For the first option, we use the python package *CRN* [Bor20]. A reaction $X >> Y$ where species X transforms into species Y has a certain speed indicating how often the reaction takes place with respect to other reactions. More complicated reactions are possible, e.g. $X + Y >> Z$, where two molecules react and a new one is created. We simulate five seconds with a step size of 0.01. Gaussian noise is used to get a more realistic setting with measurement errors. We performed the experiments with data sampled from a CRN which translates into the following system of ODEs:

$$\begin{aligned} \dot{X}^0 &= -X^0,\ X_0^0 := 500, \\ \dot{X}^1 &= X^0 - 2X^1,\ X_0^1 := 300, \\ \dot{X}^2 &= 2X^1 - 0.2X^2,\ X_0^2 := 0, \\ \dot{X}^3 &= 0.2X^2,\ X_0^3 := 0. \end{aligned} \tag{4.17}$$

CausalKinetiX

We use again a maximum number of three summands per ODE. In this case, however, there are two equations with only one summand each, which turned out to be problematic in the experiments before. Another difficulty might be the chain of effects from X^0 to X^3, which might confuse the algorithm because of the presence of direct and indirect effects. To compare the results of the algorithm on CRN data with the performance on normal ODE data, we additionally use the same sampling technique for ODEs as before.

The results of the experiments show that CausalKinetiX is not able at all to find the true model for CRN data. In contrast, when the data was sampled using *odeint*, the algorithm found the true structure of X^1, ranked the true model of X^3 second and got the variable ranking for X^2 right. Additional models included have more summands, but appear to be perfect fits as well. The only target variable wide off the mark is X^0.

We conclude that CausalKinetiX behaves well on most of the data, as long as it strictly follows the rules of deterministic ODEs plus a random noise. As soon as there is more

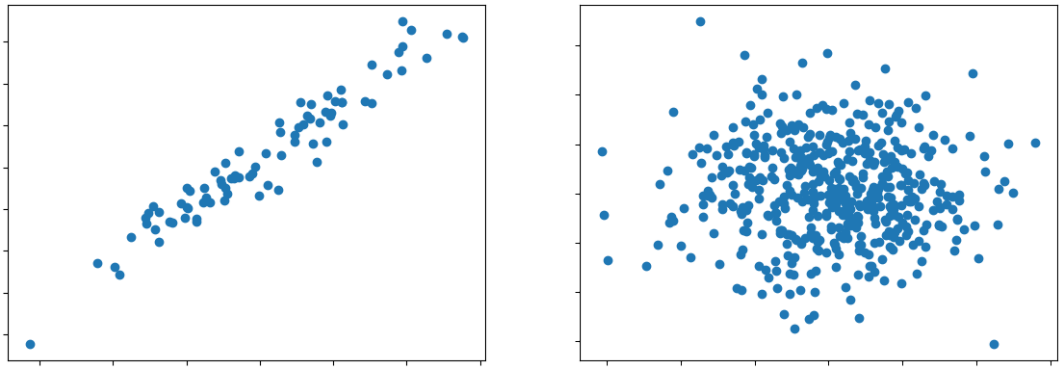


Figure 4.9: Here we can see a scatter plot of the two variables X (on x-axis) and Y (on y-axis). In the first example the variables are dependent, in the second one they are not.

randomness in the data, as is the case for chemical reaction networks, the algorithm's performance worsens drastically.

PCMCI

The other way to simulate CRNs is to use the Gillespie algorithm to simulate every reaction separately [Gil76, Gil77]. Essentially, this means that the process is sampled in discrete time steps. However, the dynamics are not described by a function, but by a stochastic process. We will see that we can use this setting to learn something about PCMCI.

At first, the results of PCMCI, run with the same parameters as before, were not satisfactory. To investigate its failure, we considered one target variable at a time. Assume that we have the chain of reactions $Z >> X >> Y$. Testing the conditional independence corresponds to fixing the value of Z and considering the values of X and Y for $Z = z$. One can visualize X and Y by plotting the values using a scatter plot. This plot can be used to confirm visually that X and Y are dependent or independent (see Figure 4.9).

As there are not sufficiently many data points of X and Y for a single value $Z = z$, we consider different hypercubes which each include a certain interval of values of Z as well as the corresponding X and Y values. This procedure is similar to the one used by the conditional independence test CMIknn. There, the size of the hypercubes is determined by the number of neighbors that are taken into consideration.

The size of the hypercubes, i.e. the number k of nearest neighbors, is a parameter which can be a lot more important than appears at first. The default option of PCMCI for the k-nearest-neighbors that are considered is a fraction of the number of samples. This leads to a large number of neighbors and to smaller variance in the decisions. However, it can also increase the bias when the local structure can only be seen in a smaller hypercube.

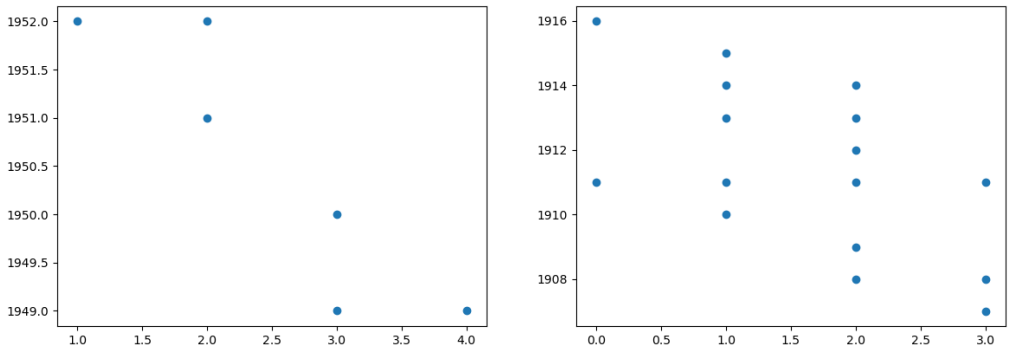


Figure 4.10: We have again a scatter plot of X (on x-axis) and Y (on y-axis). The values of X and Y do not matter for our purpose. On the left-hand side, the case of few neighbors, one could imagine a linear dependency. On the right-hand side, the case of many neighbors, one cannot.

In the case of our CRN data, the number of samples is far too big if we consider a fraction of the number of samples as k, so that the results of PCMCI were not satisfactory. Instead, we should have used only a small number of neighbors of about three to five. Indeed, we observed that in this case, PCMCI is able to find the true parent set if the reaction from parent to child is faster than the other reactions. In Figure 4.10 we can the reason for this phenomenon. If we look at only a few neighbors, there might still be some local structure that cannot be seen with more neighbors (i.e. a larger hypercube) anymore.

In chemical reaction networks, a higher reaction rate means that there are more samples in the hypercube around a certain z. In our example of the chain $Z >> X >> Y$, we observe that if the reaction $Z >> X$ is much faster than the reaction $X >> Y$, there might not be any samples for X and Y given a certain value $Z = z$. In order to get a significant number of samples for X and Y, one has to choose a large hypercube. However, there might be no structure visible in this hypercube, so that PCMCI fails to find a causal relationship from X to Y.

There are also examples where even the small hypercube is too large and where it is not possible to see any local structure, see Figure 4.11 on the left-hand side. Here, the reaction rate of $Z >> X$ is much faster than that of $X >> Y$. On the right-hand side of Figure 4.11, we see the opposite case where $Z >> X$ is slower than $X >> Y$, so that we detect a local linear structure even in a larger hypercube.

The problem that PCMCI has with the conditional independence test CMIknn can be generalized to all methods, which are based on conditional independence testing, as they have to find (local) structures in the data, and do this by considering a certain neighborhood. The latter can be smaller or bigger, such that there is a varying number of other data points included. More data means less variance, but also increases the

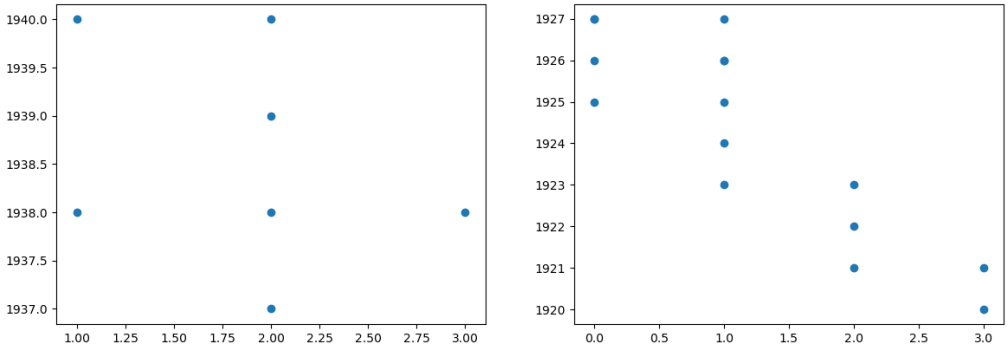


Figure 4.11: On the left-hand side there is the scatter plot of X (on x-axis) and Y (on y-axis) where the reaction rate of $Z >> X$ is faster than the one of $X >> Y$. On the right-hand side we have the opposite case. Again, the values of X and Y do not matter for our purpose.

probability that a structure which is visible in a smaller neighborhood cannot be found anymore.

4.6.5 Learnings from the Experiments

We can conclude that PCMCI is the best algorithm for discrete-time dynamical systems. Its non-parametric version can handle almost every input data of the functional form and finds even highly non-linear relations. Only in the case of data sampled from stochastic processes, e.g. the Gillespie algorithm, PCMCI had some trouble. We analyzed its behavior and concluded that conditional independence tests do not work in this setting, as long as the reaction considered is not much faster than the others.
Multivariate TE's output was mostly correct for functional data, but there were always some (minor) mistakes, making the algorithm less trust-worthy in real-world applications where no causal structure is known a priori.
CausalKinetiX worked reasonably well as long as the data was sampled from ODEs; it was not able to handle other kinds of data. A big drawback is that the algorithm did not work for multi-domain data, even though it was explicitly developed for this setting. The implementation of CCM is not usable on general datasets yet.

Chapter 5

Estimation of Causal Effects

Apart from finding causal relationships, the next big task of causal inference is to estimate the strength of causal effects. An example for this would be the *average treatment effect* (ATE), defined in Equation 5.2, which is usually used in clinical trials to rate the efficacy of a treatment.
In practice, this is done empirically via randomized controlled trials, i.e. by actually carrying out the experiments and comparing the results of the treatment groups with the control groups. In causal inference, the goal is to obtain the same knowledge, but with just using observational data, i.e. without carrying out any experiments. In the language of causal inference, the problem reduces to the estimation of the effects of interventions and all the theory that is discussed in this chapter tries to tackle it in different situations. Note that all causal relationships need to be known already, in order to estimate the strength of causal effects.

Estimating causal effects with observational data has been done in statistics long before Pearl formalized the framework of causality. The work that most of these methods can be related to is the theory of potential outcomes, pioneered by Neyman [SNDS90] in the 1920s and formalized by Rubin in the 1970s [Rub74]. Another useful tool is called instrumental variables [Wri28]. With the help of a confounding variable satisfying certain properties, it allows to estimate the causal effect from one variable to another.

The instrumental variables method is an example for approaches which assume the model to belong to a specific class. Only then one can estimate causal effects. If, on the other hand, no model structure can be assumed a priori, one has to identify the intervention. For an arbitrary model this is only possible with Pearl's do-calculus, see [Pea09] for the case of atomic interventions and [CB19] for stochastic interventions.

We will see that the methods in the i.i.d. case are quite advanced, while there is not much research for time series. However, there are attempts to make the transfer from the i.i.d. to the time series case, for example Blondel et al. [BAG17] who extended the theory of causal Bayesian networks to dynamical causal Bayesian networks and proved identifiability as well as transportability results (transportability is discussed in Section 6.3).

We will divide this chapter into three parts: Section 5.1 is about all methods and concepts that are related to Pearl's framework, in Section 5.2 the potential outcome framework of Rubin is introduced shortly, and in Section 5.3, the relevance of estimation of causal effects for dynamical systems is discussed.

The two frameworks of Pearl and Rubin are closely connected, as they both formalize (certain aspects of) causal inference. The potential outcome framework has been developed explicitly to model individual experiments. As the potential (i.e. possible) outcomes of one experiment cannot all happen at the same time (as one experiment can only have one outcome), they can also be seen as counterfactuals and can thus be included in Pearl's framework.

5.1 Interventions and Counterfactuals in Pearl's Framework

There are two important steps necessary to compute the effect of an intervention: identifying and estimating. Identification is especially important, as we want to do causal inference on arbitrary models, which are not covered by most of the other methods that compute causal effects. For arbitrary models it is not clear a priori whether a certain intervention can be computed or not. In Pearl's framework, the second step consists of estimating conditional expectations, while handling various pitfalls like unknown confounders.

