HAL
archives-ouvertes.fr

Causal Discovery between time series

Karim Assaad

▶ **To cite this version:**

Karim Assaad. Causal Discovery between time series. Artificial Intelligence [cs.AI]. Université Grenoble Alpes [2020-..], 2021. English. NNT : 2021GRALM019 . tel-03438863

HAL Id: tel-03438863
https://tel.archives-ouvertes.fr/tel-03438863

Submitted on 22 Nov 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

Spécialité : Informatique

Arrêté ministériel : 25 mai 2016

Présentée par

Karim ASSAAD

Thèse dirigée par **Eric GAUSSIER**
et codirigée par **Emilie DEVIJVER**,CR, Université Grenoble Alpes

préparée au sein du **Laboratoire d'Informatique de Grenoble**
dans **l'École Doctorale Mathématiques, Sciences et technologies de l'information, Informatique**

Découvertes de relations causales entre séries temporelles

Causal Discovery between time series

Thèse soutenue publiquement le **5 juillet 2021**,
devant le jury composé de :

Monsieur ERIC GAUSSIER
PROFESSEUR DES UNIVERSITES, UNIVERSITE GRENOBLE ALPES, Directeur de thèse
Monsieur GREGOR GOESSLER
DIRECTEUR DE RECHERCHE, INRIA CENTRE GRENOBLE-RHONE-ALPES, Président
Monsieur HERVE ISAMBERT
DIRECTEUR DE RECHERCHE, CNRS DELEGATION PARIS CENTRE, Rapporteur
Monsieur PHILIPPE LERAY
PROFESSEUR DES UNIVERSITES, UNIVERSITE DE NANTES, Rapporteur
Madame MICHELE SEBAG
DIRECTEUR DE RECHERCHE, CNRS DELEGATION ILE-DE-FRANCE SUD, Examinatrice

À mon Père.

Abstract

This thesis aims to give a broad coverage of central concepts and principles of causation and in particular the ones involved in the emerging approaches to causal discovery from time series.

After reviewing concepts and algorithms, we first present a new approach that infers a summary causal graph of the causal system underlying the observational time series while relaxing the idealized setting of equal sampling rates and discuss the assumptions underlying its validity. The gist of our proposal lies in the introduction of the causal temporal mutual information measure that can detect the independence and the conditional independence between two time series, and in making an apparent connection between entropy and the probability raising principle that can be used for building new rules for the orientation of the direction of causation. Moreover, through the development of this base method, we propose several extensions, namely to handle hidden confounders, to infer a window causal graph given a summary causal graph, and to consider sequences instead of time series.

Secondly, we focus on the discovery of causal relations from a statistical distribution that is not entirely faithful to the real causal graph and on distinguishing a common cause from an intermediate cause even in the absence of a time indicator. The key aspect of our answer to this problem is the reliance on the additive noise principle to infer a directed supergraph that contains the causal graph. To converge toward the causal graph, we use in a second step a new measure called the temporal causation entropy that prunes for each node of the directed supergraph, the parents that are conditionally independent of their child. Furthermore, we explore complementary extensions of our second base method that involve a pairwise strategy which reduces through multitask learning and a denoising technique, the number of functions that need to be estimated.

We perform an extensive experimental comparison of the proposed algorithms on both synthetic and real datasets and demonstrate their promising practical performance: gaining in time complexity while preserving accuracy.

Résumé

Cette thèse a pour but d'expliquer les concepts et principes centraux de la causalité. Nous nous intéresserons particulierement à la découverte causale à partir de séries temporelles, domaine émergent aujourd'hui avec, notamment, les données industrielles de capteurs. Dans les deux premiers chapitres, nous présentons les concepts puis les algorithmes existants dans ce domaine.

Ensuite, nous présentons une nouvelle approche qui infère un graphe récapitulatif du système causal sous-jacent aux séries temporelles tout en assouplissant le cadre idéalisé de fréquences d'échantillonnage égaux, tout en discutant ses hypothèses et sa validité. La principale nouveauté dans cette méthode réside dans l'introduction de la mesure d'information mutuelle temporelle causale qui permet de détecter l'indépendance et l'indépendance conditionnelle entre deux séries temporelles, et l'établissement d'un lien apparent entre l'entropie et le principe d'augmentation de la probabilité d'un effet sachant sa cause, lien qui peut être utilisé pour construire de nouvelles règles pour l'orientation de la direction de la causalité. De plus, à travers le développement de la première méthode, nous proposons plusieurs extensions qui permettent de gérer les causes communes cachées, de déduire un graphe causal temporel à partir d'un graphe récapitulatif et de pouvoir s'adapter aux données ordonnées (pas nécessairement temporelles).

Puis, nous nous concentrons sur la découverte de relations causales à partir d'une distribution statistique qui n'est pas entièrement fidèle au graphe causal réel et sur la distinction entre une cause commune et une cause intermédiaire même, en absence d'indicateur de temps. L'aspect clé de notre réponse à ce problème est le recours au principe du bruit additif pour déduire un supergraphe dirigé contenant le graphe causal. Pour converger vers le graphe causal, nous utilisons l'entropie de causalité temporelle qui élague pour chaque nœud du supergraphe dirigé, les parents qui en sont conditionnellement indépendants. En outre, nous explorons des extensions complémentaires de notre deuxième méthode qui impliquent une stratégie par paires et une stratégie multitâche.

Nous effectuons une comparaison expérimentale approfondie des algorithmes proposés sur des ensembles de données à la fois synthétiques et réels et nous montrons leurs performances pratiques prometteuses : gain en complexité temporelle tout en préservant la précision.

Acknowledgments

Firstly, I would like to express my sincere gratitude to my advisors Emilie Devijver and Eric Gaussier for the continuous support of my PhD study and related research, for their patience, motivation, and immense knowledge in statistics and machine learning. Their guidance helped me in all the time of research and writing of this thesis. I could not have imagined having better advisors and mentors for my PhD study.

Besides my advisors, I would like to thank the rest of my thesis committee: my reviewers Hervé Isambert and Philippe Leray, for carefully reading my manuscript, for their insightful comments, and their encouragement, and my examinators Gregor Goessler and Michele Sebag for their questions which incented me to widen my research from various perspectives.

My sincere thanks also go to Coservit, who funded my research and provided me the opportunity to join their team. In particular, I am grateful to Ali Aït-Bachir and Rachid Mokhtari for enlightening me about the different industrial challenges. Without their precious support, it would not be possible to conduct this research.

I must also acknowledge Isabelle Drouet, who taught me during the Journées d'Étude en Statistique (JES 2018), that causation is above all a philosophical concept and Jacob Runge for numerous discussions (namely during UAI 2020 and Neurips 2020) which permitted me to take a step back and look at my work with a critical eye.

I would also like to thank my colleagues in the Laboratoire d'Informatique de Grenoble and in Coservit, as well as my friends in Grenoble for all the fun we had in the last three years. In particular, I thank the members of the Causality Reading Group for stimulating discussions.

Lastly, but not least, I would like to thank my family: my father for being there whenever I need him, my mother for cheering me up whenever I'm feeling down, my sisters for supporting me in my life in general, and Dasha Bystrova for supporting me throughout writing this thesis.

Contents

List of Figures

List of Tables

List of Symbols and Notations

Time series and random variables

d number of time series, of observations and of time points

N number of observations or of time points

X, X^p, X_t^p multivariate observational time series $X = \{X^1, \cdots, X^d\}$, its pth time series, and the time point of the pth time series at time t

$X^{\mathbf{R}}, X_t^{\mathbf{R}}, X_{\mathbf{T}}^{\mathbf{R}}$ subset of time series $X^{\mathbf{R}} = \{X^{r_1}, \cdots, X^{r_K}\}$, subset of time series at time t $X_t^{\mathbf{R}} = \{X_t^{r_1}, \cdots, X_t^{r_K}\}$, subset of time series evaluated on different time points $X_{\mathbf{T}}^{\mathbf{R}} = \{X_{t_1}^{r_1}, \cdots, X_{t_K}^{r_K}\}$

$X^{(p;\lambda)}$ window representation of time series X^p with a window size λ

ξ noise

L, S latent variable, hidden selection variable

τ window size

γ_{max} maximum lag

λ_{max} maximum window size

Graphical Symbols

$\mathcal{G}$ causal graph

V set of vertex in the graph

E set of edges in the graph

U a path in the graph

$X^p - X^q$ X^p is a neighbor of X^q

$X^p \to X^q$ X^p is a cause of X^q (X^p is a parent of X^q) and X^q is an effect of X^p

$X^p \leftrightarrow X^q$ X^p and X^q have a common confounder

$\mathrm{Adj}(X^p, \mathcal{G})$ variables adjacent to X^p in $\mathcal{G}$

$\mathrm{Par}(X^p, \mathcal{G})$ set of causes of X^p (set of parents of X^p) in $\mathcal{G}$

$\mathrm{Hom}(X^p_{t-i}, X^q_t, \mathcal{G})$ set of vertex pairs (X^p_{k-i}, X^q_k) homologous to (X^p_{t-i}, X^q_t)

DAG directed acyclic graph

CPDAG completed partially directed acyclic graph

MAG maximal ancestral graph

PAG partial ancestral graph

Other Symbols

$I(X^p; X^q)$ mutual information between X^p and X^q

$\perp\!\!\!\perp, \not\perp\!\!\!\perp$ independent and not independent

$\Pr(X = x), \mathbb{E}(X)$ probability of $X = x$ and expectation of X

P probability distribution

Notations

$Sepset(X^p, X^q)$ separation set of X^p and X^q

$Dsepset(X^p, X^q)$ d-separation set of X^p and X^q

Chapter 1

Introduction

I would rather discover one causal relation than be king of Persia.

Democritus

In this chapter, we first give a brief intuitive introduction of the concept of causation and show why it is crucial for intelligent systems, while building toward the importance of the topic of causal discovery. Throughout the discussion, we introduce causal graphs and its associated subroutines related to time series which are crucial for this thesis, and finally we outline the problems raised in this thesis.

1.1 Causation

Causality takes root in the center of our universe. Humankind has long been aware of causation, as it was mentioned in the ancient Hindu scriptures: "Cause is the effect concealed, effect is the cause revealed" [Ivancevic and Ivancevic, 2007]. It was also defined in a medical context in ancient Greece: "We consider the causes of each condition to be those things which are such that, when they are present, the condition necessarily occurs, but when they change to another combination, it ceases" [Nutton, 1980]. Even Democritus famously proclaimed that he would rather discover a causal relation than be the king of presumably the wealthiest empire of his time. But what is causation? There is not a universal answer, as causation received many definitions. On the one hand, causation is regarded as a primitive concept that indicates how the world progresses, so basic a concept that it is more apt as an explanation of other concepts than as something to be explained by others more basic. On the other hand, causation is regarded as derivative and an abstraction. Either way, it seems that without a doubt our intelligence is based on it and that a leap of intuition may be needed to grasp it. So in its most intuitive form, causation is regarded as the influence by which a cause (one event, process, etc.) contributes to the production of the effect (another event, process, etc.) where the cause is partly responsible for the effect and the effect is partly dependent on the cause and can, in turn, be a cause of many other effects. Accordingly, causality is implicit in the logic and structure of ordinary language and it is embedded in our understanding mechanism that pushes humans to invoke *why* questions. Why is it dark? Where do babies come from? Why is the sea salty? What is the effect of exercise on heart rate? What is the effect of industrial pollution on the environment? What's the source of the cholera disease [Snow, 1855]? And so, as already advocated by Spirtes, Glymour and Scheines, in attempting to answer such questions, both the baby and the scientist try to turn observation into causal knowledge [Spirtes et al., 2000].

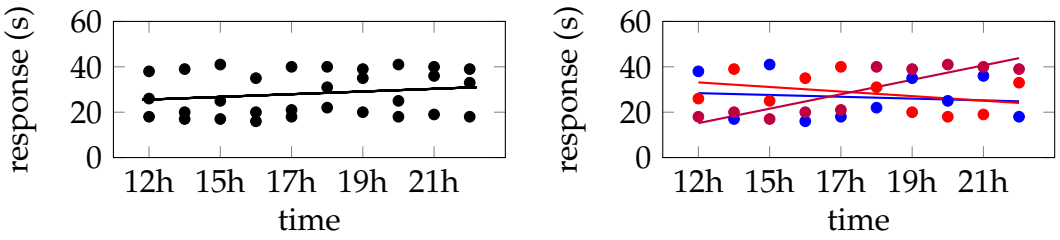
1.2 Why causality?

From the perspective of artificial intelligence (AI), causality is crucial for explanatory AI, since an effect is usually explained as a subset of its causes [Miller, 2019], while it is also necessary for invariant predictive systems. If the judgments of machines are to be of any use to us, we need them to distinguish between causal relations and mere correlations. Causality makes things invariably predictable and explainable, whereas correlation is a powerful tool to predict with no insurance about invariability and which is not sufficient to explain. Commercial and research work in AI is currently exploding everywhere, and investments in AI are accelerating at an unprecedented rate. Progress in certain fields like game playing and computer vision has been extraordinary: using learning techniques, a machine can distinguish objects in images, recognize speech, and beat humans in old Atari games and board games like Go. Nevertheless, despite these impressive achievements, the usage of machine learning methods in industry is lagging far behind. Industrial organizations still have many concerns about deploying AI systems, and these concerns are well justified. The first main concern relates to explainability [Spreeuwenberg et al., 2019]. Nowadays, in the field of AI, there is a clear tension between performance (predictive accuracy) and explainability. Often the best-performing methods such as deep learning are black boxes, meaning that they lack transparency and provide no explanation. In practical cases, intelligent systems need to collaborate with humans to solve complex problems, and like any efficient collaboration, this requires good communication, trust, clarity, and understanding [Ribeiro et al., 2016]. But how can we trust such a system if it does not provide us with an explanation? In addition, to be integrated with high-risk applications, these systems need to be held accountable for their own actions. Imagine a world in which self-driving cars replaced other means of transportation; in such a world, how can we determine who is responsible for a car crash? Was it a technical error or a human error? It is evident that a black box system would not be useful in such investigations. Lastly, intelligent systems have shown prejudice in terms of gender (promoting men for job offers) and ethnicity. Thus, fairness is an important condition that these systems should fulfill before conquering the industry. But if the decision-making processes are not explained, how can we easily verify whether the model discriminates based on gender, political affiliation, or race? As an example, risk assessment software in the US sentenced a convict to a six-year imprisonment. Since the software was unable to provide an explanation for the verdict, the prisoner's right to a decision based on accurate information had been violated

[Završnik, 2019]. The second main concern is linked to robustness [Peters et al., 2016]. Currently, machine learning techniques excel at classifying objects, but it could be duped or confounded by novel situations: a highly efficient neural network trained to classify animals could easily categorize a dog as a fish if the dog was swimming (supposing that the training data illustrated multiple scenarios of fish swimming in water, and no scenarios of dogs in water); or simple disruptions on a highway that are hardly noticed by humans can cause state-of-the-art deep learning systems to misclassify road signs [Eykholt et al., 2018]. A small change in the distribution can have severe consequences on the performance of current AI systems, and many scholars are blaming deep learning by pointing to their black box nature and alchemy [Darwiche, 2018]. Furthermore, an AI system laboriously trained to carry out one task (e.g., identifying cats) has to be taught all over again to do something else (e.g., identifying dogs). Transfer learning methods have been introduced to tackle such problems, but with these methods, the intelligent system is liable to lose some of the expertise gained in the original task. To use the old knowledge acquired from learned tasks to help learn a new task, it is first necessary to distinguish between relevant transportable predictors and irrelevant noisy information.

These shortcomings have something in common: they exist because intelligent systems do not understand causation [Pearl and Mackenzie, 2018]. While they can see that some events are associated with others, they are unable to ascertain which things directly cause others to happen. Indeed, machine learning systems perform well when learning connections between input data and output predictions, although they find it difficult to reason about cause-effect relations and adapt to environmental changes. Machine learning models that can capture causal relationships will be more explicable, generalizable, robust, and transportable. Indeed, machine learning systems based on causes and effects are explicable by construction, since the explanation is usually a subset of the causes [Miller, 2019]. In such systems, robustness is ensured, since causality models interventional distribution, making it immune to distribution changes and easily transported [Pearl and Bareinboim, 2011]. Causality allows us to easily find and transfer predictors from one domain to another: accidents are correlated with black cars in the Netherlands but perhaps with red cars in the US. Using color as a predictor does not generalize, but a causal factor such as male testosterone levels will generalize easily. In addition, causality would open doors to new horizons that machine learning could not explore. Armored with causality, intelligent systems will acquire the ability to perform new crucial tasks: answering causal questions. Many

statistical attempts have been made to answer such questions, but they are often confronted with Simpson's Paradox or Berkson's paradox [Pearl and Mackenzie, 2018]. Simpson's paradox relates to confounder bias, which implies that an association in an overall population can disappear or even change direction in subpopulations, that is, when it is conditioned by a relevant variable, typically a grouping variable. For example, in Figures 1.1a and 1.1b, the response time of servers for a given product, which was upgraded at time 17h, is plotted against time; in the former all points have the same color, while in the latter, points are distinguished by different colors depending on the their server type. In Figure 1.1a, the update did not change the response time of the servers. But by looking at the response time of each server individually in Figure 1.1b, we seem to have a different conclusion: the update did change the response time of the servers. Which is the right conclusion? Intuitively this question is simple and naive. But statistically speaking, a model cannot decide which conclusion is the right one, therefore in statistics, such problems are considered a paradox (for an example of Simpson's paradox relating to Covid-19, see von Kügelgen et al. [2020]). Berkson's paradox relates to selection bias. For exam-



(a) (b)

Figure 1.1 – Example of Simpson's paradox: the upgrade appears to be beneficial in one of the servers but ineffective over all servers.

ple, as shown in Figure 1.2, during World War II, the Royal Air Force lost many planes to anti-aircraft fire, so they decided to armor them. But where should the armor be put? The data told them that bullet holes are always on the wings and in the middle of the planes, so it seemed obvious that the extra armor should be put in these places. However, this was incorrect, because the data did not contain all the information. Indeed, airplanes with bullets in other places never returned. The data therefore misled them, because they did not take the unknown data into account. They did not ask themselves *why* or engage in counterfactual thinking, which is usually of the sort: "What would have happened if ...?". Fortunately, the mathematician Abraham Wald was able to do so [Wald and for Naval Analyses ,

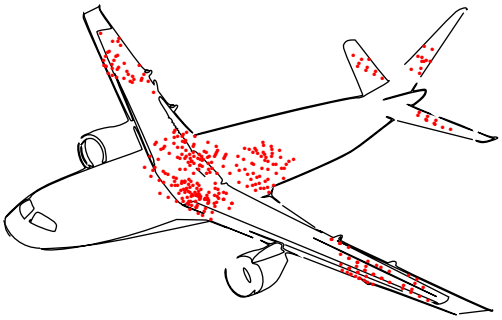


Figure 1.2 – The damaged parts (in red) of returning airplanes show locations where they can support damage and still return home.

U.S.].

One of the main reasons why causal notions were non existent in most of the machine learning literature, is that causal questions cannot be expressed through mathematical equations. Consequently, a concrete research effort began in the last few decades, and introduced a causal calculus known as do-calculus [Pearl, 2000] (for more details see Chapter 2 Section 2.8), that permits the generation of probabilistic formulas for the effect of interventions in terms of the observed probabilities, i.e., turn causal questions into statistical equations that can be estimated from the data. To enable a machine to use do-calculus efficiently, one needs to put at its disposal a causal structure that represents the causal relations of the system. Causal structures are often explicitly represented in terms of a directed graphs. Unlike other graphs with directed or undirected edges, which merely represent an independence structure, causal graphs support a very strong interpretation formalized as follows:

Definition 1 (Causal graph)**.** *A graph $\mathcal{G} = \{V, E\}$ with a set of nodes $V = X^1, \cdots, X^d$ and set of edges E, is causal, if for any directed edge $X^p \rightarrow X^q$ in E, X^p is a direct cause of X^q relative to nodes in V.*

It means that even if you randomize all other variables in $V \backslash \{X^p, X^q\}$, thereby breaking any causal connection between X^p and X^q through these

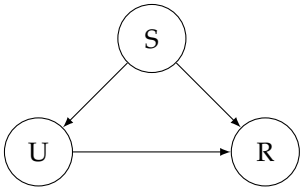


Figure 1.3 – Causal graph showing that server type is a mediator between the update and response time. R: response time, S: server type, U: update.

other variables, X^p still has a causal effect on X^q. The above definition allows for cyclic causal relations. In contrast with the typical assumption in the causal discovery literature, we do not assume here that the true causal graph is necessarily a Directed Acyclic Graph (DAG).

Using the graphical notation and do-calculus, causal questions can be treated by a machine. For the sake of illustration let's go back to the Simpson paradox example (Figure 1.1). Assume that the causal graph of the system is the one presented in Figure 1.3; the server type is a mediator between the update and the response time. Then, by simple causal reasoning, the answer becomes clear. To remove the confounding bias, we should condition on the common cause. The do-calculus is a tool that enables machines to do the exact same reasoning as we just did. The do-calculus allows an analysis of the effect of interventions or distribution changes without actually requiring a physical intervention by drawing conclusions from a causal graph. The prerequisite of this process is knowing the underlying causal graph. But what happens when we do not have a causal graph at our disposal?

1.3 Problem statement

Sometimes it is possible to construct a causal graph with the help of a human expert. But in complex systems, human experts are not always capable of providing such graphs. Another possibility involves identifying causal relations in an experimental setting, such as a randomized controlled trial where each individual in the experiment is randomly assigned to either the treatment or control group. Naturally, the interventional aspect of these procedures raise many concerns, starting from the ethical concerns, the expense, and the feasibility, namely when the data-generating process is unavailable [Spirtes and Zhang, 2016]. In such case, the only possibility left is to try to discover causal relations from observations, known as causal discovery and which denotes the inverse problem

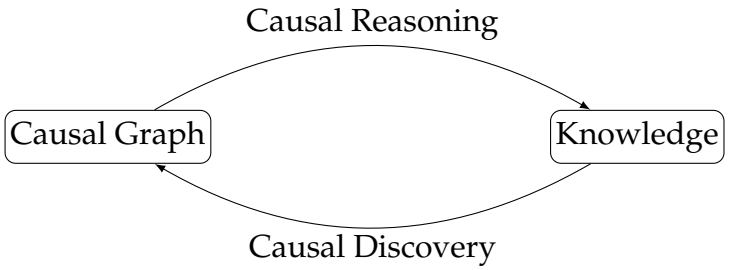


Figure 1.4 – The relationships between the causal knowledge and causal graphs, and between causal discovery and causal reasoning.

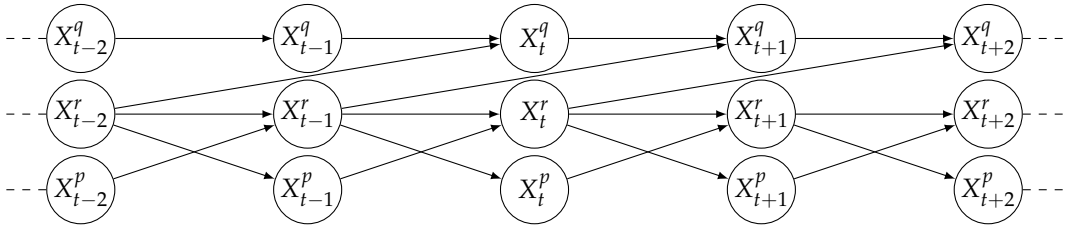
to causal reasoning (as summarized in Figure 1.4). However, causal relations are not features that can be directly read off from the data but have to be inferred. The field of causal discovery is concerned with the inference of a causal structure from its empirical implications and the assumptions that support it. It has drawn much attention in several fields, and proposals have been presented for various types of data: independent identically distributed (i.i.d.) data [Spirtes et al., 2000, Hoyer et al., 2009], time series [Peters et al., 2013, Runge et al., 2019] and images [Lopez-Paz et al., 2017]. In this thesis we mainly focus on observational [1] time series.

Time series arise as soon as observations, from sensors or experiments, for example, are collected over time. They are present in various forms in many different domains, as healthcare (through, e.g., monitoring systems), Industry 4.0 (through, e.g., predictive maintenance and industrial monitoring systems), surveillance systems (from images, acoustic signals, seismic waves, etc.) or energy management (through, e.g. energy consumption data) to name but a few.

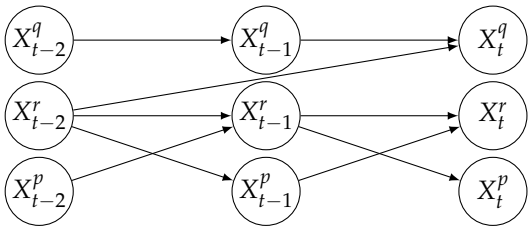
For time series, the causal graph $\mathcal{G} = (V, E)$ with V the set of vertices and E the set of edges is called a *full time causal graph* (also called *infinite dynamic causal graph* in [Malinsky and Spirtes, 2018]) and represents a complete graph of the dynamic system, through infinite vertices.

Definition 2 (Full time causal graph)**.** *Let X be a multivariate discrete-time stochastic process and $\mathcal{G} = (V, E)$ the associated* full time causal graph*. The set of vertices in that graph consists of the set of components $X^1, \ldots, X^d$ at each time $t \in \mathbb{Z}$. The edges E of the graph are defined as follows: variables X^p_{t-i} and X^q_t are connected by a lag-specific directed link $X^p_{t-i} \to X^q_t$ in $\mathcal{G}$ pointing forward in time if and only if X^p causes X^q at time t with a time lag of $i \geq 0$ for $p \neq q$*

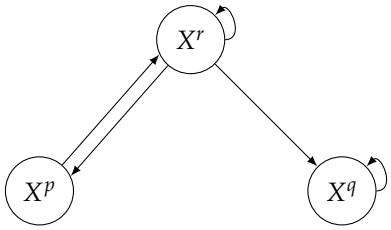
1. In observational time series, the value of a variable is always determined by its causes, hence it is never set through an intervention.



(a) Full time causal graph



(b) Window causal graph



(c) Summary causal graph

Figure 1.5 – Different causal graphs that one can infer from three time series: full time causal graph (1.5a), window causal graph (1.5b), and summary causal graph (1.5c). Note that the first one gives the more information but cannot be inferred in practice, the second one is a schematic viewpoint of the full behavior, whereas the last one gives an overview and can be deduced from the window causal graph.

and with a time lag of $i > 0$ *for* $p = q$.

In practice, inferring an infinite graph is unfeasible. However, it is very likely that causal relations between two time series will hold throughout time as such relations are generally associated with underlying physical processes. One thus relies on the so-called *Consistency throughout time* (also referred to as causal stationarity) assumption.

Definition 3 (Consistency throughout time)**.** *A causal graph* $\mathcal{G} = (V, E)$ *for a multivariate time series* X *is said to be* consistent throughout time *if all the causal relationships remain constant in direction throughout time, i.e., for two time series* X^p *and* X^q*, if* X^p_{t-i} *causes* X^q_t *then* X^p_{t-i-j} *causes* X^q_{t-j} *for all* j.

Under this assumption, and given a window size τ that can capture all causal relations that can be present in the system, the full time causal graph can be contracted to give a finite graph which we call *window causal graph*, with τ nodes for each time series [Runge et al., 2019].

Definition 4 (Window causal graph)**.** *Let* X *be a multivariate discrete-time stochastic process and* $\mathcal{G} = (V, E)$ *the associated* window causal graph *for a*

window of size τ. The set of vertices in that graph consists of the set of components $X^1, \ldots, X^d$ at each time $t, \ldots, t+\tau-1$. The edges E of the graph are defined as follows: variables X^p_{t-i} and X^q_t are connected by a lag-specific directed link $X^p_{t-i} \to X^q_t$ in $\mathcal{G}$ pointing forward in time if and only if X^p causes X^q at time t with a time lag of $0 \leq i < \tau$ for $p \neq q$ and with a time lag of $0 < i < \tau$ for $p = q$.

It is a representation of the causal graph through a time window the size of which equals the maximum lag relating time series in the full time causal graph. Figure 1.5b illustrates a window causal graph derived from the full time causal graph given in Figure1.5a with consistency throughout time. This graph summarizes the following causal relations: X^r causes X^p with a lag equal to 1 and X^q with a lag equal to 2, and X^p causes X^r with a lag equal to 1. In practice, it is often sufficient to know the causal relations between time series as a whole, without necessarily knowing the time delay between the cause and the effect. In that case, one can further compress the causal graph in a *summary causal graph* (also called *unit graph* in [Chu and Glymour, 2008]) that represents causal relation within and between time series without any time information and without referencing lags [Peters et al., 2013].

Definition 5 (Summary causal graph)**.** *Let X be a multivariate discrete-time stochastic process and $\mathcal{G} = (V, E)$ the associated* summary causal graph*. The set of vertices in that graph consists of the set of components $X^1, \ldots, X^d$. The edges E of the graph are defined as follows: variables X^p and X^q are connected if and only if there exists some time t and some time lag i such that X^p_{t-i} causes X^q_t at time t with a time lag of $0 \leq i < \tau$ for $p \neq q$ and with a time lag of $0 < i < \tau$ for $p = q$.*

An example of such a graph is given in Figure 1.5c. Note that since a summary causal graph is a summary of the full time causal graph, it can contain cycles, whereas window causal graph is always acyclic assuming that the full time causal graph is acyclic. Summary causal graphs are less sensitive to possible variations in time and errors in estimating time lags compared to full time and window causal graphs. Addressing the problem of learning summary causal graphs, without first resorting to window causal graphs, can be beneficial. For example, root cause analysis in monitoring systems mainly requires knowing what are the potential causes of a given time series metric without necessarily knowing the time delay between the cause and effect. In addition, it seems that in some complex systems (where we observe only a macro version of the real physical components or we simply don't have access to all variables), the summary

causal graph represents a more robust representation of causal relations compared to the full time causal graph, because time delays can vary whereas the causal relations themselves are immune to such variations. Moreover, it seems that expert can easily validate summary causal graphs as it provides a simple and efficient view on the causal relations that exist between time series.

Regardless of which type of graph a causal discovery algorithm seeks to infer, it will always be confronted with certain difficulties regarding temporal data, such as the timing and the frequency. To avoid these difficulties, causal discovery algorithms usually assume same timing and frequency for all time series. Although the first condition has its merits in the context of causation, the second has not. The condition that time series have the same sampling rates with identical timestamps is not necessarily met, so should be relaxed. This will be particularly an important industrial usage, where this kind of assumption is easily violated. For instance, in a monitoring system, not all metrics are collected with the same time gap. Some metrics are more important than others in respect to the maintenance of the system and so should be collected instantly, while others are of less importance to maintenance, thus, collecting them would imply a low return on investment.

Given the notion of true causal graph, full time causal graph, maximal lag $\gamma_{max} = \tau - 1$, summary causal graph, and different sampling rate, we can now formulate the main problem of this thesis.

Problem 1. *Given a maximal lag γ_{max} and an observational time series $X^1, \cdots, X^d$ with potentially different sampling rates, infer the underlying summary causal graph corresponding to the true full time causal graph.*

The approach we take to solve Problem 1 fits well within the well-known constraint-based framework which was initially presented for non temporal data. Under some assumptions (which will be discussed in depth later), the approaches that belong to this framework can identify the causal graph from non temporal data up to a Markov equivalence class [2]. Naturally, temporal data provide additional information regarding causation as it is common to suppose that a cause proceeds its effect, so for temporal data, the identification of causal graph is not restricted by the Markov equivalence class. However, in practice, in observational studies, the assumption that the cause proceeds the effect does not seems to hold, simply

2. At this point we do not need to define what is a Markov equivalence class. We just need the reader to know that constraint based approaches do not guarantee to find the true graph, instead they guarantee to find a graph that belongs to the same class of the true graph and which is known as the Markov equivalence class

because the lag between the cause and effect might be too small compared to the sampling frequency. This means that instantaneous causal relations are possible and they need to be taken into account in the process of discovering causal relations, but their symmetry with respect to time, restrict their discovery by constraint-based approaches to the Markov equivalence class. Thus, the best we can do regarding Problem 1 in the context of the constraint-based framework is to provide a partial solution which consists in discovering all lagged causal relations and only instantaneous relations that belong to the Markov equivalence class. So, another problem that we take into consideration in this thesis is the following:

Problem 2. *Given a maximal lag γ_{max} and an observational time series $X^1, \cdots, X^d$, infer the underlying summary causal graph corresponding to the true full time causal graph without restricting instantaneous relations to the Markov equivalence class.*

We stress that inferring a causal graph is an arduous task. Most algorithms, if not all, rely on assumptions that are often violated in practice. So causal discovery should be interpreted carefully. In this thesis, we consider it as a tool that assists the expert or the researcher and facilitates the search of the causal relations with respect to the given observational variables in a given system.

1.4 Thesis outline

As we saw in the previous Section, so far, artificial intelligence has dedicated its learning to finding correlations and associations from observational data, which makes it incapable of interpreting causes and effects, or understanding why these associations and correlations exist. This, in turn, limits artificial intelligence from being able to generalize its learning and to excel at tasks that involve explanation, imagination, reasoning, and planning. One way to tackle these problems would be to go beyond curve fitting and give machines the ability to discover and understand causal relations. Larry Wasserman once said that "using fancy tools like neural nets without understanding basic statistics is like doing a brain surgery before knowing how to use a band-aid", and analogically we think that using fancy tools like causal discovery methods or causal reasoning methods without understanding basic concept of causation is like doing a brain surgery without knowing why we need to make the surgery in the first place. Unfortunately, the concept of causation is not uniquely defined as it received different interpretations over time, from antiquity until today.

So to help researchers in the domain of artificial intelligence to gain some perspective regarding the concepts of causation, Chapter 2 offers a chronological review of the different philosophical views around causation [Assaad] [3].

Chapter 3 reviews and summarizes the most known existing techniques for discovering causal relations between time series and which are grouped into three families of methods: Granger causality, constraint-based approaches, and noise-based approaches. For each family of methods, it identifies the key assumptions, advantages and limitations [Assaad et al., a] [4].

Chapter 4 addresses the problem of learning a summary causal graph on time series with potentially different sampling rates. To do so, we first propose a new temporal mutual information measure defined on a window-based representation of time series. We then show how this measure relates to an *entropy reduction principle* that can be seen as a special case of the *probabilistic raising principle*. We finally combine these two ingredients in a PC-like algorithm to construct the summary causal graph [Assaad et al., b] [5]. Finally, we discuss how to extend our algorithm to handle hidden common causes and selection variables, how to infer a window causal graph from a summary causal graph, and how to apply the algorithm on sequences. The algorithm and its extensions are suited to discover lagged causal relations and instantaneous relations that fall in Markov equivalent class (for more details on Markov equivalence see Chapter 3).

Chapter 5 addresses the problem of learning instantaneous relations that do not fall in the Markov equivalent class, alongside lagged relations and instantaneous relations that fall in Markov equivalent class. To address the problem of learning such relations, we propose a hybrid method that combines the well-known constraint-based framework for causal graph discovery and the noise-based framework that gained much attention in recent years. Our method is divided into two steps. First, it uses a noise-based procedure to find the potential causes of each time series. Then, it uses a constraint-based approach to prune all unnecessary causes. A major contribution of this study is to extend the standard causation entropy measure to time series to handle lags bigger than one time step [Assaad

3. Karim Assaad, Emilie Devijver, Eric Gaussier and Ali Aït-Bachir. *A brief history of causality's principle. submitted*.

4. Karim Assaad, Emilie Devijver, Eric Gaussier and Ali Aït-Bachir. *A Survey on Causal Discovery for Time Series. submitted*.

5. Karim Assaad, Emilie Devijver, Eric Gaussier and Ali Aït-Bachir. *Entropy-based Discovery of Summary Causal Graphs in Time Series. submitted*.

et al., 2021] [6] and [Assaad et al., 2019] [7].

Chapter 6 is dedicated to numerical experiments. Our algorithms are evaluated on several datasets that shows both its efficacy and efficiency and compared with the state of art algorithms which are presented in Chapter 3.

Chapter 7 concludes the thesis by giving an overview of the results of the thesis and discusses the contributions with several remarks for future directions of research.