The do-calculus provides a tool to either identify every intervention or state that it is not possible to do so. The ID algorithm has been developed for this task by Shpitser and Pearl [SP06]. The output of the algorithm is a more or less complex expression of conditional expectations, which then have to be estimated using a suitable technique. Linear regression is probably the most famous one, but there are many which can be used in the non-linear case as well (e.g. [VM19]).

Identifiability, which has been introduced in Section 2.1.4, has been extended in several ways. Bareinboim and Pearl introduced z-identifiability [BP12a], which extends the notion to experimental data. However, the assumption that all experiments are available makes it difficult to use, so that Lee et al. [LCB19] developed g-identifiability, which includes both observational and experimental settings and does not assume that all experiments are available. The original form of the ID algorithm is only able to treat atomic interventions, i.e. fixing a variable to a certain value, but Correa and Bareinboim recently proved completeness of their method for the identification of stochastic interventions [CB19].
Unfortunately, the extensions of identifiability are not widely implemented in software packages until now. For instance, Microsoft's causal inference library DoWhy [SK19] and others still only include implementations of the standard ID algorithm.

Interventions play an important role in counterfactual analysis as well. Let us consider a structural causal model $\mathcal{M} = (V, U, F)$ where U and V are two sets of variables and F is a set of functions that determine how values are assigned to each variable $V_i \in V$.

The assignments have the form

$$v_i = f_i(v, u),$$

where V_i is assigned the value v_i given the current values v and u of all variables in V and U. The variables in V are called endogenous, the variables in U are exogenous. Let $P(u)$ be the probability distribution of the exogenous variables U, which defines a distribution $P(v)$ on V as the variables in U uniquely determine the values of the variables in V. According to Pearl [Pea], the basic counterfactual entity in structural models is the sentence 'Y would be y had X been x in situation $U = u$', denoted $Y_x(u) = y$.
Let $\mathcal{M}_x$ be the SCM $\mathcal{M}$ where the assignment $X = x$ replaces the original one. We can formally define the counterfactual Y_x by

$$Y_x(u) := Y_{M_x}(u), \tag{5.1}$$

i.e. the value of Y in the modified SCM with the same values of the exogenous variables $U = u$.

$P(u)$ also induces a probability distribution on the counterfactual events $Y_x = y$ and thus on Boolean combinations of such events. We will use this in the following. Pearl introduced three steps to compute the probability $P(Y_x = y|e)$ for some propositional evidence e [Pea09]. The first step is called *abduction*. We update $P(u)$ to obtain $P(u|e)$. The second step, *action*, consists of replacing the equations determining the variables in set X by $X = x$. Finally, in the *prediction* step, we compute the probability of $Y = y$ in the modified model.
In order to clarify the steps of a counterfactual analysis, we give an example, following the one given in [Pea].

Example 5.1.1. Let X be the level of assistance given to a student, Z the amount of time the student spends studying, and Y the student's performance on an exam. We assume to have the following linear assignments.

$$\begin{aligned} X &:= \epsilon_1, \\ Z &:= 0.5X + \epsilon_2, \\ Y &:= 0.7X + 0.4Z + \epsilon_3. \end{aligned}$$

Assuming we measure $(X, Z, Y) = (0.5, 1, 1.5)$, we can estimate what would have happened if the student had doubled his or her study time. The student's characteristics $u = (\epsilon_1, \epsilon_2, \epsilon_3)$ stay the same. The first step is *abduction*, where we have to recompute the probability distribution given the observations. We compute $(\epsilon_1, \epsilon_2, \epsilon_3)$ by replacing X, Y, and Z with the measured values. We obtain

$$\begin{aligned} \epsilon_1 &= 0.5, \\ \epsilon_2 &= 1 - 0.5 \cdot 0.5 = 0.75, \\ \epsilon_3 &= 1.5 - 0.5 \cdot 0.7 - 1 \cdot 0.4 = 0.75. \end{aligned}$$

In the *action* step, we replace the equation of Z with $Z = 2$, so that all the links to the parent variables (in this case only the link to X) are removed. Now we can do the *prediction* that we wanted to do:

$$Y_{Z=2} = 0.5 \cdot 0.7 + 2 \cdot 0.4 + 0.75 = 1.9.$$

The result is that the student's score would have improved from 1.5 to 1.9 if he or she had doubled the amount of studying time.

There are various possible applications of counterfactuals, e.g. in reinforcement learning [BWZ+18], complex learning systems [BPQC+13], or Earth system science [HPO+16]. In the latter work, the authors used them in a detection and attribution task where weather and climate events are related to anthropogenic climate change, for example the European heat wave in the summer of 2003. Ness et al. built a method for performing counterfactual inference for Markov process models in equilibrium [NPV19]. Even though this is quite a special case, it is a first step towards using causal inference for continuous-time systems in practice. For this, probabilistic programming languages such as Pyro play an important role, as Ness et al. pointed out. They can be used not only for time series data, but also for i.i.d. data.

Even though instrumental variables are much older than Pearl's framework, they can be fitted into it. The setup of instrumental variables is the following. We want to estimate the effect of X on Y, but are not able to do so right away, as there are hidden common causes. Instead, it is possible to make use of a specific confounding variable to indirectly estimate the effect. This confounding variable Z, the instrumental variable, needs to fulfill three properties: (a) Z is independent of all common causes of X and Y, (b) Z is not independent of X, and (c) Z effects Y only through X. As an example, if we have the linear case

$$\begin{aligned} X &:= \beta Z + \gamma H + N_X, \\ Y &:= \alpha X + \delta H + N_Y = \alpha(\beta Z) + (\alpha\gamma + \delta)H + N_Y, \end{aligned}$$

where H is the common cause. Here, we can regress Y on βZ to get an unbiased estimator of Y. If one tried to regress directly on X, the result would be a biased one due to the lack of independence of X and H.

There are many extensions and generalizations of instrumental variables and the concept has been widely applied in practice [IA94, BT90, DMS10, GKTCS18]. Most of the times, instrumental variables are fitted into the potential outcome framework of Rubin, which is more common in practice than Pearl's graphical causality. However, there are a few publications connecting traditional methods of causal inference to the ones that were developed more recently. For example, Rothenhäusler et al. looked at instrumental variables from an invariance-based perspective of causality, which might give new insights in the future.

5.2 Potential Outcome Framework

The language of the potential outcome framework differs from Pearl's language, mainly because its central part are counterfactuals. Let n be the number of i.i.d. experiments

(or units on whom an experiment is carried out), T the treatment, and Y the response. The simplest case is a binary treatment (i.e. treatment or no treatment) which we will consider in this thesis, but there are several extensions [Rub04, Rub05, MW15, IR15].

We consider the potential outcomes of the i-th experiment, $Y(1)$ and $Y(0)$, indicating the reaction of the response on the treatment $T = 1$ and $T = 0$, respectively. The famous 'fundamental problem of causal inference' [Hol86] highlights the fact that we can only observe one of the outcomes per experiment, so that the average treatment effect (ATE)

$$ATE := \mathbb{E}[Y(1) - Y(0)], \tag{5.2}$$

is purely hypothetical for just one experiment. For example, if somebody has a headache and takes an aspirin at time t_1, then one can measure at time t_2 whether or not the headache is gone. If the person takes no aspirin, then one can also measure at time t_2 what happened. However, one cannot just try and take the aspirin at a later stage, e.g. if the headache is still there at time t_2, as this would be a different experiment.

Apart from the ATE, we can also look at the average treatment effect of the treated (ATT) and the untreated (ATU):

$$ATT := \mathbb{E}[Y(1) - Y(0) \mid T = 1] \tag{5.3}$$
$$ATU := \mathbb{E}[Y(1) - Y(0) \mid T = 0] \tag{5.4}$$

Note that $Y(0)$ given $T = 1$ is not a contradiction, as it is merely a potential outcome, i.e. a counterfactual. It is the response of the same unit in a counterfactual world where the unit has not been treated, even though conditioning on $T = 1$ means that has been treated in the real world.

Due to the fundamental problem of causal inference, we cannot properly estimate ATE with averaging over all units, as only one treatment can be carried out per unit. Hence,

$$\overline{ACE} := \frac{1}{n} \sum_{i=1}^{n} Y_i(1) - Y_i(0). \tag{5.5}$$

is not a good estimator. Instead, Neyman [SNDS90] and Rubin [Rub74] show that

$$\widehat{ATE} := \frac{1}{|S_1|} \sum_{i \in S_1} Y(1) - \frac{1}{|S_0|} \sum_{i \in S_0} Y(0) \tag{5.6}$$

is an unbiased estimator of Equation 5.2. Here S_j is the set of units on which treatment $T = j$, $j \in \{1, 2\}$, is carried out.

In reality, we will often have observational studies where the assignment to a group (treatment or control group) is not fully random, i.e. $(Y(0), Y(1)) \perp\!\!\!\perp T$ does not hold and $E(Y(1)|Z = 1) \neq E(Y(1))$, so that $\widehat{ATE}$ is not an unbiased estimator anymore.

In this case, the so-called *strongly ignorable treatment assignment* (SITA) assumption can help us [RR83b]. We assume, similar to instrumental variables, that a confounding variable with certain properties exists, which can then be conditioned on to get an unbiased estimator again. The values $P(Z = 1|X)$ are called propensity scores [RR83b]. They are used in different methods like matching, stratification, inverse propensity score weighting, or covariate adjustment [HIKS07, Ros87, LD04, Aus11]. There is still a lot of active research to find better methods for causal inference (e.g. [SLK18]), but also on applying the framework to actual problems, even in continuous time [ZS12].

5.3 Interventions on Dynamical Systems

In the i.i.d. setting, we want to estimate something like $\mathbb{E}[Y|do(X := x)]$ from observational data. It is not entirely clear how this expression can transferred to the time series case. It makes sense to differentiate between discrete-time and continuous-time dynamical systems.

In discrete time, we can try to estimate $\mathbb{E}[Y_t|do(X_{t-\tau} := x)]$. As long as we have reasonable stationarity assumptions, we are able to estimate how the intervention on one process in a certain time step affects another process in another time step. There is some research that explicitly tries to solve this task [Li18]. In contrast to the continuous-time case, the discrete-time case can still profit from Pearl's framework, as we have already seen in Chapter 4 with PCMCI algorithm. The do-calculus can be easily used and the identifiability statements hold.

Assuming now to have a continuous-time system, the situation is very different. The expression $\mathbb{E}[Y_t|do(X_{t-\tau} := x)]$ completely loses its power, as the number of time steps is uncountable, so that probability measures give zero measure to events like $P(Y_t = y)$. Therefore, it makes more sense to look at the dynamics of the process after intervening.
However, these are only described by physical models, so that causal inference lacks the tools for this task. As observational data is not enough to compute interventions, one has to perform real experiments to obtain causal knowledge. Additionally, the lack of identifiability results (or something similar adapted for continuous-time dynamical systems) leaves us with no possibility of using observational data for estimating causal effects.

Even though the case of continuous-time dynamical systems is especially difficult to treat, there is some research that tries to connect causal and physical models and can serve as a foundation for the next steps, e.g. identifiability for continuous-time systems. For example, interventions that are defined for structural causal models can used for differential equations as well, as they are just any form of replacing parts of an equation by something else. For example, to fix the variable X to a certain value, we can require that $X(0) = \xi$ and $\dot{X} = 0$.
However, as physical models contain strictly more information than causal models, we can only build that connection for strictly stationary or converged systems. This has been one for ordinary differential equations [MJS13] as well as for random differen-

tial equations [BM18], which have random initial points and deterministic dynamics. Sokol and Hansen [SH13] give a causal interpretation of stochastic differential equations without further relating the concepts to causal models.

To summarize, estimating causal effects is one of the oldest tasks of causal inference and there has been much progress, mainly for i.i.d. data. Even though the concepts of the i.i.d. case can be transferred quite easily to discrete-time dynamical systems, they do not have the same power there, as they cannot provide information about the change of dynamics of the whole time series.
In the continuous-time case, we do not even have the theoretical foundation anymore and there is no way of estimating the effect of interventions or experiments without actually carrying them out. Therefore, the number of possible applications for time series data is limited.

Chapter 6

Causality and Machine Learning

Causality has always been a concept closely connected to data science and recently also to artificial intelligence and machine learning. Schölkopf and others even argue that it is necessary to incorporate causal knowledge into machine learning models and self-learning algorithms in order to get to the next step of artificial intelligence [Sch19, Pea19]. There are various ways in which causality is of importantance in machine learning research and applications; numerous publications show that researchers are aware of that [GCL+18].

In this chapter, we want to cluster the publications into different groups that facilitate the task of analyzing the connection of causality and machine learning. In Section 6.1, we discuss how causality can make machine learning models easier explainable and interpretable. In Section 6.2, we wrap up some approaches that use causal relationships, as well as situations where machine learning can profit from knowing the causal structure of data. We continue in Section 6.3 with evaluating how interventional knowledge can be incorporated in data science; and finish with an analysis how the most recent notion of causality, independence-based causality, can be applied in practice.