6. Karim Assaad, Emilie Devijver, Eric Gaussier and Ali Aït-Bachir. *A Mixed Noise and Constraint Based Approach to Causal Inference in Time Series*. Machine Learning and Knowledge Discovery in Databases. Research Track, pages 453–468, Cham, 2021. Springer International Publishing.

7. Karim Assaad, Emilie Devijver, Eric Gaussier and Ali Aït-Bachir. *Scaling Causal Inference in Additive Noise Models*. Volume 104 of Proceedings of Machine Learning Research, pages 22-23, Anchorage, Alaska, USA, 05 Aug 2019. PMLR.

Chapter 2

A brief history of causality's principles

> We think we have knowledge of a thing only when we have grasped its cause and we think we do not have knowledge of a thing until we have grasped its why, that is to say, its cause.
>
> Aristotle

2.1 Introduction

Throughout the history of philosophy and science, many great names explored (epistemologically or metaphysically) causal mechanisms, including Aristotle, Bacon, Galileo, Descartes, Hobbes, Hume, Kant, and Suppes, to name but a few. The studies conducted by these pioneers laid out diverse accounts of causation. Figure 2.1 presents what may be called the causation timeline: a brief non-exhaustive chronological overview [1] of these accounts and their most influential pioneers. The causation timeline is a story that began in antiquity, when Aristotle identified four types of causes that respond to the question *why*. The timeline then travels through the centuries while passing, among others, by Bacon, Galileo, and Descartes who criticize Aristotle's account while developing their own views about causation. For instance, by recognizing that nothing comes out of nothing, Descartes elaborated the causal adequacy principle, which is a perfect illustration of the deterministic way of thinking that emerged in the 18th century. After the so-called age of determinism, the causation timeline dives into Hume's regularity theory of causation, which is based on three main aspects: temporal priority, contiguity in time and space, and a constant conjunction between causes and effects. This theory influenced the majority of later thinkers. For example, Mill, inspired by Bacon and Hume, presented five methods to discover causal relations based on observations, while Mackie formulated the insufficient but non-redundant part of an unnecessary but sufficient (INUS) conditions. After Hume, for a certain period, only a few thinkers continued the search for causes and effects: Peirce (1839-1914) and Jastrow (1863-1944) used randomization in experiments in 1885 to eliminate bias and permit a valid test of significance, although the practice was not continued until Fisher developed randomized controlled trials in 1909. Wright introduced the first graphical tool (later extended to causal graphs) to address a causal question. However, aside from these few exceptions, most researchers became weary of seeking causes. Despite the undeniable importance of causality, causal notions seem strangely absent from the fundamental sciences. The main reasons are that causal relations are perceived to be too vague for a mathematically precise science and that causation is asymmetric, whereas classical mathematical operations are symmetric. Several scientists even argued that causality should be completely removed from the philosophical and scientific vocabulary. However, many scientific discoveries confirmed that such claims are absurd, as causation is one of the main pillars

1. For a complementary overview, see Drouet [2007]

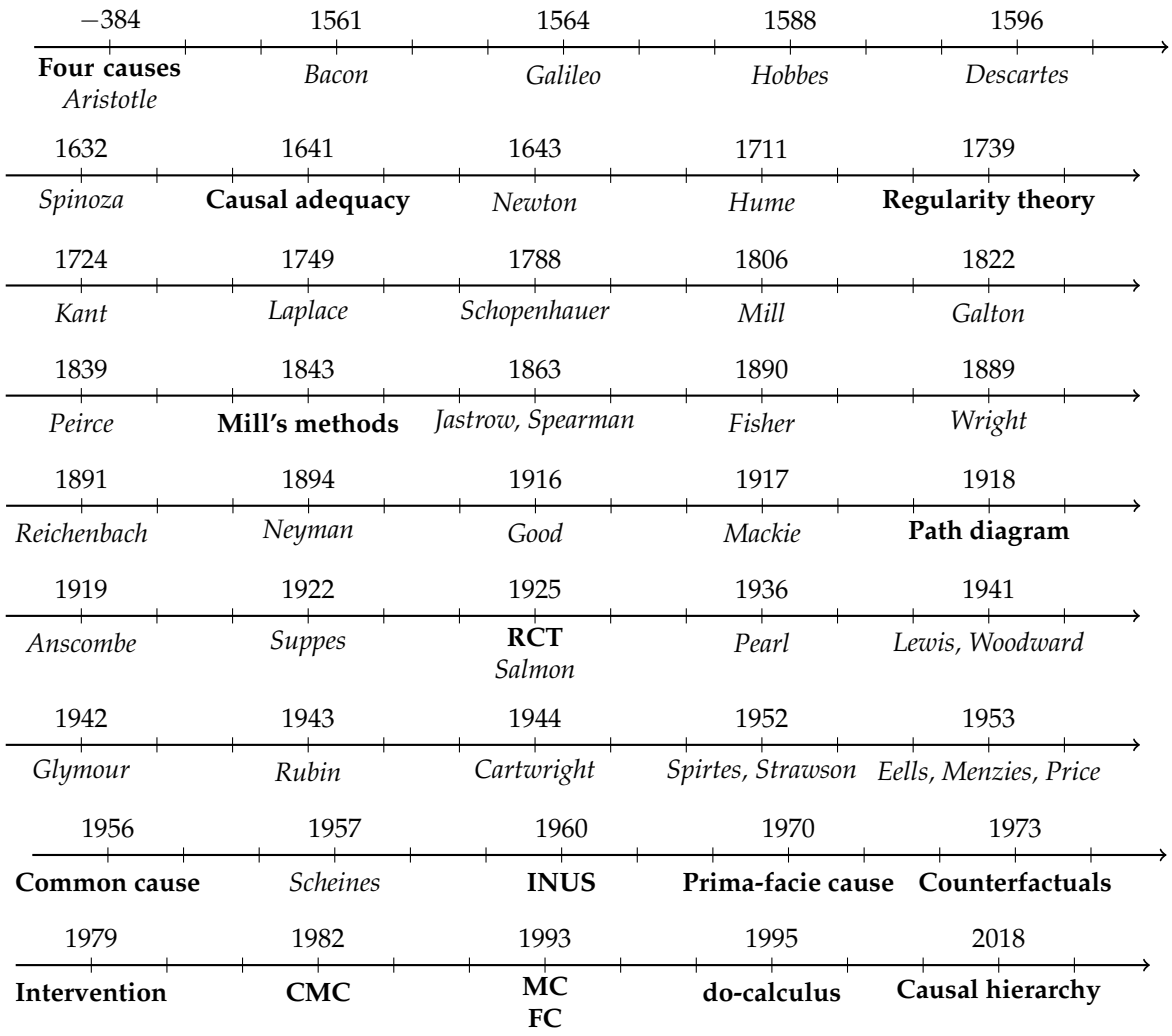


Figure 2.1 – Causation timeline. Accounts are in bold, and pioneers in italics (accounts are positioned by the date of their appearance and pioneers are positioned by their date of birth). RCT: randomized controlled trials; INUS: insufficient but non-redundant part of an unnecessary but sufficient condition; CMC: causal Markov condition; MC: Minimality Condition; FC: Faithfulness Condition.

of acquiring scientific knowledge alongside probability. Accordingly, the causation timeline proceeds with the probability theory of causation in which Reinchinbach and Suppes drew on the probability raising principle and respectively elaborated the common cause principle and the prima-facie cause. The timeline then turns to the decision-making theories of causation, namely the counterfactual theory of causation that allows us to imagine potential worlds, and the intervention theory of causation. Finally, the timeline makes a special tribute to the do-calculus, the causal Markov condition (a generalization of the common cause principle), the minimality condition, faithfulness, and the causal hierarchy, which rendered the concept of causation accessible and made it clearer.

This chapter describes several prominent versions of causality theories advocated by philosophers and scientists alike, drawing attention to their limitations.

2.2 Principle of four causes

Aristotle (384-322 BCE) viewed science through the lens of observation; as a result, he was convinced that the truth can only be sought via observations. Aristotle formulated a causal investigation, which, in his view, is the pursuit of a response to the question *why*. He claimed that "we do not think we have knowledge of each thing until we have grasped the why of it, which is to grasp its cause" Aristotle and Reeve [2018]. And since the word *cause* has several meanings, it follows that the same thing can have several types of causes (for more details, see Section 2.9 below). Aristotle thus recognized four types of causes that can be given in response to the *why* question Aristotle and Reeve [2018]:

— material cause ("that out of which"): the material of which something is made. For example, ivory is a material cause of a billiard ball;

— formal cause ("what a thing is"): the form or properties of something that makes it what it is. For example, roundness is a formal cause of a billiard ball;

— efficient cause ("primary source of the change or rest"): the proximal mechanism that provokes something to change. For example, the billiard ball manufacturer is an efficient cause of a billiard ball;

— final cause ("the end, that for the sake of which a thing is done"): the goal of something. For example, being used in billiards games is a final cause of a billiard ball.

According to Aristotle, everything in the empirical world, including the empirical world itself, is a subject to these four causes. And he believed that the final cause is the most significant explanation of everything, since the final cause reflects the final *why*. Why does the billiard ball exist? It exists to play billiard. He also found a clear association between causation and necessity. When the cause acts, the effect is acted on out of necessity.

The identification of the final cause is subject to much debate, since it necessitates an external judge. People perceive things in different ways, and perception is subjective, not universal. Thus, while one person may identify being used in billiards games as the final cause of a billiard ball, another may consider it to be the act of moving another billiard ball. Furthermore, Aristotle conceived a world in which everything has a goal as its final cause. Consider the case of a stone that fell from a roof onto someone's head, leading to their death. Did the stone have the goal of killing this individual? Or was it a mere accident? Based on Darwin's theory of natural selection, we now know that nothing has a true purpose, which means that nothing has a goal. Things come into existence by chance, and so most philosophers now agree that the final cause is superfluous [Mumford and Anjum, 2013].

Aristotle's principle of four causes seems plausible at first glance, but it turned out to be flawed. It was consistent with human logic at the time of Aristotle, but after the scientific discoveries of the last centuries, it has become outdated. However, it is still interesting that it was applied to conceive the empirical world, making a connection between causality and necessity.

Despite their agreement regarding this connection, Aristotle and other Greek philosophers, including Plato, left room for free will. Their theory thus suited the many religions that subsequently adopted it: given the existence of free will, the gods are not responsible for all the evil in the world. If humans are free to make decisions, then they have moral responsibilities that make them accountable for their actions. This explains why the free will theory was more acceptable and attractive than the one that excluded it. This idea dominated most scientific views until the time of Bacon, Galileo, Descartes, and Spinoza.

2.3 Determinism

Since the beginning of time, humans have tried to find patterns in nature. They accidentally discovered how to control fire and have since used this skill in their daily lives. They thought that they could count on knowl-

edge acquired in the past with a high degree of certainty: they thus believed that every time they needed fire, they would be able to create it [Harari, 2015]. For the most part of our history, without even knowing it, we understood deterministic causality mostly as an issue of regularity, which reflects the determination of effects by their causes. This leads us to think that causality occurs in the realm of determinism based on the idea that the past determines the future. Determinism can be defined as follows:

Definition 6 (Determinism [Hoefer, 2016])**.** *The world is governed by determinism if and only if, given a specific way that things are at time t, the way things go thereafter is fixed as a matter of natural law.*

Among the first to defend determinism were Leucippus (5th century BCE) and his associate Democritus (460 -370 BCE), the first to theorize the existence of atoms [Taylor, 1998]. To our knowledge, they were convinced that nothing happens by chance and that everything has a cause. Yet their point of view remained overshadowed by free will until the 17th century.

Francis Bacon's (1561-1626) main goal was to gain control over Nature. He considered the effect and its cause to be interchangeable: whenever the cause is present, the effect is also present, and whenever the effect is present, its cause is present too. He looked for causes that are necessary and sufficient for their effects. He agreed with Aristotle that causality can be accessed through empirical studies, although he disagreed with him on the degree of patience while conducting these studies. For Bacon, observations are necessary to uncover the concept of causes, but they are not sufficient. According to Bacon, discovering causal knowledge would entail studying the essence of events: if a first event causes a second event, then there is a logical relation that can be discovered by studying the essence of the first and the second. Furthermore, Bacon rejected Aristotle's principle of four causes on the grounds that the distribution into material, formal, efficient, and final causes is inadequate, since physics deals with material and efficient causes, whereas metaphysics deals with formal and final causes. In addition, he also believed that Aristotle's emphasis on final causes only serves to promote verbal disputes and slow down the progress of science, so he rejected it and banished all references to a divine purpose in the universe from scientific and philosophical discourses, and defined science as a search of causes. He systematically articulated a method to help find causes by scientifically observing regularities in the world. His method consists of systematically making and recording observations about the phenomenon of interest (natural and controlled), then classifying the observations according to a conceptual scheme, and

finally, by means of eliminating false causal hypotheses, inferring the true causal claim that governs the observed phenomenon. To test a potential cause-effect relation, Bacon suggested setting up experiments to manipulate nature and attempted to prove that the potential cause-effect relation is wrong. However, he never stated that his method is perfect, as he explicitly acknowledged that all methods of causal inference are fallible. Given that many different causes might be missed by our perception, Bacon stressed that these experiments must be consistently repeated before the truth can be known where possible. With Galileo, he forged the way to conduct controlled experiments as a means of gaining knowledge about nature by controlling different factors while fixing others.

Galileo Galilei (1564-1642) focused more on *how* than on *why* and viewed the explanation of causality only in the motion of the atoms. Furthermore, he identified God as the first atom: the force that constructed the world out of nothing. He thus rejected the final cause but added a first efficient cause, which led to the existence of nature and humans in the form of an immutable mathematical system. As a result, mathematical methods can access scientific knowledge independently of the first cause [Burtt, 1926].

René Descartes (1596-1650) distinguished between bodies and spirits. He accepted free will as part of the spiritual realm but denied its existence in the physical world, which was instead constrained by natural laws. Unlike Aristotle, Descartes thought that everything was inert and entirely rejected Aristotle's idea of final causes. For Descartes, people could not derive any explanations of nature that come from God, since they are unable to understand God's plan. His idea depended on the causal adequacy principle.

Definition 7 (Causal adequacy principle [Descartes et al., 1983])**.** *There must be at least as much reality in the cause as in the effect of that cause.*

By pointing out that there is nothing in the effect that is not in the cause, he concluded that something cannot come from nothing. As he put it: "if we admit that there is something in the effect that was not previously present in the cause, we shall also have to admit that this something was produced by nothing" [Descartes, 1985]. Consequently, he devised the transference model of causality, which states that when a first event causes a second event, a property of the first is communicated to the second, which in Aristotle's vocabulary means that the first event is an efficient cause of the second event. At present, Descartes' theory of separation between spirit and body (dualism) is outdated. We now know that consciousness is a function of brain activity and that the brain follows rigid laws of nature. For example, to understand why a person makes certain choices, we simply have

to look at how the brain functions, not at some spirit bubble connected to the body [Dehaene, 2014]. Nevertheless, aspects of Descartes' concept of causation can still be valid.

Thomas Hobbes (1588-1679) rejected the formal and final causes of Aristotle and introduced the notion of entire causes, which are always interpreted in terms of particular motions of particular bodies. Technically, the entire causes are the combination of material and efficient causes: the material cause represents the receptor of the agent's activity and the efficient cause represents the properties required by the agent for the production of the effect. Hobbes also associated necessity with the entire cause, which involves both the agent and the patient. Moreover, Hobbes argued that a sufficient cause should occur simultaneously with its effect: "in whatsoever instant the cause is entire, in the same instant the effect is produced. For if it be not produced, something is still wanting, which is requisite for the production of it; and therefore the cause was not entire, as was supposed." [Hobbes, 1656].

In 1677, a new definition of human freedom distinct from free will emerged. According to Baruch Spinoza (1632-1677), being free does not mean being able to act as we please, but rather recognizing *why* we act as we do given the nature of reality. Being free means understanding what causes our actions, i.e., understanding the causal mechanisms. So human freedom can be achieved by knowing the causes that determine our desires and affections. He defined servitude as the state of ignorance of the causes that determined our desires. He described freedom as the state of someone who is capable, through reason and knowledge, of understanding the causes that determine his desires. So freedom is not an exemption from causality, and it does not contradict determinism. However, Spinoza supported Aristotle on the relation between causation and necessity, writing "From a definite cause an effect follows; and if no definite cause be granted, it is impossible that an effect can follow." (Axiom III) [Spinoza, 1677]. In other words, given a cause, the effect follows out of necessity, and without it, the effect does not follow. In addition, he linked necessity to determinism "In nature there is nothing contingent, but all things have been determined from the necessity of the divine nature to exist and produce an effect in a certain way." [Spinoza, 1677, Curley, 1992].

Isaac Newton's (1643-1727) mathematical approach to natural philosophy is often deemed to have shifted the focus of science from causal explanation to pure description. For instance, he offered no assumption about the cause of gravity, but simply identified the phenomenon of gravity and described it with mathematical precision. In this sense, one could say that Newton adopted the formal causes of Aristotle as the only true objects

of science. However, in his interpretation and application of the laws of motion, his ideas of causality differed. Newton perceived a world consisting of material bodies at rest or in motion and interacting according to his three famous laws of motion [Isaac Newton et al., 1999]:

— every object in a state of uniform motion will remain in that state of motion unless an external force acts on it;

— force equals mass times acceleration ($F = m\,a$);

— for every action there is an equal and opposite reaction.

By causes, he meant the forces impressed on a body, which drives it to move differently than it would have done without them [Rynasiewicz, 2008]. Regarding the laws of motion, this means that Newton thought in terms of efficient causes.

During this period, mechanical principles came to the fore. Conservation laws tell us that the quantity of energy in a system is stable, and subsequently, the concept of causality was reduced to a transfer of energy. It is true that the transfer of energy implies the existence of causal relations in the system: for example, when billiard ball A with an initial momentum hits billiard ball B at rest, ball A will transfer some of its momentum to ball B, thus forcing it to move. However, this does not inform us about the direction of the cause. The transfer of energy can be seen in two ways; according to Newton's third law, there must be an equal but opposite force from the second object that cancels the applied force: hence, ball B at rest transfers its zero energy to the moving ball A, thus forcing it to stop. In addition, causality does not necessarily imply the existence of a transfer of energy, because not every cause-effect relation can be explained through energy transfer. For example, when Louis touched Elsa's hand, she blushed. His touch caused the blushing, but where is the transfer of energy in this scenario? Some might argue that the touching implies some physical explanation relating to the mechanical principles. In another example, Elsa who lived in France blushed when Louis who was on vacation in Germany texted her and said that he loved her. Elsa's touching the phone to read the message had nothing to do with the blushing nor did her reading the structured alphabets. The idea of Louis loving her caused her to blush. Regardless of whether the news came from a text, a call, or the touch of his hand, she would have blushed. And this could not be explained by mechanical principles (to understand how causation can deal with such an example, see Section 2.7). However, rejecting causality as a concept of energy transfer does not imply the rejection of causal determinism, which was emphasized by Pierre-Simon de Laplace (1749-1827), who observed that the world follows fundamental laws, and thus

by knowing these laws, everything in the world is determined (Laplace demon) [Laplace, 1814], i.e., every event is determined by an antecedent event along with conditions concordant with the laws of nature.

Causation thus remained heavily dependent on determinism until the analysis of David Hume who dismissed the connection between causation and necessity.

2.4 Hume's regularity theory

Impressed by Newton's work and experimental philosophy, David Hume (1711-1776) believed that the foundation of our knowledge derives from our experience of the world. He thus endeavored to understand how the mind works through observation [Hume, 1738, 1748]. He supposed that contiguity in time and space are essential for cause-effect relations. For the moving billiard ball A to be the cause of the movement of another billiard ball B after colliding with it, ball A and ball B need to touch, which means being contiguous in time and space. He argued that no direct causes can operate from a distance. Even if two distant objects seem to be causally related in nature, they are usually linked to causal chains. Furthermore, he challenged Hobbes by opposing the idea of contemporary causal relations, instead claiming that all causes, even sufficient ones, temporally precede their effects in which we formalize as follows:

Definition 8 (Temporal Priority)**.** *If X_t^p is the cause of $X_{t'}^q$, then $t < t'$.*

Returning to the example of the billiard ball, the movement of ball B (i.e., effect) happened after ball A touched it (i.e., cause). However, in his analysis, noting that causation cannot simply be an affair of contiguity and temporal priority, Hume added regularity to the equation. His refusal to give conceptual status to unobserved phenomena led him to base his conclusions about causation [2] solely on the observation of prior associations between variables. Thus, according to this interpretation, causation cannot be represented by single events; it can only arise through constantly conjoined events (in time and space). Hume agreed that causes necessitate their effect, but in his view, necessity is something imposed by humans. Necessity arises from within the human mind, conditioned by the observation of regularities in nature to form an expectation of the effect when

2. There are many interpretations of Hume's work; in this section, we focus on the most widespread. For other interpretations, see Strawson [2014].

the cause is present. This means that causality is a philosophical disposition; it does not exist in nature [3], but only in our mind, and the sole way to detect it is through regularities. When we repeatedly observe an event followed by a second event, we will believe that first event causes the second, even though the only evidence is our observation of the same event multiple times. Returning to the billiard ball example, we need to repeatedly see the movement of the first ball, followed by the collision of the two balls and the movement of the second ball before inferring the causal relation between them. But why does Hume not suppose that regularity can imply necessity? The answer can be derived from his famous theory of induction which tells us that the assumption that a regularity will continue to be repeated in the future is circular and based on the principle of uniformity in nature which is not a priori true. More formally, Hume defined causation as follows:

Definition 9 (Regularity theory of causation [Hume, 1748])**.** *We may define a cause to be an object precedent and contiguous to another, and where all the objects resembling the former are placed in like relations of precedency and contiguity to those objects that resemble the latter.*

Immanuel Kant (1724-1804) attempted to respond to Hume [Kant, 1997]. He believed that we have innate ideas, and from birth, our mind is filled with certain concepts. For Kant, objective events are not simply given; instead, they are formed by the organizing activity of the mind and especially by the imposition of the principle of causality to phenomena. Consequently, the principle of causality is, for Kant, an a priori principle; causality occurs in the mind with the aim to acquire knowledge, independently of the process of observation. Kant agreed with Hume that the necessity of causal sequences cannot be observed in the sequences themselves, as they are rather projected onto the world by the mind. However, he viewed all sequences to be a consequence, whereas Hume considered all consequences to be a mere sequence.

As Arthur Schopenhauer (1788–1860) noted (in his criticism of Hume and Kant), it is absurd to conceive all sequences as consequences: the tones of a musical composition follow each other in a certain objective order, yet it would be absurd to say that they follow each other according to the law of causality [Schopenhauer, 1889]; the night always follows the day, but it would be absurd to say that they follow each other according to the law

3. Other interpretations suggest that Hume believed that humans are capable of forming a causal concept in the mind but are incapable of accessing the true causes that exist in nature.

	Pasta	Salad	Seafood	Sick
Elsa	No	Yes	Yes	Yes
Louis	Yes	No	Yes	Yes
Paul	Yes	Yes	Yes	Yes

(a)

	Pasta	Salad	Seafood	Sick
Paul	Yes	Yes	Yes	Yes
Germaine	Yes	Yes	No	No

(b)

Table 2.1 – Illustration of Mill's methods. We report the food eaten at dinner by a group (a) and a group (b).

of causality. Apparently, regularity can exist without causality. And likewise, it appears that causality can exist without regularity. This is the case when one event causes another to happen without this particular (singular) sequence of events falling under a regularity. In nature, even though some events occur only once, we are capable of establishing their cause-effect relations: the big bang caused the existence of the universe; the design flaws of the RBMK reactor and the incompetence of worker caused Chernobyl's nuclear reactor disaster; and the assassination of Archduke Franz Ferdinand was the main immediate cause of World War I. However, there are debates surrounding these types of causation. Some call this singular causality or actual causality, which is more fundamental than general causality, which is discussed here. Actual causality focuses on specific events, whereas general causality arises from many instances of actual causality (for more details about actual causality see Section 2.7 or [Cartwright, 1989, Drouet, 2007, Halpern, 2016]). Hume did not consider the metaphysics of causality, but like Kant, he ended up with a loose notion of causality. His theory was the subject of much debate and continued to be amended.

John Stuart Mill (1806-1873) defended most of Hume's regularity view of causality except the part concerning necessity. Mill claimed that an effect invariably follows from the cause and that the cause should be taken to be the whole conjunction of the conditions that are sufficient and necessary for the effect. Inspired by Bacon's logic, Mill also defined four methods to discover causal regularities [Mill, 1843]. The first is defined as follows:

Method 1 (agreement)**.** *To find a cause of an effect, one should find a single factor common to several occurrences of the effect. If the common factor is present*

in all cases where the effect is present, we can infer it to be a cause of the effect.

For illustrative purposes, let us consider the example given in Table 2.1 (a). Here we have a group of people dining together. The next day all of them have food poisoning. Assuming we know in advance that the food eaten at dinner was the source of the food poisoning, we can apply the method of agreement to identify the exact cause. Looking at Table 2.1 (a), the only common food (common factor) is seafood which stands out as the cause of the food poisoning. Mill's second method is defined as follows:

Method 2 (difference)**.** *To find a cause of an effect, one should find a single factor present in all occurrences of an effect and absent from all occurrence of the absence of the effect.*

Now let us look at Table 2.1 (b). Here two people, Paul and Germaine, dined together, but only Paul got food poisoning. As can be seen, the only food that Paul consumed but Germaine did not is seafood, which suggests that seafood is the cause of the food poisoning. Mill's third method is a combination of the first two methods:

Method 3 (agreement and difference)**.** *To find a cause of an effect, one should find a single factor that is present in multiple occurrences in which the effect in question is present and absent from multiple occurrences in which the effect is absent.*

To illustrate this method, consider a combination of Table 2.1 (a and b) with four people dinning together. As can be seen, everyone who consumed seafood got sick (agreement), while all those who did not eat it did not get sick (difference). Mill's fourth method is a deductive method defined as follows:

Method 4 (residue)**.** *To find a cause of an effect, one should separate from a group of causally connected conditions and events those that are known to be cause-effect relations, leaving the required causal connection as the residue. We identify what is left as the cause of the remaining effect.*

Suppose that there was a bottle of wine on the dinning table, and we want to know who finished the bottle. Consider that Elsa is pregnant and cannot drink alcohol, while Louis and Germaine had their fill after the first glass was served at the beginning of the dinner. Here we can deduce that Paul is the residue, inferring that he finished the rest of the bottle. As seen here, Mill's first four methods examine cause and effect in qualitative terms. An effect either occurs or does not occur. A cause is either absent or present. Taking into account the quantitative terms, Mill introduced his fifth method for discovering causes:

Method 5 (concomitant variation)**.** *To find a cause of an effect, one should identify a causal connection between two conditions by matching variations in one condition with variations in another. So, if across a range of situations that lead to a certain effect, we find that a certain property of the effect varies with a factor that is common to those situations, we can infer that factor to be the cause.*

Now suppose that four people dined together (Elsa, Louis, Paul, and Germaine): Elsa ate a whole dish of seafood and is violently sick; Louis had half a dish of seafood and is fairly ill; Paul had one bite of seafood and felt a little queasy; and Germaine ate no seafood and did not get sick. It therefore appears that there is a direct association between the degree of seafood consumption and the severity of sickness. By the method of concomitant variation, seafood is the cause of the poisoning. Mill's methods can only help us identify causes when the potential candidates are already known. Thus, they are not capable of examining all possible interactions, which make them erroneously judge the potential cause of an event. Additionally, with these methods, we can only look for a single cause. They presuppose that from the list of factors under consideration, only one is the unique cause of the effect.

More recently, John Leslie Mackie (1917-1981) updated Hume's theory to allow for causes and effects with multiple components. According to Mackie, it is incorrect to say that striking a match causes it to light. Striking the match was not the only cause for the match to light: oxygen was present, the match was dry, there was little wind, and so on. An event can be necessary for another event to occur, but alone it may be insufficient. And the entire set of conditions (striking the match, oxygen, dry match, no wind, etc.) may not be necessary, because there may be another set of conditions that would light the match. Based on these insights, Mackie introduced the INUS conditions for identifying causes that are an insufficient but non-redundant part of an unnecessary but sufficient condition [Mackie, 1980]. By insufficient, he meant that a cause (e.g., striking a match) can not produce the effect without the presence of other factors (e.g., oxygen, dry match, no wind). By non-redundant, he meant that the cause should add something uniquely different from what the other factor adds. By unnecessary but sufficient condition, he meant that the cause should belong to a set of factors that can, if all active, light the match, although they are unnecessary in the sense another set of factors may produce the same effect.

Definition 10 (INUS conditions [Mackie, 1980])**.** *X^p is a cause of event X^q if and only if:*

— *there is a set of events $X^{\boldsymbol{R}}$ that is sufficient but not necessary for X^q;*

— X^{R} *is minimal, i.e., no subset of* X^{R} *is sufficient for* X^{q}*;*

— X^{p} *is included in* X^{R}*.*

In other words, Mackie searched for causes that are only valid under specific circumstances. In this respect, his approach is similar to Rothman's sufficient component causal model [Rothman, 1976].

In the traditional interpretation of Hume, causation is mind-dependent and there is no more to it than regularity. Hence, causation does not exist in an individual case. According to this interpretation, Hume may be considered to be a reductive realist, who jumped from the claim that we have no idea of necessary connection (epistemological claim) to the claim that there is no necessary connection (metaphysical claim). Galen Strawson (1952-present) argues that Hume would not have made a metaphysical claim, because such a claim would go against his skeptical belief that we are ignorant of the world, independently of our ideas about it. On this basis, Strawson gives a new interpretation that supports only Hume's first claim (epistemological) but not his second (metaphysical) [Strawson, 2014]. The new interpretation disagrees that causation is reductive to regularity, and consider it real and mind-independent. This suggests that causation is a real force. When the first billiard ball hits the second, there is a power, a force, a necessary connection that is independent of anything else that happened. Necessary connections are not representative; they are theoretical ideas. Thus, causation is the theoretical idea that lies behind regularity. Causation is an intrinsic [4] (not an extrinsic [5]) relation. This interpretation is interesting, although it is not accepted by all philosophers. Some argue that even with this interpretation, the only possible evidence for causal relations can come from regularities. For this reason, there is perhaps little point in insisting that causation is something different [Beebee, 2009].

Going back to the initial interpretation, Hume considered that causal representations are important, as we cannot advance very far in the world without them. But this does not mean that we must believe in a richly metaphysical idea of causal powers that produce or bring about causal regularities. Hence, with Hume's skepticism and Newton's reductionist view about causation, it was almost impossible to prevent what would happen next.

4. Meaning an extremely important and inherent characteristic of a person or thing.
5. Meaning from the outside or not related to something.

2.5 Looking the other way

Until the 20th century, causality was still associated with necessity and determinism, and since determinism was no longer an object of science, researchers began to lose interest in causality. This was the main reason why philosophers looked the other way. They also argued that causality was neither an element of reality nor an element of our reflection on reality, which is the first subject of science, and they therefore concluded that causality is not a subject of science. It is rather an epistemological obstacle and a superfluous and undesirable concept. Scientists also tended to refute the concept of causation, claiming that if it is unobservable, how can it be measured? And if it is unmeasurable, how can it be put into equations?

Shortly after Charles Darwin expounded his theory of evolution, his cousin Sir Francis Galton (1822-1911) began to draw out the implications of these ideas: if we have evolved, then mental faculties like intelligence must be hereditary. He also argued that if such faculties are hereditary, and some people have it in a greater degree than others, then our ability to choose our fate is not free but rather depends on our biological inheritance, which means that the laws of inheritance are the causes of everything that we do. In 1877, Galton started to express the need for causal claims and went on to pursue causation. He ended up discovering the concept of correlation, which Karl Pearson (1857-1936) later used to derive a formula for the slope of the regression line, known as the correlation coefficient. This discovery convinced Pearson that the concept of cause and effect is unscientific, since it is neither mathematically clear not precise. He made his point abundantly clear when he wrote: "Beyond such discarded fundamentals as matter and force lies still another fetish amidst the inscrutable arcana of even modern science, namely, the category of cause and effect" [Pearson, 1911]. Moreover, he claimed: "That a certain sequence has occurred and reoccurred in the past is a matter of experience to which we give expression in the concept causation; that it will continue to recur in the future is a matter of belief to which we give expression in the concept probability. Science in no case can demonstrate any inherent necessity in a sequence, nor prove with absolute certainty that it must be repeated." [Pearson, 1900]. Thus, Pearson agreed with Hume that the concept of cause and effect was nothing more than an affair of regularities, arguing that the entire concept should be abandoned in favor of his revolutionary correlation coefficient that is capable of summarizing regularities. Furthermore, he stated that "the ultimate scientific statement of description of the relation between two things can always be thrown back upon such a contingency table..." [Pearson, 1911]. Thus, Pearson categorically

denied the concept of causal relation beyond correlation and stated that all the knowledge required can be found in the contingency table. Since he belonged to the philosophical school of positivism, it is unsurprising that he defined science purely in terms of actual data [Pearl and Mackenzie, 2018]. And his enthusiasm and dominance made it impossible for his colleagues to contradict his ideas [Yule, 1936].

At that time, the field of statistics was not alone in opposing causation. The basic laws of physics (distinct from higher-level statistical generalizations such as the laws of thermodynamics) also appear to be time symmetric. If a certain process is allowed under the basic laws of physics, a video of the same process played backward will also depict a process that is permitted by the laws. Since cause and effect play no fundamental role in physics, some scientists argued that causality should be completely removed from the philosophical and scientific vocabulary. For example, in 1912, Bertrand Russell notoriously proclaimed that "the law of causality, like much that passes muster among philosophers, is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm." [Russell, 1912].

To sum up, most of the scientific community considered that causality was not worthy of being treated as a scientific object. They even suggested banning the term *causality* from scientific discourse. However, the second law of thermodynamics rectified this problem, as it proved the existence of the arrow of time by demonstrating that not all processes are reversible, that is, everything evolves with time to have total entropy (i.e., a maximum disorder).

2.6 Probabilistic theory of causation

Heisenberg's uncertainty principle states that it is impossible to define with precision the behavior of a particle due to the presence of an observer. Many concluded from this principle that a particle does not know the position that it will occupy until an observer appears. But recent discoveries showed that it is absurd to think so. One of the most known counterexamples showing the absurdity of this idea is Schrodinger's cat. In fact, the principle never mentioned a non-deterministic world; it simply stated that it is not determinable according to the observer. It therefore seems that the world is deterministic but indeterminable. And the explanation resides in the chaos theory, which tells us that small alterations can have severe consequences (i.e., small causes can have huge effects), which makes the world indeterminable. Assuming that everything is determined, but ev-

erything is indeterminable, the best that we can aim for is to infer with a degree of certainty. Thus, to study causal inference, the notion of probability is needed. In addition, interpreting causation as deterministic relations may lead to undesired results (mainly in systems without access to all variables), because some cause-effect relations might not always occur, although the cause increases the chance of the effect happening. Smoking causes cancer, but not every person who smokes will necessarily develop cancer. Thus, while the deterministic approach requires a sufficient or complete cause, the probabilistic approach requires an increase in the probability of the effect given the cause.

Definition 11 (Probability raising [Hitchcock, 2018]). *If X^p is a cause of X^q, then* $\Pr(X^q \mid X^p) > \Pr(X^q)$.

Hans Reichenbach (1891-1953) was one of the first to connect causality with probability. Indeed, correlation does not imply causation (see Freedman [1999] for an illustration), but Reichenbach noticed that correlation might give a certain indication about causal relations. In his view, probabilistic correlations are ultimately derived from causal relationships. Two events X^p and X^q are correlated (i.e., $\Pr(X^p \cap X^q) > \Pr(X^p)\Pr(X^q)$), because one of these events causes the other, or because an event X^r commonly causes X^p and X^q, as shown in Figure 2.2a, or because X^p causes event X^r (an intermediate), which in turn causes event X^q, as shown in Figure 2.2b. Hence, Reinchenbach introduced the common cause principle,

(a) (b)

Figure 2.2 – Two causal structures in which X^p is screened off from X^q by X^r.

which states that a correlation between two events X^p and X^q indicates that a causal relation exists between X^p and X^q ($X^p \to X^q$ or $X^q \to X^p$) or that X^p and X^q have a common cause X^r ($X^p \leftarrow X^r \to X^q$) [Reichenbach, 1956].