6.1 Explainability

One central goal of machine learning is to make 'black box models'[1] more interpretable or explainable [RSG16a, RSG16b, KW17]. The link between explainability and causality is inherent in the way humans understand the world. We can observe causality in nature, e.g. in physical or mechanical systems, and thus use it to better understand a model.

For example, if we have a machine learning model that predicts Y from a multivariate X, then it is a priori unclear whether every X_i that Y depends on is actually a causal parent. Causal inference is able to answer this question and to check whether the model uses other variables than the causal parents to predict Y. If so, one can retrain the models only using the causal parents to get a more robust and easier interpretable machine learning model.

[1] Thus named for their metaphorical inability to observe their inner workings.

Reimers et al. developed an algorithm that decides whether a feature is useful globally or only locally by using causal inference methods [RRD19]. They argue that another problem of standard algorithms in this field is that they work only locally for specific inputs. Here, the invariance of causal mechanisms can help to get more relevant, global information.

Apart from direct applications, there has been a lot of work on connecting causality and explainability on a theoretical level. Kim and Bastani [KB19] built a framework for causal interpretable models, i.e. models that are able to learn causal information from observational data while still being interpretable. This led to the term *causability*, introduced by Holzinger et al. [HLD+19], which goes side by side with explainability. It highlights the importance of recognizing human perception as an important parameter for explainability. In their recently published survey of explainability and interpretability, Roscher et al. [RBDG19] stressed the role of causality for real progress in the field of explainable machine learning. Zhao and Hastie published a paper where they connected, on a theoretic level, the back-door criterion of causality with partial dependence plots and more advanced techniques which are used in interpretable machine learning [ZH19].

6.2 Using Causal Structure

Causality is also able to make machine learning models better in different ways via exploiting causal relationships. For example, if the causal structure of data is known, only causal parents should be used for predicting a certain variable Y, as this leads to more robust and explainable methods.

Not only the relationship between observed variables can be important, but also relations between observed and unobserved ones. One of the main challenges for obtaining good datasets is selection bias, which occurs if there is an unobserved confounder that influences the data-generating process [Pea09]. For example, if one wants to conduct an empirical study in political or social sciences, there are many factors that potentially lead to selection bias [BP12b]. If people are randomly chosen and questioned in front of a university building, it will be likely that most of the participants are students and not representative of the whole society. Similarly, visiting households in the early afternoon probably leads to many closed doors, as anyone with a 9-to-5 job will not be at home. The framework of causality is able to properly address these challenges, as we have seen in Example 2.1.10, where we discussed Berkson's paradox.

Khajehnejad et al. [KTS+19] exploit causal knowledge for better decision-making. A bank may use machine learning to calculate the interest rate offered to a customer using their financial situation and other parameters which, in the past, allowed for better prediction of the credit-worthiness. This is not *per se* a good or bad thing. However, if these parameters are transparent, then one can actively intervene to get a better score. Previous debts returned in full are a reasonable predictor for solvency, which the bank could reward with a lower interest rate. Other parameters like the home address, however, do not have any causal effect on credit-worthiness. As moving to a different area does not make it any more likely to repay a loan, it shouldn't lead

to a better credit assessment.
It is likely that as soon as the parameters used to estimate credit-worthiness become transparent, people actively start trying to manipulate their scores. We enter an interventional setting where causal relations become important and correlations lose their power.

This is what Schölkopf calls 'realm of causality' [Sch19]. He gave the very intuitive example of buying a laptop case in an online shop. Due to the obvious correlation of people buying laptops and laptop cases, the algorithm might suggest you buy a laptop along with the case. However, as few people would get a laptop case before buying a laptop, this recommendation clearly is nonsense. On the other hand, suggesting to somebody who purchased a laptop also to buy a fitting case actually makes sense. The algorithm's failure is caused by an intervention in the system, which occurs with the (active) suggestion of an additional item.
Thus, purely statistical data cannot be used any more and a causal model is needed. This causal model would show that there is a causal relationship from buying a laptop to buying a laptop case, but not the other way around. Again, the machine learning model operates in the 'realm of causality' without acknowledging the fact.

A field of research where causal knowledge can be used is anomaly detection. Qiu et al. lay out different (standard) approaches used for this task which do not relate to causal inference [QLSL12]. One option is to use machine learning to predict the future of a time series using its past values. An anomaly is reported if the actual value differs significantly (for some significance level α) from the predicted value [CBK09]. Another possibility is to use clustering techniques [YT02]. These methods, however, yield no information about the source of the anomaly, which is just as important as detecting it in the first place. Causality can help with so-called dependency anomalies, i.e. anomalies that occur due to a change of temporal dependencies. One can use any algorithm that uncovers the causal structure in data and apply the scores of Qiu et al. to find out if there are anomalies somewhere in the given dataset.

Another field where investigating relations between variables is most crucial is neuroscience. Seth et al. [SBB15] give an overview of the importance of Granger causality in neuroscience. Michalareas et al. use partial directed coherence to measure causal relationships in MEG sensor measurements [MSPG13]. One of the difficulties inherent to neuroscience is that many regions of the brain are not yet fully understood, so that spurious correlations and confounding variables can mislead researchers; as such it is obvious that using causal inference methods instead of the standard statistical tools is necessary.

Zhou et al. enhanced standard Granger causality to apply it on non-stationary dynamical systems using modified Hodrick-Prescott filters to extract trend components [ZKZS13]. They use it on smart buildings to understand relationships between the sensors and allow for better predictions of energy usage and other applications. A practical consequence is that in some cases, expensive-to-maintain sensors can be made redundant, if their readings are found to be completely determined by cheaper ones.

6.3 Interventional Knowledge

One field of research, where one (implicitly) calculates the effect of interventions is reinforcement learning. However, even though interventions were used in the past, the setting has not been considered as a multi-dimensional one. Strictly speaking, every action of the agent is a violation of the common i.i.d. assumption of machine learning [Sch19] and a lot of causal and temporal information is created by these interventions. Currently, however, this information is not used and researchers instead try to get rid of it by permuting past data, which has, for example, been done with DeepQ [MKS+15, Sch19].

The ultimate goal of reinforcement learning is to imitate the learning of a human or animal. However, the latter learn by interacting with their environment. Schölkopf [Sch19] gives the example of an Atari game, which becomes easier for algorithms when the resolution is downsampled. For humans, this downsampled version proves to be much harder, as the characters are not clearly visible any more. Any child can easily play around with the joystick to find out which game character it plays and how the characters interact with each other. For a machine, this is much more difficult, as the concept of learning by intervention and transformation is not currently used [Mac71, Sch19]. For some of the various applications of causality in reinforcement learning, see [BWZ+18, LSHL18, DWC+19, ZB19].

The goal of *transportability*, a concept which has been developed by Bareinboim and Pearl [PB11, BP13, BP14, PB14], is to transfer interventional information from one domain to another. For example, let us imagine that a study has been carried out in New York, where the causal effect of an exposure X on outcome Y is estimated for every age group $Z = z$. Now, we want to use this knowledge to get insights on the effect on people living in Los Angeles. However, the distribution $P^*(X, Y, Z)$ of Los Angeles differs significantly from the distribution $P(X, Y, Z)$ in New York, due to factors like e.g. age structure. Bareinboim and Pearl suggest the transport formula

$$P^*(y \mid do(x)) = \sum_z P(y \mid do(x), z) P^*(z), \tag{6.1}$$

which can be obtained from the invariance principle of causal inference. One assumes that causal effects do not depend on specific environments, so that $P^*(y|do(x)) = P(y|do(x))$ holds. In [BP13] they give a complete algorithm that identifies effects that can be transported using the rules of do-calculus.

One problem encountered by every data scientist, is that one dataset can often stem from various, possibly very different sources. For example, consider a study that tries to analyze the maths results of all students in high school. The dataset comes from many schools and is very likely pooled together. At this point structural differences between the schools are ignored. Conceivably, one of the schools could have offered two additional lessons a week on mathematics, while most of the students at another may have practiced at home for a maths competition. In a third school, all the maths teachers might have been seriously sick, leaving the students with a biology teacher as replacement for most of the year, reducing the quality of teaching. If we were trying to estimate the effectiveness of an intervention, such as the introduction of

smart blackboards on pupils, the results would ignore many potential confounders in the data.
This is called the *data fusion problem* and has been addressed by Bareinboim and Pearl [BP16] in a general framework, including interventional and observational data with selection bias among others. Hünermund and Bareinboim focused specifically on data fusion in econometrics [HB19].

Little and Badawy [LB19] recently developed a causal version of the statistical bootstrap resampling method [ET93]. The goal is to train models with bootstrapping, while trying to get knowledge of interventions and the causal structure, which is then in turn used to improve the results of the algorithm.
This procedure is more robust than standard prediction tools, e.g. in the presence of common drivers of X and Y. Little and Badawy give the example of the MNIST dataset where a confounding variable has been introduced: the brightness of the background of the pictures. It is highly correlated with the numbers (i.e. the variable Y), such that standard machine learning algorithms tend to learn the brightness and use it as a predictor for the numbers. Only causal knowledge of the three variables pixels X, brightness U, and numbers Y can detect the spurious correlation between U and Y, and thus build a more robust model.

6.4 Independent Mechanisms

Invariance-based causality [PBM16, HDMP18, PBP19] led to new domain adaptation methods, complementary to the ones developed by Bareinboim and Pearl [SJPZ11, ZSMW13, ZGS15, GZL$^+$16, HDM17, MvOC$^+$18, RCSTP18]. The goal is to close the gap between invariance, robustness, and causality [RMBP18].

A major challenge of machine learning lies in transferring conclusions from the training data to a larger dataset (data generalization), or from one dataset to a (slightly) different one (transfer learning). Causality plays a major role in these tasks, as causal mechanisms are assumed to be invariant, allowing both for generalization and transfer of knowledge [RCSTP18]. At this point, we leave the graphical causality behind, as causal mechanisms that are invariant across domains do not necessarily have to be a relationship between two variables.

Arjovsky et al. [ABGLP19] exploit invariance of causal mechanisms for prediction tasks. The example they give is that of cow pictures with predominantly green background and camel pictures usually taken in the desert. Clearly, the background should be insignificant to distinguish the two animals, but the algorithm is quite likely to overestimate the importance of the background and would classify a cow on a beach as camel.
This can be avoided by considering different environments where the percentage of pictures with cows on green pastures and other backgrounds varies. The same is done for the camel pictures. If the algorithm now includes the background as relevant information for the classification, the results will be that classification scores differ across the environments. If environments with a lot of cows on green pasture lead to better results than others, one can infer that the background factored into the

classification. Arjovsky et al. use this fact when they say that mechanisms are only relevant if they are independent across all environments. Here, this would clearly not be the case. Indeed, no model that includes the background can get invariant results if the environments are chosen properly.

The example shows that there is a huge potential in exploiting the hypothesis of independent mechanisms. Another field where it led to a much better understanding of many machine learning models is semi-supervised learning. The idea of semi-supervised learning is to exploit Bayes' rule

$$P(Y \mid X) = \frac{P(X, Y)}{P(X)}, \tag{6.2}$$

which indicates that additional information about $P(X)$ can help in predicting $P(Y|X)$. However, the independent mechansism hypothesis tells us that $P(Y|X)$ is invariant if there is a causal relationship from X to Y. Thus, looking at the larger dataset with unlabeled samples of X does not change the prediction power. In the anti-causal direction where we use Y to predict X, on the other hand, the additional information does help [Sch19]. According to Schölkopf, this result can explain why many data scientists are not able to improve the power of their models with additional unlabeled data.

Chapter 7

Conclusion and Outlook

We talked about four different approaches to causality: Pearl's graphical causality, invariance-based causality, Wiener-Granger causality, and topological causality. The first is not originally intended, but suitable for discrete-time dynamical systems. As it needs to treat every time step of the dynamical system as random variable, it does not qualify for continuous-time dynamical systems. Unsurprisingly, the results of PCMCI, the algorithm that is based on this approach, are not reliable for ODE-based data. For discrete-time dynamical systems on the other hand, they are the most promising.

There are no implementations yet of algorithms using the invariance-based approach for dynamical systems. Li [Li18] developed an algorithm on a theoretical level, but there is no publicly available implementation at the time of writing. As there is no reason why invariance-based causality should not perform well for time series data, one can hope for promising algorithms to be developed in the upcoming years.

Wiener-Granger causality has been applied widely already in the easy case of linear relationships in the data. It is closely connected to the methods that use information theory to test for causal relationships. We used the implementation of multivariate transfer entropy of the Python package IDTxl, which can handle non-parametric and multivariate dynamical systems, finding that the greedy approach of the implementation did not work as well as PCMCI. Just as the algorithms based on Pearl's approach mainly rely on conditional independence testing, multivariate TE relies on testing for conditional mutual information. Hence, the easiest way to improve the algorithms is to improve the respective statistical tests.