Definition 12 (Common cause principle [Reichenbach, 1956]). *X^r is a common cause of X^p and X^q if*

— $\Pr(X^p X^q) > \Pr(X^p)\Pr(X^q)$;

— $\Pr(X^p X^q \mid X^r) = \Pr(X^p \mid X^r)\Pr(X^q \mid X^r)$;

— $\Pr(X^p X^q \mid\sim X^r) = \Pr(X^p \mid\sim X^r) \Pr(X^q \mid\sim X^r)$;

— $\Pr(X^p \mid X^r) > \Pr(X^p \mid\sim X^r)$;

— $\Pr(X^q \mid X^r) > \Pr(X^q \mid\sim X^r)$.

The first condition ensures that X^p and X^q are unconditionally dependent. The next two conditions express that when the presence (or absence) of the common cause is taken into account (by conditioning), the correlated events X^p and X^q are rendered probabilistically independent. In Reichenbach's terminology, X^r screens X^p off from X^q. The last two conditions describe the probability raising of the effect given their causes. Reichenbach left open the question as to whether the causal relations occurs at the population or individual level (general or actual causality, respectively). His work inspired many thinkers who followed in his wake. For example, Irving J. Good (1916-2009) introduced a causal calculus based on common causes to construct causal nets and attempted a quantitative characterization of the strengths of these nets [Good, 1961]. Wesley C. Salmon (1925-2001) used the combination of screening off and the common cause principle to offer a solid basis for causal explanation, although he later revised his position by recognizing another type of common cause, known as the interactive fork [Salmon, 1984]. The interactive fork depicts causal interactions whose effects remain correlated even in the presence of the common cause, which contradicts the common cause principle. However, Peter Spirtes (1952-present), Clark Glymour (1942-present), and Richard Scheines (1957-present) studied Salmon's position and postulated that interactive forks do not exist [Spirtes et al., 2000] (at least in the macroscopic world). As can be seen, Reichenbach avoided references to time in his analysis of causality. However, this avoidance was not unintentional: he excluded it, because he wanted to analyze time in terms of causality. He nevertheless failed to achieve this goal, because, as he postulated, the intermediate causes would also screen off two dependent variables. Consequently, Spirtes, Glymour, and Scheines showed that Reichenbach erroneously focused on common causes and intermediate causes while disregarding unshielded colliders [6], which seems to be the key to determining the direction of the cause between two variables [Spirtes et al., 2000, Pearl, 2000]. An unshielded collider gives rise to distinctive probabilistic relationships compared to a common cause and intermediate cause. A collider creates dependency through conditioning, while a common cause and an

6. The treatment of colliders can be traced back to Arthur Cecil Pigou [Pigou, 2017]. A collider is unshielded if there is no direct causal relation between the direct causes of the collider variable. Otherwise, it is shielded.

intermediate cause remove the dependency. An example of an unshielded collider is provided in Figure 2.3.

Definition 13 (Unshielded Collider)**.** *X^p and X^q cause X^r if $X^p \perp\!\!\!\perp X^q$ and $X^p \not\perp\!\!\!\perp X^q \mid X^r$.*

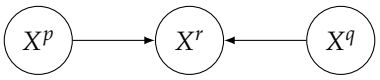


Figure 2.3 – Unshielded collider

After Reichenbach, Patrick Suppes (1922-2014) presented one of the most known theories of probabilistic causality. He agreed with Hume, that causes must, by definition, precede their effects in time, and in this way, he opposed Reichenbach's approach. Nevertheless, like Reichenbach, Suppes did not attempt to introduce any quantitative measures of causal strength, instead framing his definitions in terms of measures of probability. He noticed that the probability raising of $\Pr(X^q \mid X^p)$ is necessary but insufficient to claim that X^p causes X^q. Thus, he started by defining a prima-facie cause that encodes temporal priority and probability raising along with the stipulation that the cause has a nonzero probability.

Definition 14 (Prima facie cause [Suppes, 1970])**.** *An event $X^p_{t'}$ is a prima facie cause of X^q_t when the following conditions hold:*

— $t' < t$;

— $\Pr(X^p_{t'}) > 0$;

— $\Pr(X^q_t | X^p_{t'}) > \Pr(X^q_t)$.

The first condition in the definition is the temporal priority introduced by Hume (see Section 2.4). The second condition is a safe requirement that the prima facie cause has some nonzero probability of occurring. The third condition requires for the conditional probability of the effect, given the prima facie cause, to be greater than the probability of the unconditioned effect (probability raising). A prima facie cause does not need to be a genuine cause. And the above definition alone cannot differentiate when it is a genuine or a spurious cause. To strengthen his theory, Suppes offered another definition that allows for the detection of spurious causes:

Definition 15 (Spurious cause [Suppes, 1970])**.** *An event X^p_{t-i} is a spurious cause of X^q_t if and only if X^p_{t-i} is a prima facie cause of X^q_t and there is an event X^r_{t-i-j} that occurs prior to than the prima facie cause and the effect such that*

— $\Pr(X^p_{t-i}, X^r_{t-i-j}) > 0$;

— $\Pr(X^q_t \mid X^p_{t-i}, X^r_{t-i-j}) = \Pr(X^q_t \mid X^r_{t-i-j})$;

— $\Pr(X^q_t \mid X^p_{t-i}, X^r_{t-i-j}) \geq \Pr(X^q_t \mid X^p_{t-i})$.

According to this definition, X^p is a spurious cause of event X^q if it is a prima facie cause of X^q and it is screened off from X^q by an event X^r (or partition of events $X^{\mathbf{R}}$) that precede X^p. And thus a prima facie cause is genuine if it is not spurious, that is, if no earlier event undermines its effectiveness. Suppes also introduced another version of this definition by dropping the last condition. Under this setting, a prima facie cause is genuine, if it is not spurious, that is, given prior events, the knowledge of the prima facie cause remains predicatively informative.

Definition 16 (Indirect cause [Suppes, 1970])**.** *An event X^p_{t-i-j} is an indirect cause of X^q_t if and only if X^p_{t-i-j} is a prima facie cause of X^q_t and there is an event X^r_{t-i} that occurs between the prima facie cause and the effect such that*

— $\Pr(X^p_{t-i-j}, X^r_{t-i}) > 0$;

— $\Pr(X^q_t \mid X^p_{t-i-j}, X^r_{t-i}) = \Pr(X^q_t \mid X^r_{t-i})$;

— $\Pr(X^q_t \mid X^p_{t-i-j}, X^r_{t-i}) \geq \Pr(X^q_t \mid X^p_{t-i-j})$.

Suppes also extended his framework to quantitative causal relations. And in a similar manner, Granger proposed a probabilistic concept of causality that is defined in terms of the incremental probability of a time series [Granger, 1980].

Instead of seeking single causes that better explain the effect, and partitioning it into genuine/spurious causes, Ellery Eells (1953-2006) considered a quantity to denote how significant each cause is for its effect [Eells, 1991]. In other words, in his approach, he did not seek to find any single more powerful cause, but rather to measure, overall, how well the cause predicts the effect. First, to address the influence of common confounders, Eells defined the set of causal background contexts that constitute a subset of possible factors relevant to the effect and occurring at any time prior to the effect (unlike Suppes who only considered those occurring prior to the potential cause). Eells then defined the average degree of causal significance as the average difference of the probability in each context, weighted by the probability of that background context occurring.

Definition 17 (Average degree of causal significance [Eells, 1991])**.** *The average degree of causal significance of a factor X^p on a factor X^q is given by:*

$$\sum_i \Pr(X^{\boldsymbol{R}_i})(\Pr(X^q \mid X^{\boldsymbol{R}_i}, X^p) - \Pr(X^q \mid\sim X^p, X^{\boldsymbol{R}_i})),$$

where $X^{\boldsymbol{R}_i}$ is the causal background context appropriate for assessing X^p's causal role for X^q.

Nancy Cartwright (1944-present), observed that causal laws are irreducible to laws of association, weather probabilistic or deterministic. Statistical or probabilistic analyses of causality, which typically require for the cause to increase or alter the probability of the effect, cannot succeed, because causes increase the probability of their effects only in situations that exhibit causal homogeneity with respect to that effect [Cartwright, 1979]. In addition, such approaches cannot deal with actual causation, that is, the type of causation that affects individuals based on how events actually play out. Eells suggested a probabilistic account to address this issue [Eells, 1991]. However, as we shall see in the next section, the dominant approach used for actual causation necessitates a whole new concept.

2.7 Counterfactuals theory

Hume's regularity theory can be summed up by his words: "We may define a cause to be an object followed by another, and where all the objects, similar to the first, are followed by objects similar to the second." [Hume, 1748]. Hume then continues by stating "Or, in other words, where, if the first object had not been, the second never had existed." [Hume, 1748]. Even though he seems to take the two definitions as equivalent, it seems that they are not. The first definition tells us that causation is based on regularity, while the second states that causation implies a necessary connection defined in terms of counterfactuals. A counterfactual account of causation would tell us that to prove that an event causes another, we need to prove that if former would not have existed, the latter would not have happened.

For a certain time, only a few thinkers followed the footsteps of Hume's second definition to develop an account of causation based on counterfactuals because counterfactuals themselves were unclear. The true potential of the counterfactual approach to causation did not become clear until counterfactual conditionals became properly understood through the development of *possible worlds semantics* by Rudolph Carnap [Carnap, 1949]

in 1949. Along with the semantics came a consistent set of propositions that allow measuring the similarity between possible worlds. For instance, there should be a weak ordering of the worlds in which a possible world is closer to actuality than another if it resembles the actual world more than the other possible worlds, whereas the actual world should be the closest to actuality, as it resembles itself more than any other world does. In the light of this understanding of counterfactuals, we can define counterfactual dependence as follows:

Definition 18 (Counterfactual dependence [Lewis, 1973])**.** *An event counterfactually depends on another if and only if, if the latter were not to occur, the former would not occur.*

Furthermore, counterfactual dependence was considered to be useful for detecting causation. For example, Figure 2.4 shows two different scenarios: in Figure 2.4a a brick blocks the main path of ball A, and in Figure 2.4b, it does not. Straight black lines represent the actual paths of balls A and B, while dashed red lines represent the path that ball A would have taken in the absence of ball B [Gerstenberg et al., 2014]. Here, the causal question relates to whether ball B caused ball A to go through the gate. In the first scenario (a), the answer is yes, because in the imagined world where ball B did not exist, ball A would have continued its main trajectory and crashed into the block. In the second scenario (b), the answer is no, because in the imagined world where ball B did not exist, ball A would have continued its main trajectory and continued straight to its goal.

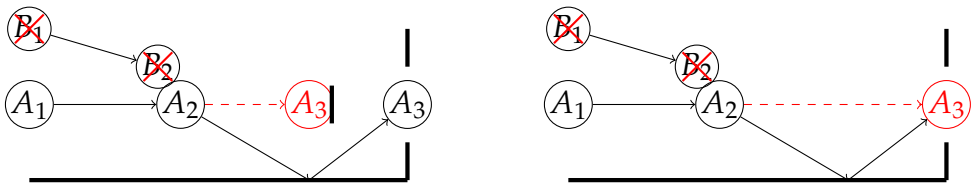


Figure 2.4 – Two illustrations of counterfactual reasoning. Straight black lines illustrate what actually happened, whereas dashed red lines illustrate what would have happened in the absence of ball B. The index on balls A and B represents the time index.

However, counterfactual dependence alone could not form a theory of causation, because it is sufficient but unnecessary for causation. Two events can be causally related without one counterfactually depending on the other, mainly because a cause might sometimes be accompanied by

backups, which can bring about the effect in the absence of the cause. This type of problem usually involves early preemption, which occurs when the process running from the preempted alternative is cut short before the main process running from the preempting cause can be completed. This is illustrated in Figure 2.5, which shows an example similar to that presented in Figure 2.4a with one distinction. In this case, a third ball C would have bumped into ball A in the absence of ball B, because the existence of ball B interrupted the alternative process in which C would have been the cause. In this scenario, ball B is a preempting cause of A reaching the goal, although ball A would have reached the goal even if ball B did not exist because of the preempted backup ball C.

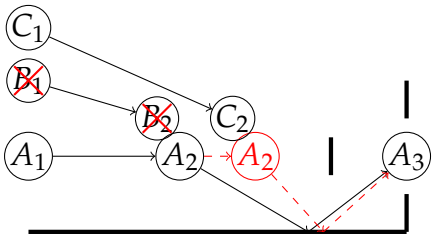


Figure 2.5 – An illustration of early preemption. Straight black lines represent what actually happened, whereas dashed red lines represent what would have happened if B did not exist. The index on balls A, B, and C represents the time index.

Many solutions have been proposed to the problem of necessity between causation and counterfactual dependence [Lyon, 1967]. However, the most famous account is that of David Lewis (1941-2001) who solved it by defining causation in terms of chains of counterfactual dependence.

Definition 19 (Causation in terms of chains of counterfactual dependence [Lewis, 1973])**.** *An event causes another if and only if, there is a chain of counterfactual dependence leading back from the latter to the former.*

Returning to our example, a chain of counterfactual dependence exists between ball B bumping into A and ball A reaching the goal. Supposing t is the time when A reached the goal and $t - i$ the time when ball B bumped into ball A, we can thus say that ball A would not have changed trajectory at time $t - i$ if ball B had not existed, A would not have been in the same position as it was at time $t - i + \epsilon$ if ball B had not existed, and so on until reaching the goal. However, there is no such chain leading back from ball A reaching the goal to ball C bumping into A. The main issues with this approach are late preemption [Lewis, 1986] and trumping preemption

[Schaffer, 2000]. On the one hand, late preemption occurs when the process running from the preempted cause is cut short by the main process running to completion and bringing about the effect before the preempted potential cause has the opportunity to do so. For illustrative purposes, let us consider the same example as in Figure 2.5 but now suppose that ball B bumps into ball A, and a split second later, ball C bumps into ball A without changing its new trajectory. The chain of counterfactual dependencies leading back from ball B bumping into ball A is matched by the chain of counterfactual dependencies leading back from ball C bumping into ball A. On the other hand, trumping preemption occurs when the process running from the preempted cause happens at the same time as the main process running to completion and bringing about the effect, but the preempted potential cause has no influence on the effect relative to the real cause. To understand this, let us take an example analyzed by Lewis himself: a major and a sergeant simultaneously shout "Advance" to a group of soldiers. The soldiers hear them both and advance. Since the soldiers obey the superior officer, they advance because the major orders them to do so, not because of the sergeant. Thus, the major's command preempts or trumps that of the sergeant. To overcome these difficulties, Lewis amended his counterfactual theory [Lewis, 2000] by introducing the notion of alteration.

Definition 20 (Alteration [Lewis, 2000])**.** *An event that is identical to another except that it occurs at a slightly different time, place or in a slightly different manner.*

And by means of the notion of alteration, Lewis defined influence as follows:

Definition 21 (Counterfactual influence [Lewis, 2000])**.** *An event causes another if and only if there is a chain of stepwise influence (alteration) leading back from latter to the former.*

Returning to the late preemption example, if an alteration of ball B bumping into ball A occurred, then an alteration of ball A reaching the goal would occur. But an alteration of ball C bumping into ball A would have no effect on ball A reaching the goal. Using the major and sergeant example, altering only the major's command would correspondingly alter the response of the soldiers. By contrast, altering only the sergeant's command would make no difference as the soldiers would continue to obey the major.

Although Lewis' new approach addressed the problems of early preemption, late preemption, and trumping preemption, it was not immune

to criticism. The fact that causation is represented by a chain suggests that it is transitive. If causation is transitive, then it can infinitely propagate back to the beginning of time. Using the transitivity of stepwise influence, we can argue that the big bang might be the cause of ball A reaching the goal. Thus, it seems that this theory will produce too many causes for each effect. In response to these attacks, Lewis along with Donald Davidson (1917-2003) highlighted the importance of distinguishing causation from explanation. There is no explanation without causation, although there is causation without explanation. Causation does not depend on us: if humans disappear, explanation will cease to exist, but this is not the case with causation. Despite the existence of causes going back to the beginning of time, only a few of them are adequate to provide an explanation.

Two alternative counterfactual approaches to causation (in parallel to Lewis' theory) emerged in applied fields. The first started when Jerzy Neyman (1894-1981) [Splawa-Neyman et al., 1990, Rubin, 1990] sought to solve a problem about the best variety of crops to plant in a given field. With such problems, experimentation is possible but inefficient, since crop planting depends on the season. For each plot, there can only be one entry, which means that the problem at hand generates a sparse matrix, where all the zeros can be referred to as counterfactuals. This led Donald Rubin (1943-present) to view causal inference as a missing data problem [Rubin, 1974] and formalize it as the potential outcome framework. The main idea is about figuring out what is the causal effect of some treatment on some outcome. By treatment, we refer to something that you might be exposed to whether it happens to be in the environment or whether it is something that a clinician would give you as an actual formal treatment. By outcome, we refer to what we would have observed under each possible treatment option. The second counterfactual framework finds its origin in the work of Sewall Wright (1889-1988) [Wright, 1921], who introduced a method known as *path analysis* to explain the patterns of inheriting different color markings in guinea pigs. His method attempts to measure the direct influence of each correlation and find the degree to which the variation of a given effect is determined by each particular cause. For this purpose, diagrams of variables connected by arrows are constructed, showing the different correlations within the system. Based on these diagrams and the correlations observed between variables, equations are constructed and then solved. The coefficients (also known as path coefficients) represent the direct effects of the variables on each other. The work of Wright was later generalized by Judea Pearl (1936-present) and other pioneers in the domain of causality [Pearl, 2000] who subsequently introduced the framework of structural causal models. The potential outcome framework and

the structural causal models framework are not exclusive, as it was proven that one can be theoretically translated to the other [Pearl, 2010]. However, we will focus on the framework of structural causal models here.

Structural causal models combine features of structural equation models used in the social sciences, the graphical models, and the counterfactual account of causation. They deal with examples of late preemption using a certain procedure to test the existence of a causal relation. This procedure involves searching for an intrinsic process that connects the potential cause and effect and then suppresses the influence of their extrinsic surroundings by fixing them as they actually are. Finally, a counterfactual test is applied the potential cause. For the example given in Figure 2.5, the system can be described using the following set of variables:

— $BE = 1$ if ball B exists, 0 otherwise;

— $BA = 1$ if B bumps into ball A, 0 otherwise;

— $CE = 1$ if ball C exists, 0 otherwise;

— $CA = 1$ if C bumps into ball A, 0 otherwise;

— $BG = 1$ if ball A reaches the goal, 0 otherwise.

To test whether B bumping into ball A caused ball A to reach the goal, we should examine the process running from BE through BA to BG while fixing variable CA (extrinsic to this process) to its initial value, and then changing the value of variable BE to see weather it changes the value of BG. The last steps involve evaluating the counterfactual statement that "if ball B had not bumped into ball A and if ball C had not bumped into ball A, then ball A would not have reached the goal". It is easy to see that this counterfactual is true. By contrast, when we apply a similar procedure to test whether ball C bumping into ball A caused ball A to reach the goal, we are required to consider the counterfactual statement "if ball C had not bumped into ball A and if ball B had bumped into ball A, then ball A would not have reached the goal". This counterfactual is false. The difference in the truth value of these two counterfactual statements explains that ball B bumping into ball A, and not ball C bumping into ball A, caused ball A to reach the goal.

2.8 Manipulability and intervention theory

Peter Menzies (1953–2015) and Huw Price (1953-present) developed the manipulability theory of causation (also known as the agency theory of

causation [7]) [Menzies and Price, 1993], which considers free human action to be the model for understanding causation, both in everyday life and in experimental analysis.

Definition 22 (Causation by manipulation [Menzies and Price, 1993])**.** *An event is a cause of a distinct event in the case that bringing about the occurrence of the first would be an effective means by which a free agent could bring about the occurrence of the second.*

This approach tends to be reductionist and circular, thus leading to a subjective conception of causation. It is reductionist, because it grounds causation in non-causal human manipulation. It is circular, because the notion of *bringing about* is itself a causal notion. Finally, it is subjective, because it is only related to human agents. In addition, manipulation, as defined in this theory, cannot be conducted in the entire range of applications, since not all causal knowledge is manipulable by a normal agent.

According to James Woodward (1941-present), the concept of agency is not independent of the notion of causation. However, they lack a special connection with the notion of human agency. Woodward argued that there is nothing logically special about human action or agency. Human interventions can be regarded as nothing more than events in the natural world. He thus introduced the intervention theory of causation [Woodward, 2003], which is an evolution of the manipulability theory. An intervention is a causal process that acts in a surgical, targeted, and exogenous manner. Given a cause X^p and its effect X^q, the intervention on X^p must completely disrupt the causal relationship between X^p and its previous causes so that the value of X^p is set entirely by the intervention. In addition, the intervention must not directly cause X^q via a route that does not go through X^p. Furthermore, the intervention itself should not be caused by any cause that affects X^q via a route that does not go through X^p. Finally, the intervention should leave unchanged the values taken by any causes of X^q except those on the direct path from X^p to X^q (should this exist) unchanged. Given these requirements, X^p causes X^q when some possible intervention on X^p changes the value of X^p along with an associated regular change in the value of X^q.

Definition 23 (Causation by intervention [Woodward, 2003])**.** *A necessary and sufficient condition for X^p to be a direct cause of X^q with respect to some variable set V is that there is a possible intervention on X^p that will change X^q*

7. The early version of the agency theory of causation was developed by von Wright in 1971.

(or the probability distribution of X^q) when all other variables are kept fixed at some value by interventions.

According to Pearl, to be able to give optimal solutions to causal questions, we need to discover how to include interventions in equations. He thus argues that the entire enterprise of probabilistic causation has been misguided from the very beginning, because the central notion that a cause raises the probability of its effect cannot be expressed in the language of probability theory [Pearl, 2000]. In particular, the inequality $\Pr(X^q \mid X^p) > \Pr(X^q)$, which philosophers invoked to define causation as well as its many variations and nuances, fails to capture the intuition behind *probability raising*, which is inherently a manipulative or counterfactual notion. Figure 2.6 shows that the statistical implications of the structures represented in Figure 2.2 are inherited from intervention. For common and intermediate causes, when we intervene on C, we remove the dependency between A and B. However, for the collider (Figure 2.3), when we intervene on C, we do not create a dependency between A and B as in the case of conditioning on C. To rectify the mistakes of the past, Pearl introduced

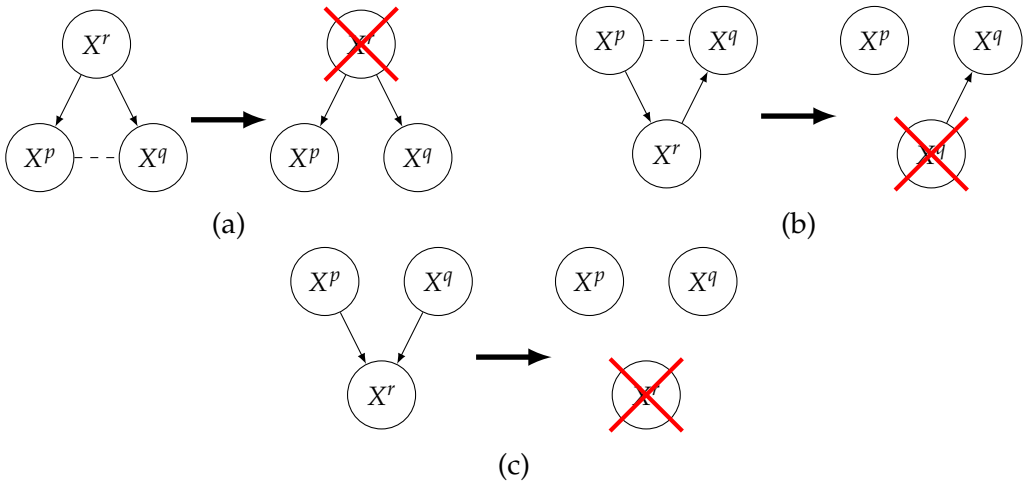


Figure 2.6 – The consequence of an intervention on X^r in the case of a common cause (a), an intermediate cause (b), and an unshielded collider (c). Dashed lines represent correlations, and red crosses denote interventions.

a new type of algebra based on causal relationships [Pearl, 2000, Pearl and Bareinboim, 2011]: the do-calculus. Its main purpose is to integrate interventions into mathematical equations. This consists of three inference rules that allow us to map the interventional and observational distributions whenever certain conditions hold in the causal graph $\mathcal{G}$. Thus, ac-

cording to Pearl, the correct formulation of probability raising should read: $\Pr(X^q \mid do(X^p)) > \Pr(X^q \mid X^p)$.

In parallel with the development of do-calculus, Spirtes, Glymour, and Scheines developed a similar approach to causality [Spirtes et al., 2000] in which they certainly consider interventions as an important part of it. However, their work contains less of an emphasis on the calculus of interventions and focuses more on finding bridge principles that allow to infer causation from observation. Reinchenbach's common cause principle told them that if a common cause (for example, X^r in Figure 2.2a), was not observed, one would infer a spurious correlation and thus a causal relation between its effects (X^p and X^q in Figure 2.2a) as these latter variables are independent only when conditioned on their common cause (X^r in Figure 2.2a). To avoid such spurious correlations, they started by assuming that all common causes are measured and observed; this assumption is known by causal sufficiency.

Definition 24 (Causal sufficiency)**.** *A set of variables is said to be* causally sufficient *if all common causes of all variables are observed.*

As a consequence, if one wants to focus on a few variables, one needs to make sure that all their common causes are also taken into account. To represent causal relations, similarly to Pearl, Spirtes, Glymour, and Scheines used graphical notions. However, for large graphs, it is not obvious how to conclude that two nodes are conditionally independent. Thus, they used *d-separation* (previously introduced by Pearl [1988]), a tool that allows an algorithmic check of conditional independencies in the graph; i.e., d-separation introduce the probability distributions into the graph, which describes the set of independencies in a DAG $\mathcal{G}$ in terms of whether, for two vertices X^p and X^q, there is some set of vertices X^r blocking connections between them. Formally speaking:

Definition 25 (d-connection and d-separation)**.** *If $\mathcal{G}$ is a directed graph in which X^p, X^q and X^r are disjoint sets of vertices, then X^p and X^q are* d-connected *by X^r in $\mathcal{G}$ if and only if there exists an undirected path U between some vertex in X^p and some vertex in X^q such that for every collider X^c on U, either X^c or a descendant of X^c is in X^r, and no non-collider on U is in X^r. Otherwise, X^p and X^q are* d-separated *given X^r.*

One statistical consequence can be drawn out from d-separations and d-connections: if X^r d-separates X^p and X^q then we can say that X^r renders X^p and X^q statistically independent $X^p \perp\!\!\!\perp X^q \mid X^r$ in any distribution that factorizes according to $\mathcal{G}$. Nevertheless, if taken no further, d-

Figure 2.7 – Faithful vs unfaithful graphs.

separation alone is just mathematics connecting DAGs and probability distributions and need not involve causation at all. To make a bridge between d-seperation and causation, Spirtes, Glymour, and Scheines developed the *causal Markov condition* [Spirtes et al., 2000], a generalization of Reichenbach's common cause principle, which was first introduced by Harri Kiiveri and Terry Speed [Kiiveri and Speed, 1982]. The causal Markov condition states that information about a variable is found only in its direct causes and not in its effects or any indirect causes.

Definition 26 (Causal Markov condition)**.** *A causal graph $\mathcal{G} = (V, E)$ with its probability distribution P is said to satisfy the* causal Markov condition *if every vertex in $\mathcal{G}$ is independent with respect to P of its nondescendents given its parents.*

Under this assumption, d-separation becomes the correct connection between causal structure and probabilistic independence. But what about probabilistic dependence? Consider two DAGs $\mathcal{G}$ and $\mathcal{G}'$, such that $\mathcal{G}$ is a supergraph of $\mathcal{G}'$, i.e., $\mathcal{G}$ and $\mathcal{G}'$ have the same vertices, and the set of arrows in $\mathcal{G}'$ is a subset of the set of arrows in $\mathcal{G}$. In such case, if the probability distribution P is Markov to a DAG $\mathcal{G}'$, then P is Markov to $\mathcal{G}$. Hence, the causal Markov condition alone cannot differentiate between the two DAGs. To tackle this problem, the *minimality condition* [8] was introduced.

Definition 27 (Minimality condition)**.** *A DAG $\mathcal{G}$ compatible with a probability distribution P is said to satisfy the* minimality condition *if P is Markov to $\mathcal{G}$ but not Markov to any proper subgraph of $\mathcal{G}$.*

The causal Markov condition alone puts no constraint on the distributions that the structure could produce, so independencies that cannot be explained by the causal Markov condition or d-separation can be obtained. This limits the possibility to infer a causal graph from probabilities alone. Adding the minimality condition is also not sufficient to restrict the set of

8. Here we use the definition introduced by Spirtes et al. [2000], not the one introduced by Pearl [2000].

possible causal structures. To see that, let us assume that we have the following (conditional) independence and dependence relations between the three variables X^p, X^q, and X^r:

$$X^p \perp\!\!\!\perp X^r,\ X^q \not\perp\!\!\!\perp X^p,\ X^q \not\perp\!\!\!\perp X^r,\ X^p \not\perp\!\!\!\perp X^r | X^q,\ X^q \not\perp\!\!\!\perp X^r | X^p,\ X^q \not\perp\!\!\!\perp X^p | X^r.$$

The two graphs given in Figure 2.7 are compatible with the probability distribution given above as $P(X^p, X^q, X^r)$ factorizes in both cases as

$$P(X^p)(X^r)(X^q|X^p, X^r).$$

They furthermore satisfy the minimality condition as removing any edge on one of the two graphs changes the factorization of the joint probability. This said, the graph in Figure 2.7b states that X^p and X^r are unconditionally dependent whereas the probability distribution states they are unconditionally independent. This can be seen as problematic and we say in such a case that the graph is *unfaithful* according to the following definition.

Definition 28 (Faithfulness ([Spirtes et al., 2000]))**.** *We say that a graph $\mathcal{G}$ and a compatible probability distribution P are faithful to one another if all and only the conditional independence relations true in P are entailed by the Markov condition applied to $\mathcal{G}$ using d-separation.*

The faithfulness condition serves as a methodological tool to infer causal graphs and, in many studies, one aims at inferring faithful graphs with respect to the (conditional) independence relations observed in the data. Two causal consequences can be drawn out from d-separations and d-connections when assuming the causal Markov condition and faithfulness: if X^r d-separates X^p and X^q then X^r is not a collider of X^p and X^q according to $\mathcal{G}$; and if X^r d-connects X^p and X^q then X^r is a collider X^p and X^q according to $\mathcal{G}$. Given these assumptions and assuming and no temporal information is available, the causal structure can be retrieved from data up to a Markov equivalent class. Two DAGs are called *Markov equivalent* if they encode the same independence relationships between the observational variables, i.e., same structure and same colliders.

This main theory [9] presented by Spirtes, Glymour, and Scheines is based on DAGs which are easily interpretable but lack the tools to entail knowledge about hidden confounders or selection variables as illustrated in Figure 2.8: by neglecting the hidden confounder L, a causal dependence between X^q and X^r might appear, and by neglecting the selection variable

9. Pearl also contributed to this theory.

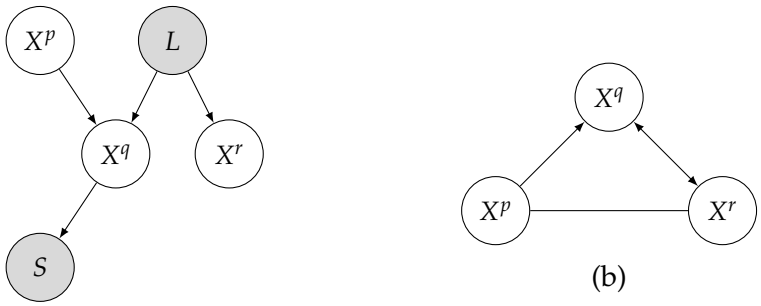


Figure 2.8 – Illustration of hidden confounder (L) and selection bias (S): on the left when observing all the variables, and on the right its representation in a MAG.

S, X^p has a spurious influence on X^r. However, they encountered this problem by using *maximal ancestral graphs* ([Richardson and Spirtes, 2002], MAGs [10]) which enable one to focus on the independence structure over the observational variables that results from the presence of latent variables and selection bias without explicitly including them in the graph. Permitting bi-directed edges ($\longleftrightarrow$) in the graph allows one to graphically represent the existence of an unobserved common cause of observational variables, and permitting undirected edges ($-$) allows to represent unobserved selection variables that have been conditioned on rather than marginalized over, as illustrated in Figure 2.8. MAGs are maximal in the sense that no additional edge may be added to the graph without changing the independence model [Richardson and Spirtes, 2002]. However, this is not enough to infer causal relations with the absence of causal sufficiency because d-separation is only adapted to DAGs. A natural extension of d-separation that can be applied to ancestral graph has been introduced, namely the m-separation. Since the notion of collider and non collider now include bi-directed edges and undirected edges, m-separation can be defined as follows:

Definition 29 (m-separation)**.** *Let $\mathcal{G}$ be a maximal ancestral graph in which X^p, X^q and X^r are disjoint sets of vertices. X^p and X^q are* m-connected *given X^r if there exists an undirected path U between some vertex in X^p and some vertex in X^q such that every collider on U is an ancestor of a vertex in X^r, and no non-collider on U is in X^r. Otherwise, X^p and X^q are* m-separated *given X^r.*

10. In contrast to DAGs, MAGs cannot be trivially be used for causal reasoning [Meganck et al., 2007, Zhang, 2008b]

Using these notions, Spirtes, Glymour, and Scheines introduced two algorithms PC and FCI (which are still the most known algorithms in causal discovery), that can discover causal relations from non temporal data up to a Markov equivalence class, under the assumption of causal sufficiency for PC, and without the causal sufficiency assumption for FCI. These two algorithms will be further discussed in the next chapter.

2.9 Pluralism and hierarchy

Causation is used in verbal discussions in several senses. So why limit causality to one account? This idea goes back to Aristotle. Striking a match caused it to light, one billiard ball caused another to move, Louis caused Elsa to blush, the big bang caused the universe to exist, design flaws of the RBMK reactor and the incompetence of workers caused Chernobyl's nuclear reactor disaster, and so on. As seen in every account of causation, we were able to find an example to challenge it. So perhaps a universal law of causation does not exist. Elizabeth Anscombe (1919-2001) [Anscombe, 1993] argued that a general account of causality is impossible. Maybe it is too much to expect that one theory can encompass all notions of causality. It could be that causation is a general account of different instances with their own laws. Just as many species constitute a genus, perhaps many types of causality constitute universal causality.

Alternatively, Pearl provided an interesting insight unveiled by the logic of causal reasoning. It involves a classification of causal information in terms of the types of questions that each class can answer. This classification forms a three-level hierarchy known as the Pearl causal hierarchy (PCH), which was recently proven to be hierarchically correct (Bareinboim et al, 2020, to be published) in the sense that questions at level i ($i = 1, 2, 3$) can only be answered if information from level j ($j \geq i$) is available. The three-levels are association, intervention, and counterfactual.

Association involves purely statistical relationships defined by observational data. For instance, observing a phone freezing makes it more likely that its messaging system will fail. This association can be inferred directly from the observational data using conditional expectation. Since this layer does not require causal information, it is placed at the bottom of the hierarchy. This process can refer to seeing, observing, or detecting regularities and patterns. All machine learning applications (prediction, clustering, etc.) belong to this level, which can respond to the following questions: "What happens if I see ...?", "How are the variables related?",

"How would an event change my belief in another event?", "What does the increase in CPU tell us about the RAM?", "What does a symptom tell me about a disease?", or "What does a survey tell me about the election results?".