The algorithms based on topological causality are able to treat the case of deterministic systems that are non-separable. Separability is one of the most important assumptions that is needed for Granger causality and also a huge drawback from the theoretical perspective, as it is unlikely to hold in practice [SMY+12, HLSP17]. In our experiments, convergent cross mapping did not work properly for most of the data, but there are some extensions of CCM and other, similar, algorithms have been developed recently, so further progress may yet be achieved in the future. An important critique to bear in mind, however, is that of Stark et al. [SBDH97], pointing out that systems in the real world do not function in a purely deterministic way, leading to problems with the traditional theory of chaotic dynamical systems. Another approach that connects the ideas of Granger and topological causality is needed, which works for random dynamical systems that are non-separable.

	Continuous-time dynamical systems	Discrete-time dynamical systems	i.i.d. data
Observational data - deterministic	Convergent cross-mapping?	Convergent cross-mapping	
Observational data - stochastic	CausalKinetiX	PCMCI, multivariate TE	FCI,GES, etc.
Observational and experimental data	CausalKinetiX		NonlinearICP

Table 7.1: This table shows for what kind of data which algorithm can be used.

A singular perfect solution exists neither for finding causal relationships, the case we just mentioned, nor for other tasks, like estimating the effect of causal interventions. However, there are many situations, in which it is very helpful or necessary to include causal knowledge. Data is often sampled from different domains and interventional/experimental data is not treated accordingly. Mostly, it is simply discarded, making the task of learning good models harder than it should be. The goal is to not suffer from heterogeneous data but to embrace it as a good opportunity for learning more robust models.

The most important step is not to forget that causality plays a huge role in many applications. Especially in today's machine learning community, researchers commonly attempt to find the best models without being able to explain what the algorithm does and why it says that certain variables are important for the prediction. If there is no causal relationship from X_i to Y, then a robust model should not use this source variable. On the training data, and even on similar test data, the prediction results appear worse at first, but the algorithm can generalize better to other datasets.

Causality is so important for humans and the way we learn and see things, that one often does not see the immediate connection of the given (data science) task and causality. For example, many dynamical systems are given by differential equations which explicitly specify the causal structure of the data. Hence, causal models are not the solution for everything. Physical or mechanical models often contain even more information about the causal structure and the dynamics of the system. Causal models can often be learned from data, which is an advantage over physical models, but their main aim is to improve statistical ones.
Therefore, if one has a dataset of samples of a dynamical system and ODEs that describe this system well, then the maximum of causal knowledge is reached already. In this case, one could only find out whether the dataset is really from i.i.d. data or in fact a collection of data from different environments. In the latter case, causal inference has the tools to include the additional knowledge in a sensible way.

Even though Pearl's approach has been criticized for not being applicable in prac-

tice, it provides, at least, one tool that is very useful and generally applicable: the do-calculus. It gives a formal language that includes interventional and observational data. As such, it is the foundation of the next level of data science, which is going to be able to handle and profit from experimental data.

Even though causality slowly enters in the mainstream of data science, there is still a lack of publications that apply causal inference methods to real-world problems, especially in the case of dynamical systems. Granger causality in its linear form is the only method that has been widely adopted. However, at least in theory, there are far more advanced approaches. In the following, we want to outline a few situations where we think that causality can bring significant progress. The focus is on the adoption of invariance-based methods for time series, as both graphical and Wiener-Granger causality are not really able to handle continuous-time dynamical systems, and this will continue to be a big problem of causal inference.

First, we consider the process of estimating a differential equation from data. Even though there is a natural way of defining interventions on dynamical systems and there is extensive theory about controlling mechanical systems, researchers might not always use all the experimental knowledge they have. In order to get to more robust models, one can carry out different experiments on the system, collect the data, and see whether the ODE is able to describe the experiments well across all environments. If not, the chosen model (given by a collection of ODEs) might not be sufficiently robust.

The second field, where we think that causality will play a major role in the future, is anomaly detection, or, more general, the detection of (physical) mechanisms, which are not explained by available ODE models. We can use the ideas of Qiu et al. [QLSL12] and connect them to recent developments in causal inference.
If one is able to collect data from different environments, say of wind turbines in different countries/regions/places where the same anomaly is observed, then one could explicitly focus on it and try to find the source processes $\mathbf{X}$, so that $P(Y|\mathbf{X})$ is invariant across all environments. Here Y denotes the anomaly. If there is such an $\mathbf{X}$, then there is a causal relationship (a causal mechanism) according to the principle of independent mechanisms. This approach is not only able to detect and predict anomalies, but also to fully explain why they happen.

Even though significant progress has been made in predictive maintenance, there is still room for further improvements. We can think of a machine that is used in very different environments, e.g. on oil platforms and in the car industry. Currently, data scientists try to collect all the data they can get from the machine and pool them together, in order to get a dataset which is as large as possible. However, in the pooling procedure really important information gets lost. Assuming that there are similar circumstances in the different environments that lead to a defect of the machine, we might not be able to detect them after two or more different probability distributions are averaged out.

One could imagine a situation where one component of machine A starts to malfunction. This influences a neighboring machine B, which in turn starts to show a

different behavior. This has an impact on some of the sensors of machine A. These sensors might be seen as a predictor for the malfunction of machine A, even though their change is merely caused by machine B.
If there is always a machine B next to our machine A, then, while not being robust, the model will at least work well. Otherwise, the predictions will not be satisfactory and considering another environment may help. We assume that in the second environment, there are no neighboring machines influencing the sensors, such that it is obvious that they cannot be good predictors.
However, this knowledge can only be used if one treats the two environments separately. If one pools all the data together, the algorithm might learn to use the sensors as a predictor, even though the predictions are poor on some of the data. A more robust algorithm can only be obtained by treating the data as heterogeneous.

To give another data-driven field which can profit from causality, we consider Earth system processes that are still far from being understood. For them, it is very important to avoid spurious correlations and instead find actual causal relationships, as otherwise no real knowledge is gained. The modern challenge of climate change makes it more and more important to understand how the climate works and how different meteorological processes relate to each other, but hardly any experiments can be carried out because they are too expensive, infeasible, or too dangerous. Even though there has been progress, which led, for example, to the algorithm PCMCI [RNK$^+$19, RBB$^+$19], it is important to also include the notion of invariance-based causality, in order to properly address data which is often sampled from very different environments.

The author believes that the biggest progress lies in developing algorithms which use invariance-based causality. Not only since it is the newest (and as of yet least developed) approach, but also because it is able to include all the data, both observational and experimental, to create better models. The goal is to create models which learn in a way that is more human than any other AI algorithm, as they actively use knowledge that has been created by transforming and experimenting, just as humans and animals do.

Bibliography

[ABGLP19] Martín Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *ArXiv*, abs/1907.02893, 2019.

[AEI18] Susan Athey, Dean Eckles, and Guido W. Imbens. Exact p-values for network interference. *Journal of the American Statistical Association*, 113(521):230–240, 2018.

[ARG+16] OO Aalen, K Røysland, JM Gran, R Kouyos, and T Lange. Can we believe the dags? a comment on the relationship between causal dags and mechanisms. *Statistical Methods in Medical Research*, 25(5):2294–2314, 2016.

[AS17] Peter M. Aronow and Cyrus Samii. Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics*, 11(4):1912–1947, 2017.

[Aus11] Peter C. Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46(3):399–424, 2011.

[BAG17] Gilles Blondel, Marta Arias, and Ricard Gavaldà. Identifiability and transportability in dynamic causal networks. *International Journal of Data Science and Analytics*, 3(2):131–147, 2017.

[Ber46] Joseph Berkson. Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*, 2(3):47–53, 1946.

[BFT19] G W Basse, A Feller, and P Toulis. Randomization tests of causal effects under interference. *Biometrika*, 106(2):487–494, 2019.

[BM18] Stephan Bongers and Joris M. Mooij. From random differential equations to structural causal models: the stochastic case. *ArXiv*, abs/1803.08784, 2018.

[Bol89] Kenneth A. Bollen. *Structural Equations with Latent Variables*. Wiley series in probability and mathematical statistics. Applied probability and statistics. Wiley, 1989.

[Bor20] Enrico Borba. Python crn, 2020. `https://github.com/enricozb/python-crn`, Last accessed on 2020-03-12.

[BP12a] Elias Bareinboim and Judea Pearl. Causal inference by surrogate experiments: Z-identifiability. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, UAI'12, page 113–120. AUAI Press, 2012.

[BP12b] Elias Bareinboim and Judea Pearl. Controlling selection bias in causal inference. In Neil D. Lawrence and Mark Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 100–108. PMLR, 2012.

[BP13] Elias Bareinboim and Judea Pearl. A general algorithm for deciding transportability of experimental results. *Journal of Causal Inference*, 1(1):107–134, 2013.

[BP14] Elias Bareinboim and Judea Pearl. Transportability from multiple environments with limited experiments: Completeness results. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 280–288. Curran Associates, Inc., 2014.

[BP16] Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

[BPQC+13] Léon Bottou, Jonas Peters, Joaquin Quinonero-Candela, Denis X. Charles, D. Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14:3207–3260, 2013.

[BS16] Anjali Raja Beharelle and Steven L. Small. Imaging brain networks for language: Methodology and examples from the neurobiology of reading. In Gregory Hickok and Steven L. Small, editors, *Neurobiology of Language*, chapter 64, pages 805 – 814. Academic Press, 2016.

[BS17] Lionel Barnett and Anil K. Seth. Detectability of granger causality for subsampled continuous-time neurophysiological processes. *Journal of Neuroscience Methods*, 275:93 – 121, 2017.

[BT90] Roger J. Bowden and Darrell A. Turkington. *Instrumental Variables*. Cambridge University Press, 1990.

[BWZ+18] Lars Buesing, Théophane Weber, Yori Zwols, Sébastien Racanière, Arthur Guez, Jean-Baptiste Lespiau, and Nicolas Manfred Otto Heess. Woulda, coulda, shoulda: Counterfactually-guided policy search. *ArXiv*, abs/1811.06272, 2018.

[BZDP20] Simon Behrendt, David Zimmermann, Thomas Dimpfl, and Franziska Peter. Rtransferentropy: Measuring information flow between time series with shannon and renyi transfer entropy, 2020. `https://CRAN.`

R-project.org/package=RTransferEntropy, Last accessed on 2020-03-12.

[CB19] Juan D. Correa and Elias Bareinboim. From statistical transportability to estimating the effect of stochastic interventions. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 1661–1667. International Joint Conferences on Artificial Intelligence Organization, 2019.

[CBK09] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3), 2009.

[CGS15] Bree Cummins, Tomáš Gedeon, and Kelly Spendlove. On the efficacy of state space reconstruction methods in determining causality. *SIAM Journal on Applied Dynamical Systems*, 14(1):335–381, 2015.

[CH02] Stephen Cole and Miguel Hernán. Fallibility in estimating direct effects. *International journal of epidemiology*, 31:163–5, 2002.

[Chi02] David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002.

[CL55] Earl A. Coddington and Norman Levinson. *Theory of Ordinary Differential Equations*. McGraw-Hill, 1955.

[CM14] Diego Colombo and Marloes H. Maathuis. Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.*, 15(1):3741–3782, 2014.

[CMH13] Tom Claassen, Joris M. Mooij, and Tom Heskes. Learning sparse causal models is not np-hard. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, UAI'13, pages 172–181. AUAI Press, 2013.

[CMKR12] Diego Colombo, Marloes H. Maathuis, Markus Kalisch, and Thomas S. Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *Ann. Statist.*, 40(1):294–321, 2012.

[CT06] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.

[CWPW86] C R Charig, D R Webb, S R Payne, and J E Wickham. Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy. *BMJ*, 292(6524):879–882, 1986.

[Daw10] A. Philip Dawid. Beware of the dag! In Isabelle Guyon, Dominik Janzing, and Bernhard Schölkopf, editors, *Proceedings of Workshop on Causality: Objectives and Assessment at NIPS 2008*, volume 6 of *Proceedings of Machine Learning Research*, pages 59–86. PMLR, 2010.

[Did08] Vanessa Didelez. Graphical models for marked point processes based on local independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):245–264, 2008.

[DMS10] Vanessa Didelez, Sha Meng, and Nuala A. Sheehan. Assumptions of iv methods for observational epidemiology. *Statistical Science*, 25(1):22–40, 2010.

[DWC+19] Ishita Dasgupta, Jane Wang, Silvia Chiappa, Jovana Mitrovic, Pedro Ortega, David Raposo, Edward Hughes, Peter Battaglia, Matthew Botvinick, and Zeb Kurth-Nelson. Causal reasoning from meta-reinforcement learning. *ArXiv*, abs/1901.08162, 2019.

[ED10] Michael Eichler and Vanessa Didelez. On granger causality and the effect of interventions in time series. *Lifetime Data Analysis*, 16:3–32, 2010.

[Eic12] Michael Eichler. *Causal Inference in Time Series Analysis*, chapter 22, pages 327–354. John Wiley and Sons, Ltd, 2012.

[Eic13] Michael Eichler. Causal inference with multiple time series: principles and problems. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, 371(1997):1–17, 2013.