Intervention involves not only seeing what is but changing what we see by acting on the system. A typical question at this level would be: "What happens if we intentionally break the messaging system?". Such questions cannot be answered from monitoring data alone, because they involve a change in system behavior in response to the intentional intervention. The freezing of the phone under this intervention may differ substantially from what happened in the past. It can refer to doing or intervening. Other questions that can be answered at this level take the form of "What happens if I do ...?", "How can I make an event happen?", "If I shut down the server, will the system stabilize?", "If I take an aspirin, will my headache go away?", or "What will happen if we ban cigarettes?". There is an obvious distinction between intervention and conditioning. When we intervene on a variable, we fix its value and thus change the system; as a result, the values of the other variables might change. When we condition on a variable, we change nothing; we merely narrow our focus to the subset of cases in which the variable takes the value of interest. In the case of conditioning, our perception about the world changes, not the world, but in the case of intervening, the world itself changes. If X^p is the cause of X^q, then $\Pr(X^q \mid do(X^p))$ is equivalent to $\Pr(X^q \mid X^p)$, since the new world refers to a subset of the world in which we find ourselves.

In counterfactual, in addition to seeing and changing what we see, there is imagining: namely, imagining what would happen (also known as retrospecting and understanding). This level can answer the following types of questions: "What if I had done ...?", "Why?", "Was it ball B that caused ball A to move?", "What if ball B had not occurred?", "What if I had acted differently?", "Would the frozen phone unfreeze if the messaging system worked?", "Was it the aspirin that stopped my headache?", "Would Kennedy still be alive if Oswald had not killed him?", or "What if I had not smoked for the last 2 years?".

2.10 Conclusion

In this chapter, we provided a brief chronological overview of the causation principle, starting in Ancient Greece, passing through the Renaissance, and ending in the contemporary period. In the last two centuries, many scientists have reflected on the subject, but most were too skepti-

cal to take it seriously. However, in the last few decades, enlightening progress has been made to understand causation as well to infer it under certain conditions. Among others, we focused on the probability raising principle, the prima-facie cause, the causal Markov condition, minimality, and faithfulness which constitute the building block of this thesis.

Chapter 3

A survey on causal discovery for time series

> Resemblance, Contiguity and Causation are the only ties of our thoughts, they are really to us the cement of the universe, and all the operations of the mind must, in a great measure, depend on them.
>
> David Hume

3.1 Introduction

In recent years, causal discovery from statistical data has attracted increasing interest [Hoyer et al., 2009, Mooij et al., 2009, Kalainathan et al., 2018, Bloebaum et al., 2018]. This chapter consists of a review of existing methods that infer causal discovery from time series data and which can be divided into three categories: Granger causality that considers that a cause has unique information about the future values of its effect Granger [1969]; constraint-based approaches that filter unwanted associations via independence test [Spirtes et al., 2000, Pearl, 2000]; and noise-based approaches that as the name suggests, use noise to infer causal relations Hoyer et al. [2009].

Compared to causal inference from static data, the temporal information present in time series serves as a strong constraint for deciding the direction of a causal relation as "a cause precedes its effects" (see Definition 8). This said, the consideration of time series induces new difficulties as the relations between time series can occur across different time lags. Furthermore, observations in time series are often strongly correlated with the recent history of the time series, which may induce spurious correlations. We review in this chapter the most known families of methods that detect causal relations between time series and we describe several algorithms from each family.

In the remainder, we consider d-variate time series X where, for a fixed t, each X_t is a vector $(X_t^1, \cdots, X_t^d)$ in which each variable X_t^p represents a measurement of the p-th time series at time t.

3.2 Granger causality

Granger causality is one of the oldest concepts in causal inference, based on a statistical version of Hume's regularity theory [Hume, 1738] which states that causal relations can be inferred by the experience of constant conjunctions between a cause that precedes its effects. Probabilistic versions of Hume's regularity theory, based on the probability raising principle (conditioning on a cause increases the probability for the effect to appear), have been investigated by different authors, among which one can cite Reichenbach [1956], Suppes [1970] and Eells [1991]. Granger [1969] proposed a statistical version that can be stated as:

Definition 30 (Granger Causality [Granger, 1980])**.** *A time series X^p Granger-causes X^q if past values of X^p provide unique, statistically significant information about future values of X^q.*

For a given effect, the unique information contained in its causes and not in other variables allows to optimally forecast the effect from its causes only. As such, Granger causality assumes causal sufficiency (see Def. 24). In addition, the temporal precedence constraints it relies on prevents one from inferring the direction of "instantaneous" causal relations. Indeed, modifying Granger causality by regressing X_t^q using the past values of X^q and X^p, as well as X_t^p to take into account instantaneous effects, does not allow to decide which variable is the cause and which the effect, as already noted by Granger [1988]. Please bear in mind that instantaneous effects in this context are due to the fact that, with discrete time stamps, small temporal differences between a cause and its effect are observed as being instantaneous.

However, despite these downsides, Granger causality is generally considered as a valuable tool that can improve the performance of prediction and was proven to be effective in many fields such as econometrics [Hiemstra and Jones, 1994], neuroscience [Brovelli et al., 2004, Ding et al., 2006], climate analysis [Papagiannopoulou et al., 2017, Zhang et al., 2011] to name but a few.

We provide below a more detailed description of standard Granger causality and its recent extensions.

3.2.1 Standard PairWise Granger causality

In its simplest version, under the assumption of stationary linear systems and to assess whether X^p Granger-causes X^q, one considers the following autoregression model:

$$X_t^q = a^{q,0} + \sum_{i=1}^{\gamma_o} a^{q,i} X_{t-i}^q + \xi_t^q, \tag{Mres}$$

and its augmented version:

$$X_t^q = a^{q,0} + \sum_{i=1}^{\gamma_o} a^{q,i} X_{t-i}^q + \sum_{i=1}^{\gamma_o} a^{p,i} X_{t-i}^p + \xi_t^q, \tag{Mfull}$$

where $(\xi_t^q)_t$ are uncorrelated random variables with zero mean and variance σ^2, $(a_{q,i})_{1\leq i\leq \gamma_o}$ and $(a^{p,i})_{1\leq i\leq \gamma_o}$ are real coefficients, and γ_o corresponds to the optimal lag value. The model (Mres) is an autoregressive model and is called the *restricted model*. It uses only past values of X^q to predict its current value. The model (Mfull) is an augmented version of the autoregressive model and is called the *full model*. It uses both past values

of X^q and X^p to predict the current value of X^q. If the full model is significantly more accurate than the restricted model, one can conclude that X^p Granger-causes X^q. From a statistical viewpoint, a statistical test such as the F-test can be used to determine whether the full model is significantly better than the restricted one, the null hypothesis stating that X^p does not Granger-cause X^q. In practice, the optimal lag γ_o can be estimated using any information criterion, as the Akaike or Schwartz information criteria ($\gamma_o \in \{1, \cdots, \gamma_{max}\}$).

In a multivariate setting, a pairwise analysis can be performed using the bivariate approach. This approach does however not fully capture Granger's original ideas which assume that all relevant information is included in the analysis [Eichler, 2008]. Furthermore, a pairwise approach may lead to ambiguous results in terms of differentiating direct from mediated causal relations [Ding et al., 2006], detecting for example a spurious correlation in a chaining of three times series, which can be removed by conditioning on the common dependencies. To address these problems, a direct extension of Granger causality to multivariate time series has been proposed.

3.2.2 MultiVariate Granger causality

To overcome the problem of common confounders, all relevant information needs to be included in the analysis. Let $X = (X^1, X^2, \cdots, X^d)$ be a d-dimensional time series. The multivariate Granger causality, or conditional Granger causality [Geweke, 1982, Chen et al., 2004, Barrett et al., 2010], makes use of the following restricted and full models, both based on a vector autoregressive extension of the autoregressive model of the pairwise case:

$$X_t^q = a^{q,0} + \sum_{\substack{r=1 \\ r \neq p}}^{d} \sum_{i=1}^{\gamma_{\max}} a^{r,i} X_{t-i}^p + \xi_t^q, \qquad \text{(mvMres)}$$

$$X_t^q = a^{q,0} + \sum_{r=1}^{d} \sum_{i=1}^{\gamma_{\max}} a^{r,i} X_{t-i}^r + \xi_t^q, \qquad \text{(mvMfull)}$$

where $(\xi_t^q)_t$ are uncorrelated random variables with zero mean and variance σ^2, $(a^{r,i})_{1 \leq i \leq \gamma_{\max}, 1 \leq r \leq d}$ and $a^{q,0}$ are real coefficients, and $\gamma_{\max}$ is as before the optimal lag. Here the full model (mvMfull) uses all observational time series whereas the restricted model (mvMres) uses all time series except X^p. Analogously to the bivariate case, if the full model is significantly

more accurate than the restricted model (through a statistical test), one concludes that X^p Granger-causes X^q. If the conditional Granger causality is sound and usually yields better results for inferring a causal graph from observational multivariate time series, its computation overload is such that in practice many studies rely on the pairwise version.

Many proposals, extended Granger causality to non-linear relations among which and not surprisingly, several investigated the use of deep networks. The temporal causal discovery model `TCDF` represents such an attempt. Because of the popularity of deep neural networks, we detail it below.

3.2.3 A deep-learning method to detect Granger causality

The temporal causal discovery framework (TCDF), introduced by Nauta et al. [2019], learns complex non linear causal relations between time series using deep neural networks. It is based on an attention mechanism within dilated [1] depthwise [2] convolutional networks. It consists of d independent attention-based CNNs $(N^q)_{1\leq q\leq d}$, all with the same architecture but with a different target time series X^q as illustrated in Figure 3.1. Along with the prediction of the target, each neural network outputs its attentions scores and its kernel weights which allow a causal interpretation of the results: high attention on a time series X^p while forecasting a time series X^q indicates that the former contains relevant information that helps forecasting better the latter.

For $1 \leq q \leq d$, the attention scores $(a^{q,p})_{1\leq p\leq d}$ of the attention mechanism indicate which time series contains the most valuable information for prediction, and detect which ones are potentially causally associated with the target time series X^q. To interpret the attention scores causally, the softmax function σ is applied and followed by a straightforward semi-binarization that truncates all attention scores that fall below a threshold s^q. To determine s^q, `TCDF` starts by ranking the attention scores from high to low and then searches for the largest gap [3] between two adjacent attention scores. The threshold s^q is then equal to the biggest attention score associated with that gap. To distinguish causality-based from correlation-based attention, a causal validation step is applied: potential

1. A dilated convolution applies a kernel over an area while skipping values with a certain step size. This step size increases exponentially from a hidden layer to another depending on a chosen dilation coefficient c.

2. A depthwise convolution is a type of convolution where a single convolutional filter is applied for each input channel. In this case, each channel is a time series.

3. Additional constraints can be added; for more details see Nauta et al. [2019].

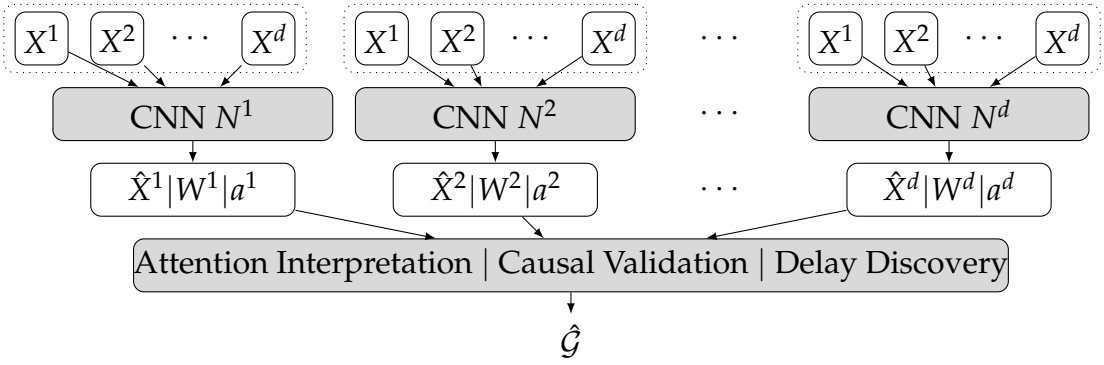


Figure 3.1 – Neural network associated to TCDF: d independent CNNs $(N^q)_{1 \leq q \leq d}$, all having time series $X^1 \cdots X^d$ of length T as input. For $1 \leq q \leq d$, the network N^q predicts X^q by $\hat{X}^q$, and also outputs the kernel weights $(W^{q,p,k})_{1 \leq p \leq d, 1 \leq k \leq K}$ (where K represents the kernel size) and attention scores $(a^{q,p})_{1 \leq p \leq d}$. After attention interpretation, causal validation, and delay discovery, a temporal causal graph is constructed.

Figure 3.2 – How TCDF deals with hidden confounders. A red edge (a) indicates a wrong causal relation discovered by TCDF, whereas a green double edge (b) indicates that a true causal relation is discovered. Numbers correspond to delays.

causes are validated if the loss of a network, when removing the chronicity of a time series using permutation, increases significantly when a variable is permuted. Once all causal relations have been established for time series X^q, TCDF detects their time delays by interpreting the kernel weights $(W^{q,p,k})_{1 \leq p \leq d, 1 \leq k \leq K}$ which consist of d rows and K columns (where K is the kernel size). Each row is associated to one input time series and each column shows the importance of each time delay of associated time series.

As can be seen in Figure 3.1, TCDF can learn self-causation since it includes the past of X^q when fitting N^q for $1 \leq q \leq d$. It is also able to detect hidden confounders if they have equal delays to their effects with no additional cost by simply assuming causal relations cannot be instantaneous.

For example, TCDF is able to detect the presence of hidden confounder in Figure 3.2 (b) but not in Figure 3.2 (a).

One of the main drawbacks of TCDF is the number of hyperparameters it relies on (number of hidden layers, kernel size, dilation coefficient, number of epochs, loss function, and learning rate) and the difficulty to set them. In addition, unlike other methods, there is no direct way to set the maximum number of lags as increasing the number of hidden layers (or the kernel size or the dilation coefficient) leads to an increase in the number of time steps seen by the sliding kernel, and so to an increase in the maximum delay.

3.3 Constraint-based approaches

Constraint-based approaches exploit conditional independencies to build a skeleton between variables. This skeleton is then oriented according to a set of rules that define constraints on admissible orientations. Central to these approaches is the notion of v-structures, or colliders, as these are the only structures which can be oriented without ambiguity (an example of a v-structure is given in Figure 2.3, page 34). We first cover here the main algorithms assuming causal sufficiency, corresponding to situations when all possible causes are observed, then we deal with situations without causal sufficiency, *i.e.*, with hidden causes.

3.3.1 With causal sufficiency

The goal here is to exploit conditional independencies, obtained from observational data, to construct the underlying causal graph which is typically represented by a directed acyclic graphs (DAGs) in causally sufficient situations. The underlying causal graph is however not unique as several DAGs can be used to represent the same set of conditional independencies. For example, the models in Figure 3.3, borrowed from Verma and Pearl [1991], all represent the same independence relation "X^p is independent from X^q given X^r" ($X^p \perp\!\!\!\perp X^q | X^r$). This leads to the notion of *Markov equivalence class* which corresponds to a set of DAGs that encode the same set of conditional independencies. Verma and Pearl [1991] have shown that two DAGs are Markov equivalent if and only if they have the same skeleton and the same v-structures. This notion of equivalence only relies on the orientation of *compelled* edges, that is edges participating to v-structures or whose change in orientation would lead to new v-structures and can be represented by partially directed acyclic graphs (PDAGs), in

which some edges are not oriented [4]. Given an equivalence class of DAGs, Andersson et al. [1997] and Chickering [2002] introduce the completed PDAG (CPDAG) as the PDAG that consists of a directed edge for every compelled edge in the equivalence class, and an undirected edge for all other edges. It turns out that a CPDAG uniquely represents a Markov equivalence class. Thus, the goal of constraint-based, causal discovery algorithms can finally be formulated as: construct, from observational data, the CPDAG that represents the Markov equivalence class of a true causal graph.

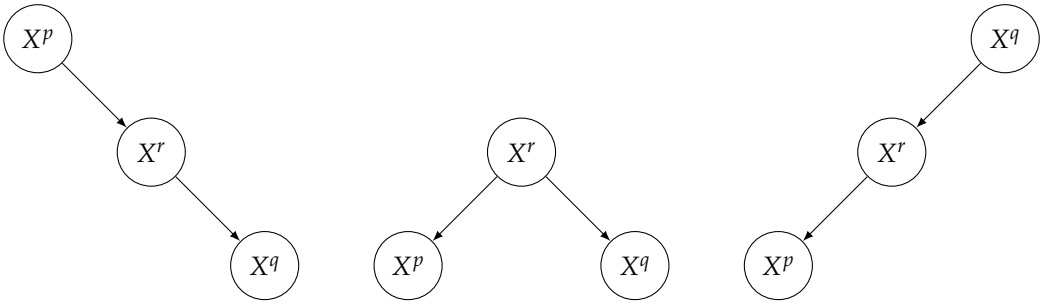


Figure 3.3 – Three Markov equivalent structures.

One of the oldest constraint-based algorithms is the SGS algorithm [Spirtes et al., 1990], which has been proved to be consistent under *i.i.d.* observations assuming causal sufficiency. SGS starts with a full undirected graph connecting all variables. For each pair of vertices (X^p, X^q), it finds (if possible) some subset of vertices that makes them conditionally independent (the smallest such subset is referred to as Sepset(X^p, X^q)) removes the edge between them if it is the case. It then orients undirected edges by subsequently employing orientation rules to derive causal conclusions. The main drawback is that the number of conditional independencies that needs to be tested in a fully connected graph grows exponentially with the number of variables, which makes SGS not usable in practice. The Peter-Clark (PC) algorithm was introduced [Spirtes et al., 2000] to address this issue.

4. This extension is particularly useful when dealing with situations in which it is difficult, or even impossible, to decide on an orientation. For example, one cannot in general determine the direction of causality between two non-temporal variables solely from their observations.

Peter-Clark algorithm for non-temporal data

The PC algorithm aims at optimizing the number of computations necessary to assess whether two variables are conditionally independent or not by considering conditioning variables that are likely to be parents of the two variables. Even if it grows exponentially with the maximal degree of the graph, large sparse graphs can be easily inferred using the PC algorithm.