[Ell15] Jordan Ellenberg. *How Not to Be Wrong: The Power of Mathematical Thinking*. Penguin Press, 2015.

[ET93] Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Number 57 in Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, 1993.

[FBO+14] Karl J. Friston, André M. Bastos, Ashwini Oswal, Bernadette van Wijk, Craig Richter, and Vladimir Litvak. Granger causality revisited. *NeuroImage*, 101:796 – 808, 2014.

[Fis15] R. A. Fisher. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4):507–521, 1915.

[FKS+06] Kaye Middleton Fillmore, William C. Kerr, Tim Stockwell, Tanya Chikritzhs, and Alan Bostrom. Moderate alcohol use and reduced mortality risk: Systematic error in prospective studies. *Addiction Research & Theory*, 14(2):101–132, 2006.

[Fre83] John R. Freeman. Granger causality and the times series analysis of political relationships. *American Journal of Political Science*, 27(2):327–358, 1983.

[GCL+18] Ruocheng Guo, Lu Cheng, Jundong Li, P. Richard Hahn, and Huan Liu. A survey of learning causality with data: Problems and methods. *ArXiv*, abs/1809.09337, 2018.

[Gil76] Daniel T Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4):403 – 434, 1976.

[Gil77] Daniel T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, 1977.

[GKTCS18] Zijian Guo, Hyunseung Kang, T. Tony Cai, and Dylan S. Small. Confidence intervals for causal effects with invalid instruments by using two-stage hard thresholding with voting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):793–815, 2018.

[Gra69] Clive W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969.

[Gre10] Sander Greenland. Overthrowing the tyranny of null hypotheses hidden in causal diagrams. In R. Dechter, H. Geffner, and J.Y. Halpern, editors, *Heuristics, Probability and Causality: A Tribute to Judea Pearl*, pages 365–392. College Publications, 2010.

[GZL+16] Mingming Gong, Kun Zhang, Tongliang Liu, Dacheng Tao, Clark Glymour, and Bernhard Schölkopf. Domain adaptation with conditional transferable components. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2839–2848. PMLR, 2016.

[GZS+15] Mingming Gong, Kun Zhang, Bernhard Schoelkopf, Dacheng Tao, and Philipp Geiger. Discovering temporal causal relations from subsampled data. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1898–1906. PMLR, 07–09 Jul 2015.

[Had02] Jacques Hadamard. Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton University Bulletin*, pages 49–52, 1902.

[HB19] Paul Hünermund and Elias Bareinboim. Causal inference and data-fusion in econometrics. *ArXiv*, abs/1912.09104, 2019.

[HDM17] Christina Heinze-Deml and Nicolai Meinshausen. Conditional variance penalties and domain shift robustness. *ArXiv*, abs/1710.11469, 2017.

[HDMP18] Christina Heinze-Deml, Nicolai Meinshausen, and Jonas Peters. Invariant Causal Prediction for Nonlinear Models. *Journal of Causal Inference*, 6(2):1–35, 2018.

[HGC94] David Heckerman, Dan Geiger, and David Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Mach Learning*, 20:293–301, 1994.

[HH08] Michael G Hudgens and M. Elizabeth Halloran. Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842, 2008.

[HIKS07] Daniel Ho, Kosuke Imai, Gary King, and Elizabeth Stuart. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15:199–236, 2007.

[HLD+19] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. Causability and explainability of artificial intelligence in medicine. *WIREs Data Mining and Knowledge Discovery*, 9(4), 2019.

[HLSP17] Daniel Harnack, Erik Laminski, Maik Schünemann, and Klaus Richard Pawelzik. Topological causality in dynamical systems. *Phys. Rev. Lett.*, 119(9), 2017.

[HMC99] David Heckerman, Christopher Meek, and Gregory F. Cooper. A Bayesian Approach to Causal Discovery. In Gregory F. Cooper and Clark Glymour, editors, *Computation, Causation, and Discovery*, pages 141–165. The MIT Press, 1999.

[Hol86] Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

[HPJ+16] Antti Hyttinen, Sergey Plis, Matti Järvisalo, Frederick Eberhardt, and David Danks. Causal discovery from subsampled time series data by constraint optimization. In Alessandro Antonucci, Giorgio Corani, and Cassio Polpo Campos, editors, *Proceedings of the Eighth International Conference on Probabilistic Graphical Models*, pages 216–227, 2016.

[HPO+16] A. Hannart, J. Pearl, F. E. L. Otto, P. Naveau, and M. Ghil. Causal counterfactual theory for the attribution of weather and climate-related events. *Bulletin of the American Meteorological Society*, 97(1):99–110, 2016.

[HSPVB07] Katerina Hlaváčková-Schindler, Milan Paluš, Martin Vejmelka, and Joydeep Bhattacharya. Causality detection based on information-theoretic approaches in time series analysis. *Physics Reports*, 441(1):1 – 46, 2007.

[Hua10] Tzee-Ming Huang. Testing conditional independence using maximal nonlinear conditional correlation. *The Annals of Statistics*, 38(4):2047–2091, 2010.

[HV06] Yimin Huang and Marco Valtorta. Pearl's calculus of intervention is complete. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'06, pages 217–224. AUAI Press, 2006.

[HZL+18] Biwei Huang, Kun Zhang, Yizhu Lin, Bernhard Schölkopf, and Clark Glymour. Generalized score functions for causal discovery. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '18, page 1551–1560. Association for Computing Machinery, 2018.

[IA94] Guido W. Imbens and Joshua D. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994.

[Imb03] Guido W. Imbens. Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, 93(2):126–132, 2003.

[Imb19] Guido Imbens. Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. Working Paper 26104, National Bureau of Economic Research, 2019.

[IR15] Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press, 2015.

[JBC16] Ryan G. James, Nix Barnett, and James P. Crutchfield. Information flows? a critique of transfer entropies. *Phys. Rev. Lett.*, 116(23), 2016.

[JBGWS13] Dominik Janzing, David Balduzzi, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Quantifying causal influences. *Ann. Statist.*, 41(5):2324–2358, 2013.

[JV18] Mohammad Ali Javidian and Marco Valtorta. A proof of the front-door adjustment formula. *CoRR*, abs/1806.10449, 2018.

[KB19] Carolyn Kim and Osbert Bastani. Learning interpretable models with causal guarantees. *ArXiv*, abs/1901.08576, 2019.

[KHM+20] Markus Kalisch, Alain Hauser, Martin Maechler, Diego Colombo, Doris Entner, Patrik Hoyer, Antti Hyttinen, Jonas Peters, Nicoletta Andri, Emilija Perkovic, Preetam Nandy, Philipp Ruetimann, Daniel Stekhoven, Manuel Schuerch, and Marco Eigenmann. pcalg: Methods for graphical models and causal inference, 2020. `https://CRAN.R-project.org/package=pcalg`, Last accessed on 2020-03-12.

[KSG04] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Phys. Rev. E*, 69(6), 2004.

[KTS+19] Moein Khajehnejad, Behzad Tabibian, Bernhard Schölkopf, Adish Singla, and Manuel Gomez-Rodriguez. Optimal decision making under strategic behavior. *ArXiv*, abs/1905.09239, 2019.

[KW17] Sanjay Krishnan and Eugene Wu. Palm: Machine learning explanations for iterative debugging. In *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics*, number 4 in HILDA'17. Association for Computing Machinery, 2017.

[Lau96] S. L. Lauritzen. *Graphical Models.* Oxford University Press, 1996.

[LB19] Max A. Little and Reham K. I. Badawy. Causal bootstrapping. *ArXiv*, abs/1910.09648, 2019.

[LCB19] Sanghack Lee, Juan D. Correa, and Elias Bareinboim. General identifiability with arbitrary surrogate experiments. In *UAI*, 2019.

[LD04] Jared K. Lunceford and Marie Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, 23(19):2937–2960, 2004.

[Li18] Shu Li. *Estimating Causal Effects from Time Series*. PhD thesis, ETH Zurich, 2018.

[LSHL18] Chaochao Lu, Bernhard Schölkopf, and José Hernández-Lobato. Deconfounding reinforcement learning in observational settings. *ArXiv*, abs/1812.10576, 2018.

[Mac71] Saunders MacLane. *Categories for the Working Mathematician*, volume 5 of *Graduate Texts in Mathematics*. Springer-Verlag New York, 1971.

[Mee95] Christopher Meek. Causal inference and causal explanation with background knowledge. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, pages 403–410. Morgan Kaufmann Publishers Inc., 1995.

[MJS13] Joris Mooij, Dominik Janzing, and Bernhard Schölkopf. From ordinary differential equations to structural causal models: the deterministic case. *Uncertainty in Artificial Intelligence - Proceedings of the 29th Conference, UAI 2013*, 04 2013.

[MKS+15] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen. King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[MN19] Oliver J. Maclaren and Ruanui Nicholson. What can be estimated? identifiability, estimability, causal inference and ill-posed inverse problems. *ArXiv*, abs/1904.02826, 2019.

[MSPG13] George Michalareas, Jan-Mathijs Schoffelen, Gavin Paterson, and Joachim Gross. Investigating causality between interacting brain areas with multivariate autoregressive models of meg sensor data. *Human Brain Mapping*, 34(4):890–913, 2013.

[MvOC+18] Sara Magliacane, Thijs van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M. Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 10869–10879. Curran Associates Inc., 2018.

[MW15] Stephen L. Morgan and Christopher Winship. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press, 2nd edition, 2015.

[NPV19] Robert Ness, Kaushal Paneri, and Olga Vitek. Integrating markov processes with structural causal modeling enables counterfactual inference in complex systems. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 14211–14221. Curran Associates, Inc., 2019.

[NWM+19] Leonardo Novelli, Patricia Wollstadt, Pedro Mediano, Michael Wibral, and Joseph T. Lizier. Large-scale directed network inference with multivariate transfer entropy and hierarchical statistical testing. *Network Neuroscience*, 3(3):827–847, 2019.

[PB11] Judea Pearl and Elias Bareinboim. Transportability of causal and statistical relations: A formal approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 540–547, 2011.

[PB14] Judea Pearl and Elias Bareinboim. External validity: From do-calculus to transportability across populations. *Statistical Science*, 29(4):579–595, 2014.

[PBM16] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

[PBP19] Niklas Pfister, Stefan Bauer, and Jonas Peters. Learning stable and predictive structures in kinetic systems. *Proceedings of the National Academy of Sciences*, 116(51):25405–25411, 2019.

[PBP20] Niklas Pfister, Stefan Bauer, and Jonas Peters. Causalkinetix: Learning stable structures in kinetic systems, 2020. `https://CRAN.R-project.org/package=CausalKinetiX`, Last accessed on 2020-03-12.

[Pea] Judea Pearl. Causal and counterfactual inference. Forthcoming section in The Handbook of Rationality, MIT Press.

[Pea85] Judea Pearl. Bayesian networks: A model of self-activated memory for evidential reasoning. In *Proc. of Cognitive Science Society (CSS-7)*, 1985.

[Pea88] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., 1988.

[Pea93] Judea Pearl. [bayesian analysis in expert systems]: Comment: Graphical models, causality and intervention. *Statistical Science*, 8(3):266–269, 1993.

[Pea09] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition, 2009.

[Pea19] Judea Pearl. The seven tools of causal inference, with reflections on machine learning. *Commun. ACM*, 62(3):54–60, 2019.

[PF08] Graciela De Pierris and Michael Friedman. Kant and hume on causality. In Edward Zalta, editor, *Stanford Encyclopedia of Philosophy*. 2008.

[PJS17] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference - Foundations and Learning Algorithms*. Adaptive Computation and Machine Learning Series. The MIT Press, 2017.

[PMJS14] Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15:2009–2053, 2014.

[PV91] Judea Pearl and Thomas Verma. A theory of inferred causation. In *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning*, KR'91, pages 441–452. Morgan Kaufmann Publishers Inc., 1991.

[QLSL12] H. Qiu, Y. Liu, N. A. Subrahmanya, and W. Li. Granger causality for time-series anomaly detection. In *2012 IEEE 12th International Conference on Data Mining*, pages 1074–1079, 2012.

[RBB+19] J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, M. Glymour, C. andKretschmer, M.D. Mahecha, E.H. van Nes, J. Peters, R. Quax, M. Reichstein, B. Scheffer, M. Schölkopf, P. Spirtes, G. Sugihara, J. Sun, Ka. Zhang, and J. Zscheischler. Inferring causation from time series with perspectives in earth system sciences. *Nature Communications*, 10(1), 2019.

[RBDG19] Ribana Roscher, Bastian Bohn, Marco F. Duarte, and Jochen Garcke. Explainable machine learning for scientific insights and discoveries. *ArXiv*, abs/1905.08883, 2019.

[RCSTP18] Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19(1):1309–1342, 2018.

[RMBP18] Dominik Rothenhäusler, Nicolai Meinshausen, Peter Bühlmann, and Jonas Peters. Anchor regression: heterogeneous data meets causality. volume abs/1801.06229, 2018.

[RNK+19] Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, 5(11), 2019.

[Ros87] Paul R. Rosenbaum. Model-based direct adjustment. *Journal of the American Statistical Association*, 82(398):387–394, 1987.

[Roy12] Kjetil Roysland. Counterfactual analyses with graphical models based on local independence. *The Annals of Statistics*, 40(4):2162–2194, 2012.

[RPD+15] J Runge, V Petoukhov, JF Donges, N Jajcay, M Vejmelka, D Hartman, N Marwan, M Palus, and J Kurths. Identifying causal gateways and mediators in complex spatio-temporal systems. *Nature Communications*, 6(8502), 2015.

[RR83a] P. R. Rosenbaum and D. B. Rubin. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society. Series B (Methodological)*, 45(2):212–218, 1983.

[RR83b] Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[RRD19] Christian Reimers, Jakob Runge, and Joachim Denzler. Using causal inference to globally understand black box predictors beyond saliency maps. In *International Workshop on Climate Informatics (CI)*, 2019.

[RSG16a] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-agnostic interpretability of machine learning. *ArXiv*, abs/1606.05386, 2016.

[RSG16b] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144. Association for Computing Machinery, 2016.

[RSZ06] Joseph Ramsey, Peter Spirtes, and Jiji Zhang. Adjacency-faithfulness and conservative causal inference. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'06, pages 401–408. AUAI Press, 2006.

[Rub74] Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

[Rub04] Donald Rubin. Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics*, 31:161–170, 02 2004.

[Rub05] Donald B Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

[Run15] Jakob Runge. Quantifying information transfer and mediation along causal pathways in complex systems. *Phys. Rev. E*, 92(6), 2015.

[Run18a] Jakob Runge. Causal network reconstruction from time series: From theoretical assumptions to practical estimation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(7):075310, 2018.

[Run18b] Jakob Runge. Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 938–947. PMLR, 2018.

[Run20] Jakob Runge. Tigramite - causal discovery for time series datasets, 2020. `https://github.com/jakobrunge/tigramite/blob/master/README.md`, Last accessed on 2020-03-12.

[RW99] James M. Robins and Larry A. Wasserman. On the impossibility of inferring causation from association without background knowledge. In C. Glymour and G. Cooper, editors, *Computation, Causation, and Discovery*, pages 305–321. AAAI Press/The MIT Press, 1999.

[SB14] Jie Sun and Erik M. Bollt. Causation entropy identifies indirect influences, dominance of neighbors and anticipatory couplings. *Physica D: Nonlinear Phenomena*, 267:49–57, 2014.

[SBB15] Anil K. Seth, Adam B. Barrett, and Lionel Barnett. Granger causality analysis in neuroscience and neuroimaging. *Journal of Neuroscience*, 35(8):3293–3297, 2015.

[SBDH97] J. Stark, D.S. Broomhead, M.E. Davies, and J. Huke. Takens embedding theorems for forced and stochastic systems. *Nonlinear Analysis: Theory, Methods, and Applications*, 30(8):5303 – 5314, 1997. Proceedings of the Second World Congress of Nonlinear Analysts.

[Sch70] Tore Schweder. Composable markov processes. *Journal of Applied Probability*, 7(2):400–410, 1970.

[Sch94] David A. Schum. *The Evidential Foundations of Probabilistic Reasoning*. Northwestern University Press, 1994.

[Sch00] Thomas Schreiber. Measuring information transfer. *Phys. Rev. Lett.*, 85:461–464, 2000.

[Sch19] Bernhard Schölkopf. Causality for machine learning. *ArXiv*, abs/1911.10500, 2019.

[SGS00] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. The MIT press, 2nd edition, 2000.

[SH13] Alexander Sokol and Niels Hansen. Causal interpretation of stochastic differential equations. *Electronic Journal of Probability*, 19, 2013.

[Sim72] Christopher A. Sims. Money, income, and causality. *The American Economic Review*, 62(4):540–552, 1972.

[SJPZ11] Bernhard Schölkopf, Dominik Janzing, Jonas Peters, and Kun Zhang. Robust learning via cause-effect models. *ArXiv*, abs/1112.2738, 2011.

[SK19] Amit Sharma and Emre Kiciman. DoWhy: A Python package for causal inference, 2019. `https://github.com/microsoft/dowhy`, Last accessed on 2020-03-12.

[SLK18] Patrick Schwab, Lorenz Linhardt, and Walter Karlen. Perfect match: A simple method for learning representations for counterfactual inference with neural networks. *ArXiv*, abs/1810.00656, 2018.

[SMY+12] George Sugihara, Robert May, Hao Ye, Chih-hao Hsieh, Ethan Deyle, Michael Fogarty, and Stephan Munch. Detecting causality in complex ecosystems. *Science (New York, N.Y.)*, 338, 2012.

[SNDS90] Jerzy Splawa-Neyman, D. M. Dabrowska, and T. P. Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4):465–472, 1990.

[SP06] Ilya Shpitser and Judea Pearl. Identification of conditional interventional distributions. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence, UAI 2006*, pages 437–444, 2006.

[Ste16] David I. Stern. Economic growth and energy. In *Reference Module in Earth Systems and Environmental Sciences*. Elsevier, 2016.

[Str20] Eric Strobl. Rcit and rcot, 2020. `https://github.com/ericstrobl/RCIT`, Last accessed on 2020-03-12.

[SVZ19] Eric V. Strobl, Shyam Visweswaran, and Kun Zhang. Approximate Kernel-Based Conditional Independence Tests for Fast Non-Parametric Causal Discovery. *Journal of Causal Inference*, 7(1), 2019.

[SYC+20] George Sugihara, Hao Ye, Adam Clark, Ethan Deyle, Steve Munch, Jun Cai, Jane Cowles, Yair Daon, Andrew Edwards, Os Keyes, James Stagge, Masayuki Ushio, Ethan White, and Takeshi Abe. Redm: Applications of empirical dynamic modeling from time series, 2020. `https://cran.r-project.org/web/packages/rEDM/index.html`, Last accessed on 2020-03-12.

[Tib96] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[TW20] Ruey S. Tsay and David Wood. Mts: All-purpose toolkit for analyzing multivariate time series (mts) and estimating multivariate volatility models, 2020. `https://CRAN.R-project.org/package=MTS`, Last accessed on 2020-03-12.

[VM19] Jaroslav. Vondřejc and Hermann G. Matthies. Accurate computation of conditional expectation for highly nonlinear problems. *SIAM/ASA Journal on Uncertainty Quantification*, 7(4):1349–1368, 2019.

[Wie56] Norbert Wiener. The Theory of prediction. In Edwin Beckenbach, editor, *Modern Mathematics for Engineers*, pages 165–190. Dover Books on Engineering, 1956.

[WL10] Halbert White and Xun Lu. Granger Causality and Dynamic Structural Systems. *Journal of Financial Econometrics*, 8(2):193–243, 2010.

[Wol18] Patricia Wollstadt. Idtxl - theoretical introduction, 2018. `https://github.com/pwollstadt/IDTxl/wiki/Theoretical-Introduction`, Last accessed on 2019-10-14.

[Wol20a] Patricia Wollstadt. Idtxl, 2020. `https://github.com/pwollstadt/IDTxl`, Last accessed on 2020-03-12.

[Wol20b] Patricia Wollstadt. Idtxl, 2020. `https://github.com/pwollstadt/IDTxl/wiki/Runtimes-and-Benchmarking`, Last accessed on 2020-03-12.

[WPP+13] Michael Wibral, Nicolae Pampu, Viola Priesemann, Felix Siebenhühner, Hannes Seiwert, Michael Lindner, Joseph T. Lizier, and Raul Vicente. Measuring information-transfer delays. *PLOS ONE*, 8(2):1–19, 2013.

[Wri21] Sewall Wright. Correlation and causation. *Journal of agricultural research*, 20(7):557–585, 1921.

[Wri28] Philip Green Wright. *The tariff on animal and vegetable oils*. The Macmillan Company, 1928.

[YT02] Kenji Yamanishi and Jun-ichi Takeuchi. A unifying framework for detecting outliers and change points from non-stationary time series data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, page 676–681. Association for Computing Machinery, 2002.

[ZB19] Junzhe Zhang and Elias Bareinboim. Near-optimal reinforcement learning in dynamic treatment regimes. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 13401–13411. Curran Associates, Inc., 2019.

[ZGS15] Kun Zhang, Mingming Gong, and Bernhard Schoelkopf. Multi-source domain adaptation: A causal view. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pages 3150–3157, 2015.

[ZH19] Qingyuan Zhao and Trevor Hastie. Causal interpretations of black-box models. *Journal of Business & Economic Statistics*, 0(0):1–10, 2019.

[ZHZ+17] Kun Zhang, Biwei Huang, Jiji Zhang, Clark Glymour, and Bernhard Schölkopf. Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 1347–1353, 2017.

[ZKZS13] Y. Zhou, Z. Kang, L. Zhang, and C. Spanos. Causal analysis for non-stationary time series in sensor-rich smart buildings. In *2013 IEEE International Conference on Automation Science and Engineering (CASE)*, pages 593–598, 2013.

[ZPJS11] Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, pages 804–813. AUAI Press, 2011.

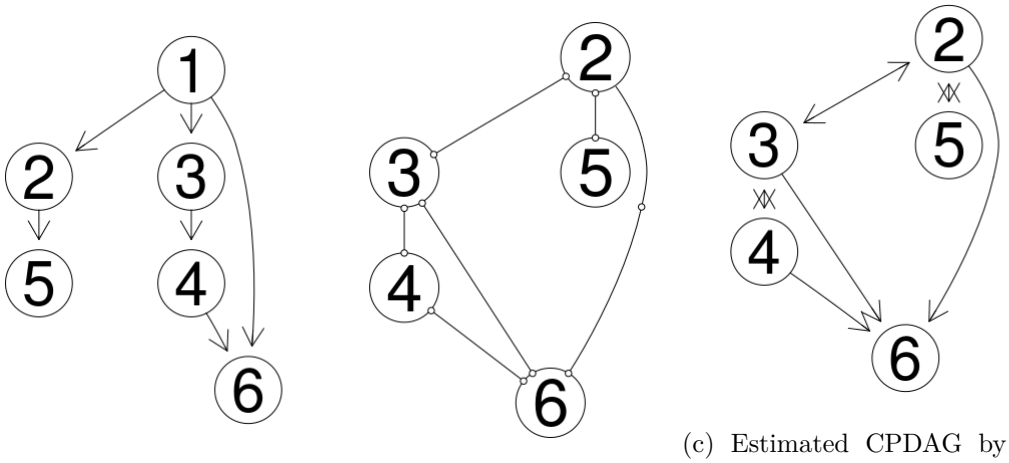
[ZS12] Mingyuan Zhang and Dylan S. Small. Effect of vitamin a deficiency on respiratory infection: Causal inference for a discretely observed continuous time non-stationary markov process. *Canadian Journal of Statistics*, 40(4):646–662, 2012.

[ZSMW13] Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 819–827. PMLR, 2013.

Appendix A

Finding Causal Relationships - IID Case - Experiments

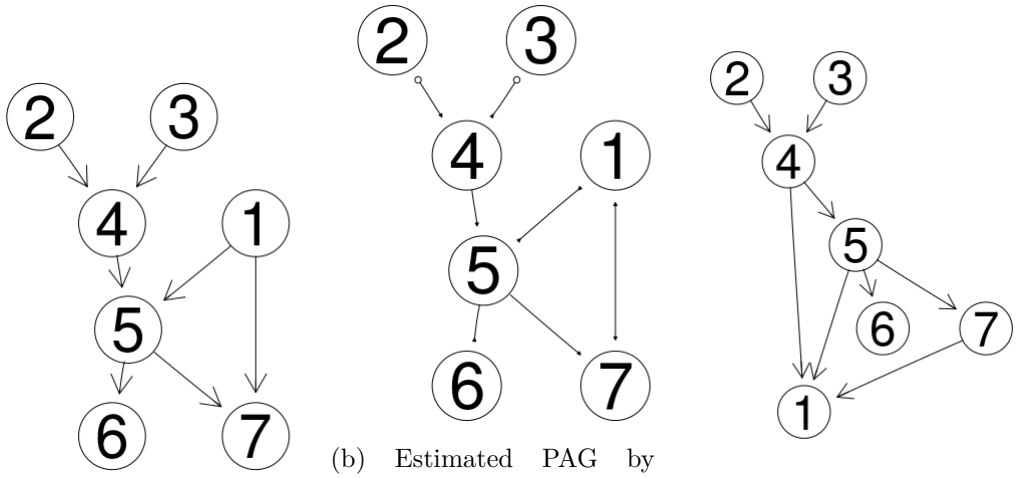
A.1 Latent Variables



(a) True DAG (b) Estimated PAG by FCI (c) Estimated CPDAG by GES

Figure A.1: Multivariate Gaussian data, where the first variable has been deleted, so that the first node is hidden for FCI and GES.

A.2 Different Distributions

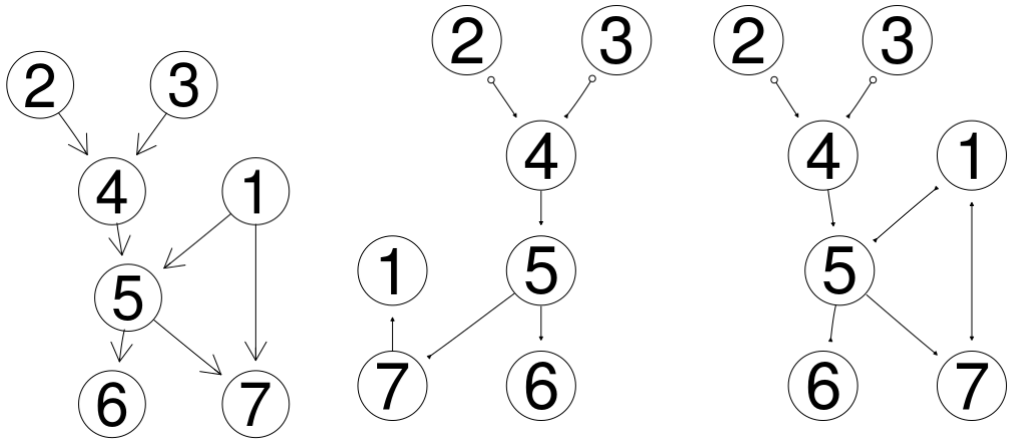


(a) True DAG

(b) Estimated PAG by FCI with Fisher's z-transformation

(c) Estimated CPDAG by GES

Figure A.2: The results of FCI with Fisher's z-transformation and GES for the second dataset with Gaussian distribution.

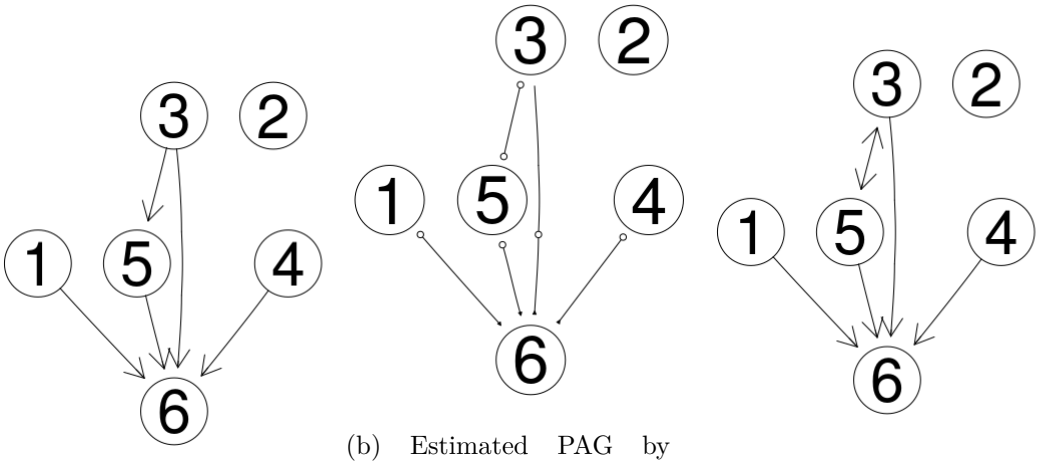


(a) True DAG

(b) Estimated PAG using RCIT

(c) Estimated PAG using RCoT

Figure A.3: The results of FCI with RCIT and RCoT for the second dataset with Gaussian distribution.

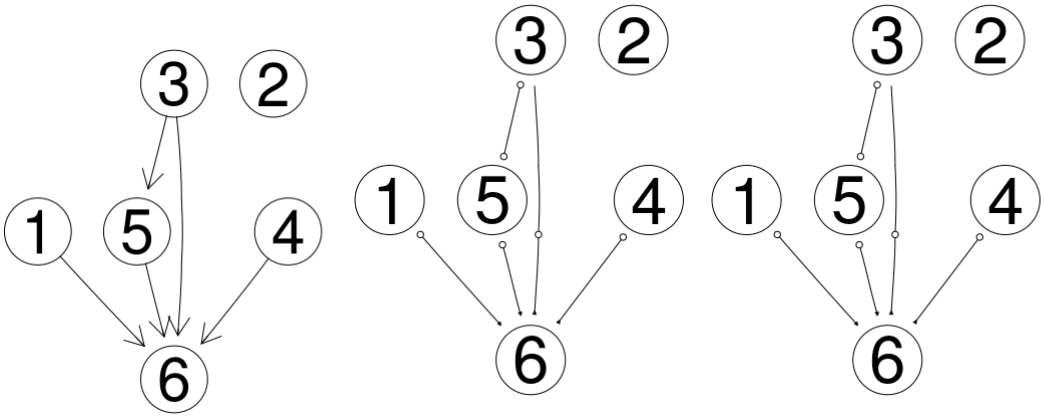


(a) True DAG

(b) Estimated PAG by FCI with Fisher's z-transformation

(c) Estimated CPDAG by GES

Figure A.4: The results of FCI with Fisher's z-transformation and GES for the third dataset with Gaussian distribution.

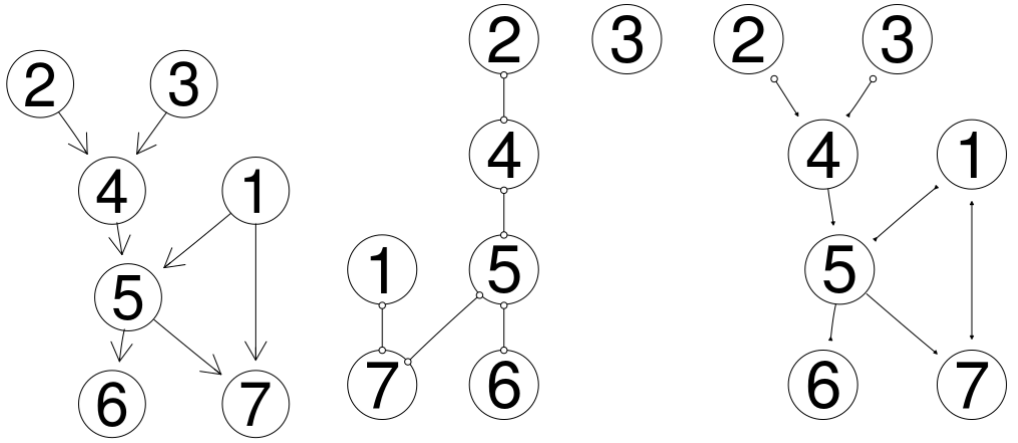


(a) True DAG

(b) Estimated PAG using RCIT

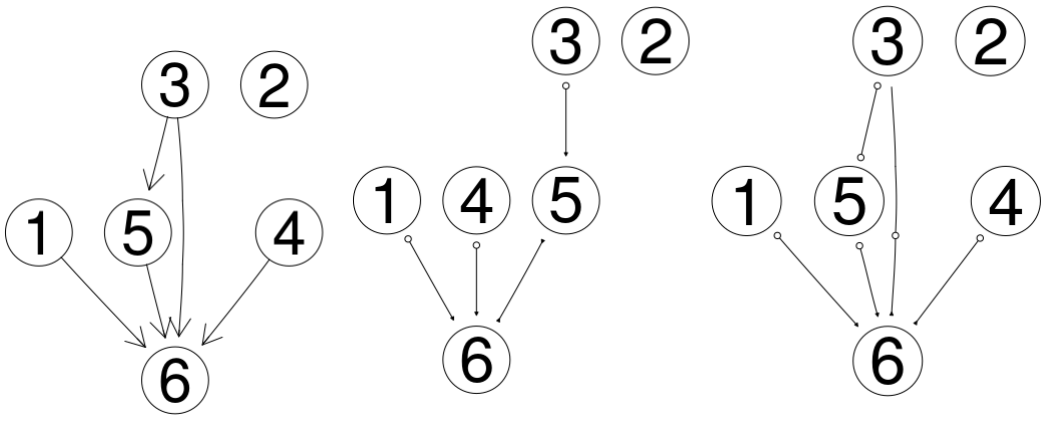
(c) Estimated PAG using RCoT

Figure A.5: The results of FCI with RCIT and RCoT for the third dataset with Gaussian distribution.



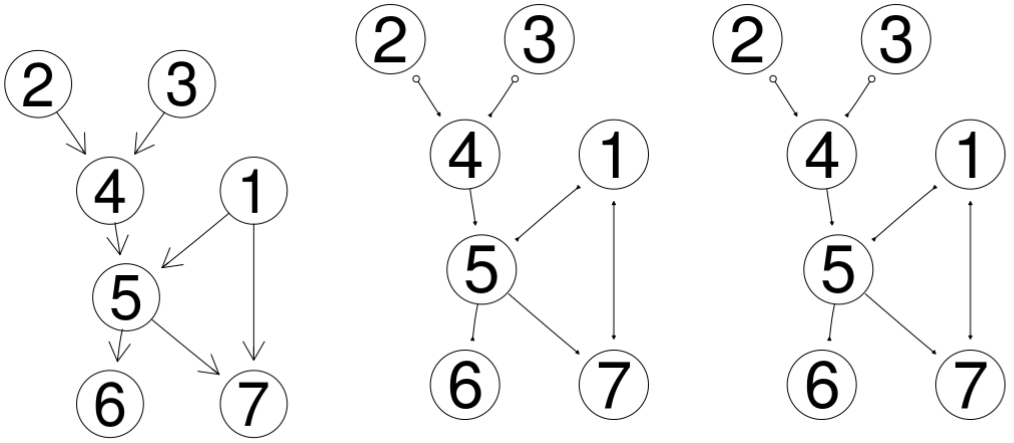
(a) True DAG (b) Estimated PAG using RCIT (c) Estimated PAG using RCoT

Figure A.6: The results of FCI with RCIT and RCoT for the second dataset with Cauchy distribution.



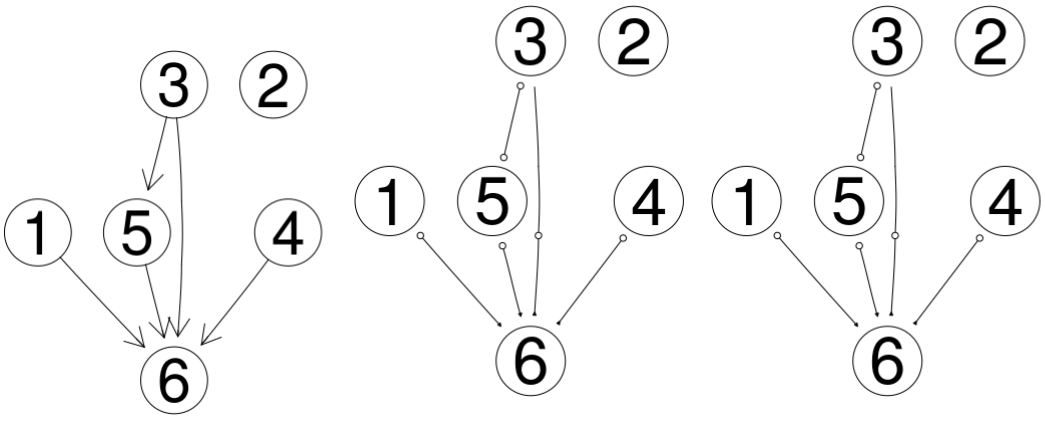
(a) True DAG (b) Estimated PAG using RCIT (c) Estimated PAG using RCoT

Figure A.7: The results of FCI with RCIT and RCoT for the third dataset with Cauchy distribution.



(a) True DAG (b) Estimated PAG using RCIT (c) Estimated PAG using RCoT

Figure A.8: The results of FCI with RCIT and RCoT for the second dataset with t-distribution.



(a) True DAG (b) Estimated PAG using RCIT (c) Estimated PAG using RCoT

Figure A.9: The results of FCI with RCIT and RCoT for the third dataset with t-distribution.

Appendix B

Finding Causal Relationships - Dynamical Systems - Experiments

B.1 Functional Data

B.1.1 Linear Case

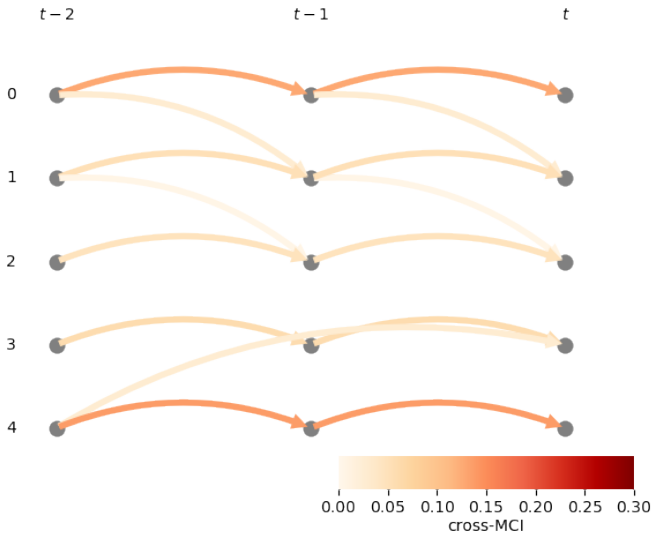


Figure B.1: The output of PCMCI for the dataset described by the functions in Equation 4.13 with Cauchy-distributed noise.

B.1.2 Non-Linear Case

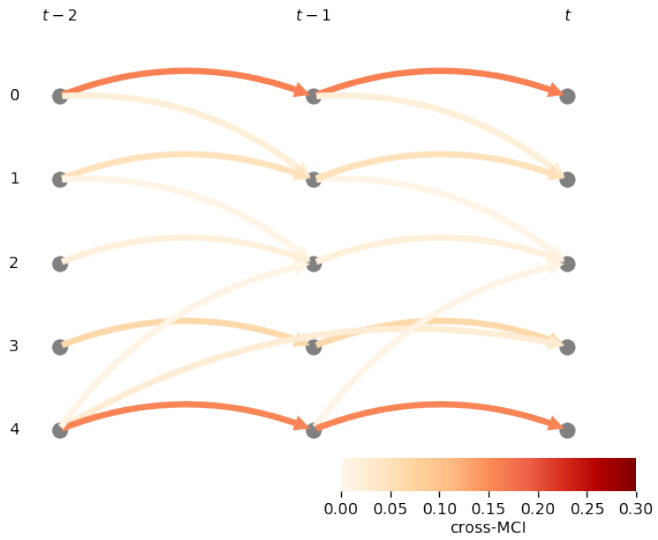


Figure B.2: The output of PCMCI for the dataset described by the functions in Equation 4.13 with t-distributed noise.

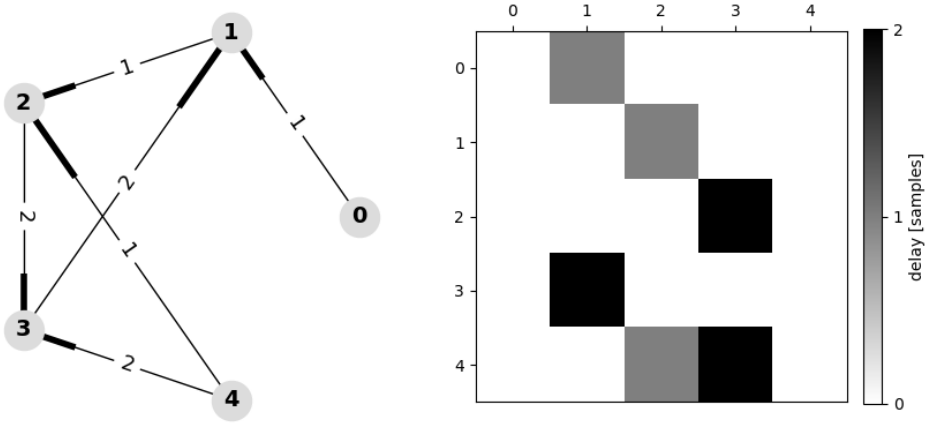


Figure B.3: The output of multivariate transfer entropy for the dataset described by the functions in Equation 4.13 with Cauchy-distributed noise.

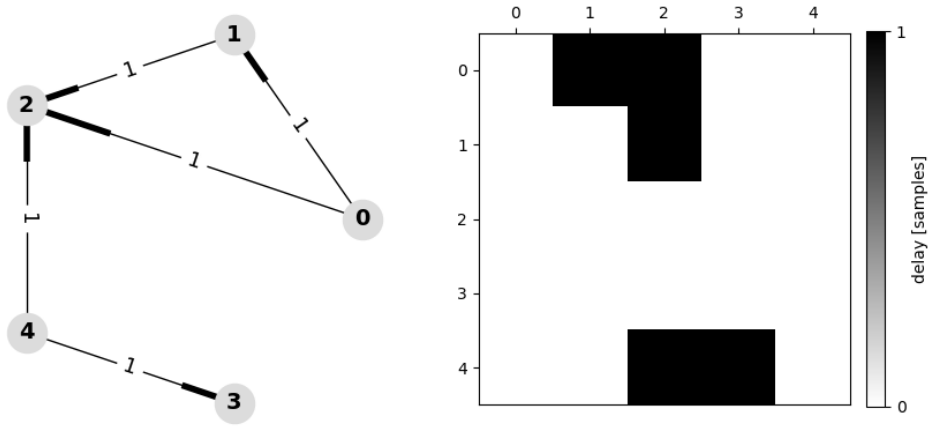


Figure B.4: The output of multivariate transfer entropy for the dataset described by the functions in Equation 4.13 with t-distributed noise.

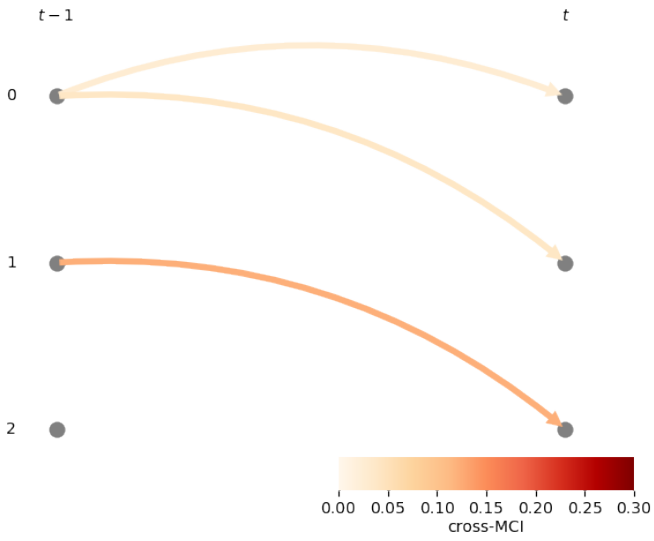


Figure B.5: The output of PCMCI for the dataset described by the functions in Equation 4.14 with Gaussian noise.

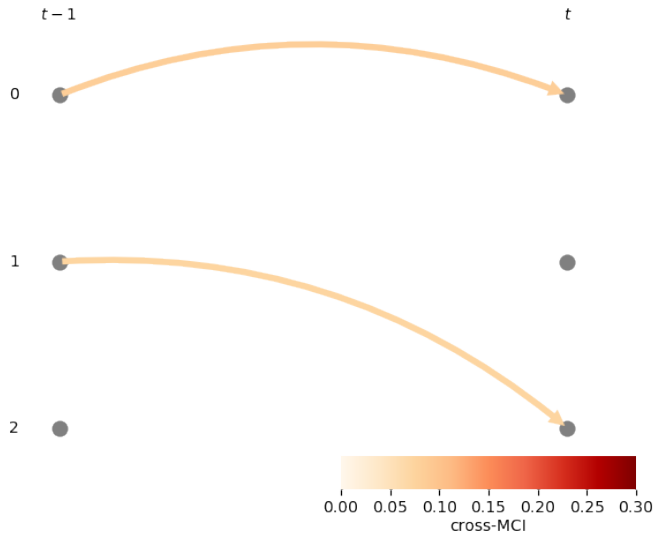


Figure B.6: The output of PCMCI for the dataset described by the functions in Equation 4.14 with Cauchy-distributed noise.

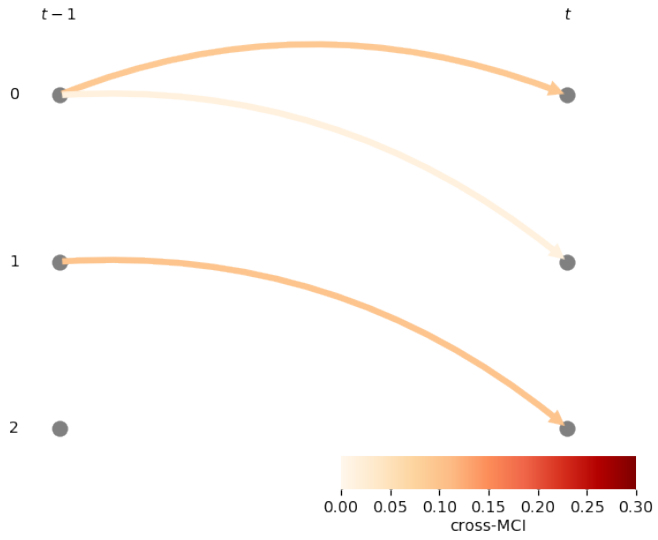


Figure B.7: The output of PCMCI for the dataset described by the functions in Equation 4.14 with t-distributed noise.

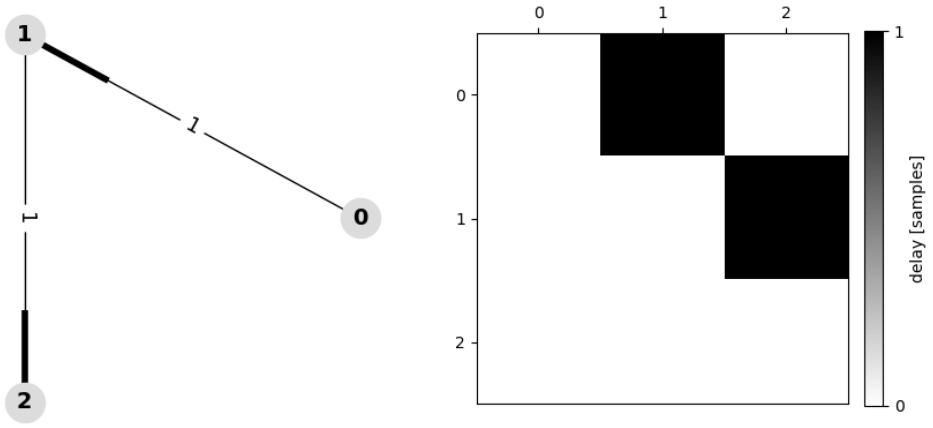


Figure B.8: The output of multivariate transfer entropy for the dataset described by the functions in Equation 4.14 with Gaussian noise.

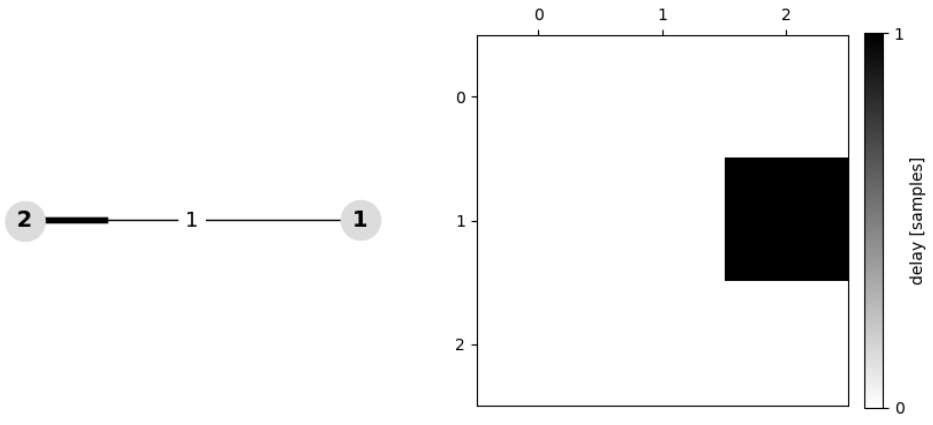


Figure B.9: The output of multivariate transfer entropy for the dataset described by the functions in Equation 4.14 with Cauchy-distributed noise.

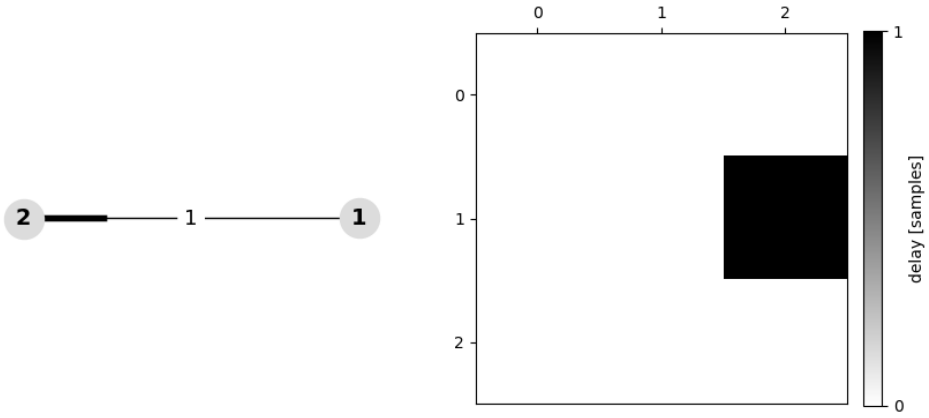


Figure B.10: The output of multivariate transfer entropy for the dataset described by the functions in Equation 4.14 with t-distributed noise.