Starting with a complete undirected graph $\mathcal{G}$, the algorithm checks the dependency for all pairs of vertices and removes or keeps links according to whether or not the two vertices are considered to be independent. Then it checks the conditional independencies between dependent vertices by first computing it for each adjacent pair X^p and X^q in $\mathcal{G}$ and for each vertex X^r (other than X^p) adjacent to X^q in $\mathcal{G}$. If X^r is able to remove the dependency between X^p and X^q then the algorithm removes the edge between them and adds X^r to their separation set Sepset(p, q). Then, it gradually increases the number of variables to condition on, and proceeds as above till a conditional independence is found or all sets of vertices adjacent to X^q have been considered for the conditioning.
Once the skeleton has been constructed, the algorithm applies series of rules [Spirtes et al., 2000, Colombo and Maathuis, 2014a], starting by identifying v-structures using the so-called *origin of causality*.

PC-Rule 0 (Origin of causality)**.** *For every triple $X^p - X^r - X^q$ such that X^p and X^q are not adjacent and $X^r \notin Sepset(p, q)$, orient the triple as $X^p \to X^r \leftarrow X^q$.*

Triples of the form $X^p - X^r - X^q$ such that X^p and X^q are not adjacent are usually referred to as *unshielded* triples in the causality literature. We do not use this term here so as to remain as simple as possible in our exposition of the PC algorithm but will use it in the remainder of the thesis.

When all v-structures have been identified using the above rule, the PC algorithm orients as many of the remaining undirected edges as possible, by repeating the following rules until no other changes can be made.

PC-Rule 1. *In a triple $X^p \to X^q - X^r$ such that X^p and X^r are not adjacent, orient $X^q - X^r$ as $X^q \to X^r$.*

PC-Rule 2. *If there exist a direct path from X^p to X^q and an edge between X^p and X^q, then orient $X^p \to X^q$.*

PC-Rule 3. *Orient $X^p - X^q$ as $X^p \to X^q$ whenever there are two paths $X^p - X^r \to X^q$ and $X^p - X^s \to X^q$.*

A different orientation in PC-Rule 1 would lead to new v-structures, which is not possible as the *origin of causality* should identify all v-structures. A different orientation in PC-Rule 2 would lead to a cycle, whereas a different orientation in PC-Rule 3 would lead to either a cycle or a new v-structure when orienting the remaining undirected edges.

From a theoretical viewpoint, the above procedure is correct, sound and complete [Meek, 1995, Andersson et al., 1997] in the set of Markov equivalence graphs.

Theorem 1 (Theorem 5.1 in [Spirtes et al., 2000])**.** *Let the distribution of V be faithful to a DAG $\mathcal{G} = (V, E)$ and assume that we are given perfect conditional independence information about all pairs of variables (X^p, X^q) in V given subset $X^S \subseteq V \setminus \{X^p, X^q\}$. Then, the output of the PC-algorithm is the CPDAG that represents $\mathcal{G}$.*

Consistency of the PC algorithm has been discussed in Spirtes et al. [2000], Robins et al. [2003]: if the model is only faithful, uniform consistency cannot be achieved, but pointwise consistency can. Kalisch and Bühlmann [2007], Zhang and Spirtes [2002] provide assumptions which render the PC-algorithm uniformly consistent, for a number of nodes and neighbors increasing in a limited way with respect to the sample size

The main weakness of the original PC algorithm is that it is order dependent and thus not stable. To tackle this issue, Colombo and Maathuis [2014a] proposed to measure all conditional independencies for a given cardinal before removing links in the undirected graph. This simple modification renders the main procedure order-independent.

In the following, we detail three popular methods for time series based on PC algorithms. Other methods, as for example FASK [Sanchez-Romero et al., 2019], have also been proposed using different orientation rules. They are however beyond the scope of the current survey.

Temporal Extension with Momentary Conditional Independence Tests

The `PCMCI` algorithm [Runge et al., 2019] is able to detect time lagged causal relations in a window causal graph (see Section 1.3). The method is divided into three steps. First, a partially connected graph $\mathcal{G}$ is constructed, such that all pairs of nodes (X^p_{t-i}, X^q_t) are directed as $X^p_{t-i} \to X^q_t$ if $i > 0$. The second step removes all unnecessary edges based on conditional independencies, as done in PC, and takes into account the assumption of consistency through time to remove homologous edges: for each edge $X^p_{t-i} \to X^q_t$ removed, all edges included in $\mathrm{Hom}(X^p_{t-i}, X^q_t, \mathcal{G})$ are removed as well, where $\mathrm{Hom}(X^p_{t-i}, X^q_t, \mathcal{G})$ represents the set of instants ho-

mologous to X^p_{t-i} and X^q_t, *i.e.*, instants in X^p and X^q shifted by a lag of i from p to q (see Section 1.3). As the conditioning is based only on the parents of X^q_t, it does not control false positives correctly for large autocorrelation in X^p_{t-i}. Thus, the third step takes into account those autocorrelations by using the momentary conditional independence test (MCI). MCI conditions on the parents of X^q_t and the parents of X^p_{t-i} while testing $X^p_{t-i} \to X^q_t$. It is defined as follows:

$$\mathrm{MCI}(X^p_{t-i}; X^q_t) = \mathrm{I}\left(X^p_{t-i}; X^q_t \,|\mathrm{Par}(X^q_t)\backslash\{X^p_{t-i}\}, \mathrm{Par}(X^p_{t-i})\right),$$

and estimates an interpretable notion of causal strength as it quantifies the causal effect of a hypothetical perturbation in $X^p_{t-\gamma_{\max}}$ on X^q_t. Thus, the value of the MCI statistics allows to rank causal links in large-scale settings. The method depends on the significant rate α, which can be selected using Akaike information criterion or cross validation. The computational time is polynomial in the number d of time series and the maximum lag γ_{max}.

`PCMCI` has been shown to be consistent [Runge et al., 2019]. Note that both stages of `PCMCI` can be flexibly combined with any kind of conditional independence tests. We rely in our experiments (Chapter 6) on two measures used in Runge et al. [2019], namely the partial correlation and the mutual information.

Instantaneous causal relations [5], which were not supported in the initial algorithm, have been integrated in Runge [2020] by conducting separately the edge removal for lagged conditioning sets and instantaneous conditioning sets. Lagged relations are treated as in `PCMCI` and instantaneous relations are inferred using the PC-rules.

Temporal extension using transfer entropy

Even if PC-based methods optimize the number of conditional independencies to be computed, the conditioning sets might go up to the size of the entire network. In this respect, regardless of the dimensionality of the sample space, the combinatorial search itself can be computationally infeasible for moderate to large networks. One way to overcome this

5. The difference in time between two events associated with two time series may not be observed if the sampling frequencies of the time series are small. It is thus possible that two events that occurred at different time instants will be seen as instantaneous in the observational time series. Instantaneous causal relations, sometimes called contemporaneous causal relations, correspond to causal relations between causes and effects that occur at different time instants yet appear instantaneous.

issue would be to use an asymmetric measure such as transfer entropy [Schreiber, 2000], which can be defined as follow:

$$\mathrm{TE}(X^p \to X^q) = h(X^q_{t+1} \mid X^q_t) - h(X^q_{t+1} \mid X^q_t, X^p_t)$$

where $h(. \mid .)$ denotes the conditional entropy. However, this metric is limited to pairwise relations and assumes that nodes are self causal. Therefore, Sun et al. [2015] introduced the *causation entropy* (CE), a generalization of the conditional transfer entropy to multivariate time series which relaxes the self causation assumption. Causation entropy from a set of nodes $\mathbf{P}$ to the set of nodes $\mathbf{Q}$ conditioned on the set of nodes $\mathbf{R}$ is defined as:

$$\mathrm{CE}(X^{\mathbf{P}} \to X^{\mathbf{Q}} \mid X^{\mathbf{R}}) = h(X^{\mathbf{Q}}_{t+1} \mid X^{\mathbf{R}}_t) - h(X^{\mathbf{Q}}_{t+1} \mid X^{\mathbf{R}}_t, X^{\mathbf{P}}_t),$$

where $\mathbf{P}, \mathbf{Q}, \mathbf{R}$ are all subsets of $\{1, \cdots, d\}$. Sun et al. [2015] proved that the set of nodes that directly causes a given node is the unique minimal set of nodes that maximizes causation entropy. They propose the `oCSE` (optimal causation entropy) algorithm, to find, for each node X^p_t, the smallest set that maximizes the causation entropy. As they detect only causation relations with time-lag of size 1, they consider stationary first-order Markov processes with the following dynamics:

$$X^q_t = f_q(a^1 X^1_{t-1}, a^2 X^2_{t-1}, \cdots, a^d X^d_{t-1}, \xi^p_t),$$

where for all $p \in \{1, \ldots, d\}$, a_p is the weight of the link from X^p to X^q. Note that the parents of X^q_t can only be attributed to the time $t-1$, known as the temporally Markov assumption: for all t, $\Pr(X_t \mid X_{t-1}, X_{t-2}, \cdots) = \Pr(X_t \mid X_{t-1})$. `oCSE` starts by identifying nodes that form a superset of the causal parents (including indirect and spurious causal connections): iteratively, it adds the node with the largest CE, conditioning on the set of parents (which recursively increases). Then, the second step consists in eliminating from the set of parents the ones deemed insignificant. This algorithm strikes a tradeoff between computational cost and data efficiency.

The second stage of the algorithm is order dependent so results might vary depending on which of the potential parents is treated first.

3.3.2 Without causal sufficiency

As explained in Chapter 2, hidden confounders and unobserved selection variables can be represented by maximal ancestral graphs (MAGs). They play the role of DAGs in situations when not all variables are observed. As shown in Figure 2.8, page 47, the fact that two variables are

related through a common confounder is represented in a MAG by a double arrow, whereas the dependence between two variables induced by an unobserved selection variable is represented by an undirected edge. The equivalence between MAGs is slightly more complex than the one between DAGs and makes use of the notion of discriminating paths.

Definition 31 (Discriminating path)**.** *In a MAG, a path U between X^p and X^q is a* discriminating path *for X^r if U includes at least three edges, X^r is a non-endpoint vertex and is adjacent to X^q, X^p is not adjacent to X^q, and every vertex between X^p and X^r is a collider and a parent of X^q.*

Ali et al. [2005] and Zhang [2007] showed that two MAGs are Markov equivalent if and only if they have the same adjacencies, the same unshielded colliders, and if a path U is a discriminating path for a vertex X^r in both graphs, then X^r is a collider on the path in one graph if and only if it is a collider on the path in the other. As shown in Richardson [1996], a Markov equivalent class of MAGs can be described by a partially ancestral graph (PAG) which can contain up to six types of edges: undirected ($-$), single arrow ($\rightarrow$ or $\leftarrow$), double arrow ($\leftrightarrow$), undirected on one side and undetermined on the other ($-\!\circ$ or $\circ\!-$), directed on one side and undetermined on the other ($\circ\!\rightarrow$ or $\leftarrow\!\circ$), and undetermined on both sides ($\circ\!-\!\circ$). In MAGs, the separation subset that ensures independence between two vertices X^p and X^q can include vertices that are neither parents of X^p nor of X^q. This leads to the notion of possible d-separation sets, in short Possible-Dsep, introduced in Spirtes et al. [2000]. We introduce here a symmetric version of Possible-Dsep sets that may lead to a slower algorithm than the one based on the original asymmetric version of Spirtes et al. [2000] but that simplifies the exposition of the overall procedure.

Definition 32 (Possible-Dsep [Spirtes et al., 2000, Zhang, 2008a])**.** *The* Possible-Dsep *set of two time series X^p and X^q is the set of time series X^r that are such that $X^p \neq X^r$ (or $X^q \neq X^r$) and there is an undirected path U between X^p and X^r (or between X^q and X^r) such that every vertex on U is an ancestor of X^p or X^q and, except for the endpoints, is a collider on U.*

In the graph presented in Figure 3.4, which displays two hidden common causes (L^1, L^2) between X^p and X^q and X^v and X^w, the set $\{X^q, X^r, X^u, X^v\}$ is a Possible-Dsep set for X^p and X^w. It separates these two time series. Note that X^q or X^v alone does not separate X^p and X^w as there is still a path relating X^p and X^w. X^q and X^v together neither separate X^p and X^w as conditioning on X^q creates a dependence between X^r and X^p, and similarly for X^v and X^w, so that X^p and X^w become dependent.

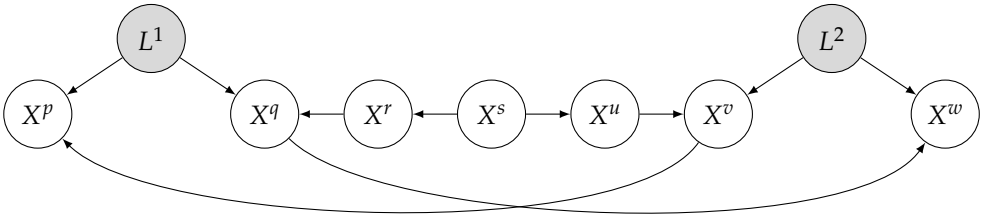


Figure 3.4 – Causal graph with two hidden common causes.

We now present the standard causal inference algorithm for non-temporal data without causal sufficiency, referred to as FCI for fast causal inference, prior to describing extensions to time series.

Fast causal inference algorithm for non-temporal data

FCI starts, as the PC algorithm, by initializing the skeleton with all possible edges and by removing the edges that are either independent or conditionally independent, first when conditioning with Sepsets and then with Possible-Dsep sets. Ten orientation rules, described in, *e.g.*, Zhang [2008a], are applied recursively [6]. As in PC, all colliders are first identified by Rule 0. One then orients as many of the remaining undirected edges as possible, by repeating Rules 1 to 4.

FCI-Rule 0 (Origin of causality)**.** *For each unshielded triple $X^p \ast\!\!-\!\!\circ X^r \circ\!\!-\!\!\ast X^q$, if $X^r \notin Sepset(p,q)$, then orient the unshielded triple as a collider: $X^p \ast\!\!\rightarrow X^r \leftarrow \ast X^q$.*

FCI-Rule 1. *In an unshielded triple $X^p \ast\!\!\rightarrow X^r \circ\!\!-\!\!\ast X^q$, if $X^r \in Sepset(p,q)$ then orient the unshielded triple as $X^p \ast\!\!\rightarrow X^r \ast\!\!\rightarrow X^q$.*

FCI-Rule 2. *If there exists a triple $X^p \rightarrow X^r \ast\!\!\rightarrow X^q$ or a triple $X^p \ast\!\!\rightarrow X^r \rightarrow X^q$ with $X^p \ast\!\!-\!\!\circ X^q$, then orient the pair as $X^p \ast\!\!\rightarrow X^q$.*

FCI-Rule 3. *If there exists an unshielded triple $X^p \ast\!\!\rightarrow X^r \leftarrow\!\!\ast X^q$ and an unshielded triple $X^p \ast\!\!-\!\!\circ X^s \circ\!\!-\!\!\ast X^q$, and $X^s \ast\!\!-\!\!\circ X^r$ then orient the pair as $X^s \rightarrow X^r$.*

FCI-Rule 4. *If there exists a discriminating path between X^p and X^q for X^r, and $X^r \circ\!\!-\!\!\ast X^q$; then orient $X^r \circ\!\!-\!\!\ast X^q$ as $X^r \rightarrow X^q$; otherwise orient the triple as $X^s \longleftrightarrow X^r \longleftrightarrow X^q$.*

6. In stating the 10 orientation rules, the meta-symbol $\ast$ is used as a wildcard that may stand for all three possible edge marks: $-$, $\rightarrow$, $-\!\!\circ$.

The remaining rules make use of the notions of *uncovered path*, *potentially directed path*, and *circle path*. An uncovered path is a path in which every consecutive triple is unshielded. A potentially directed path of length l is a path, which we assume to be represented, after re-indexing the vertices, as $X^1, \cdots, X^l$, that is such that an edge between two consecutive vertices X^{i-1} and X^i has no arrow on X^{i-1}'s side and has either an arrow or a circle on X^i's side. A circle path is a potentially directed path in which every edge on the path is of the form $\circ\!\!-\!\!\circ$.

If selection bias is considered, FCI-Rules 5 to 7 are applied recursively to discover selection variables. Then, FCI-Rules 8 to 10 are applied recursively to pick up directed edges missed by FCI-Rules 0 to 4.

FCI-Rule 5. *For every remaining $X^p \circ\!\!-\!\!\circ X^q$, if there is an uncovered circle path $U = \langle X^p, X^r, \cdots, X^s, X^q \rangle$ between X^p and X^q such that X^p and X^s are not adjacent and X^q and X^r are not adjacent, then orient $X^p \circ\!\!-\!\!\circ X^q$ and every edge on U as undirected edges (-).*

FCI-Rule 6. *If $X^p - X^r *\!\!-\!\!\circ X^q$ (X^p and X^q are not necessarily adjacent), then orient the triple as $X^p - X^r \rightarrow X^q$.*

FCI-Rule 7. *If $X^p - \circ X^r \circ\!\!-\!\!* X^q$, and X^p and X^q are not adjacent, then orient the triple $X^p - \circ X^r \rightarrow X^q$.*

FCI-Rule 8. *If $X^p \rightarrow X^r \rightarrow X^q$ or $X^p - \circ X^r \rightarrow X^q$, and $X^p \circ\!\!\rightarrow X^q$, then orient $X^p \rightarrow X^q$.*

FCI-Rule 9. *If $X^p \circ\!\!\rightarrow X^q$, and U is an uncovered potentially directed path from X^p to X^q such that X^q and X^r are not adjacent, then orient the pair as $X^p \rightarrow X^q$.*

FCI-Rule 10. *Suppose $X^p \circ\!\!\rightarrow X^q$, $X^r \rightarrow X^q \leftarrow X^s$, U_1 is an uncovered potentially directed path from X^p to X^r, and U_2 is an uncovered potentially directed path from X^p to X^s . Let μ be the vertex adjacent to X^p on U_1 (μ could be X^r), and ω be the vertex adjacent to X^p on U_2 (ω could be X^s). If μ and ω are distinct, and are not adjacent, then orient $X^p \circ\!\!\rightarrow X^q$ as $X^p \rightarrow X^q$.*

From a theoretical viewpoint, FCI is correct, sound, complete [Zhang, 2008a] and consistent [Colombo et al., 2012]. One of the disadvantages of FCI, however, is that the conditional independence tests given subsets of Possible-Dsep sets can become very large even for sparse graphs. RFCI [Colombo et al., 2012] was introduced to solve this problem. This algorithm avoids searching for Possible-Dsep sets by performing additional tests. The number of these additional tests and the size of their conditioning sets remain reasonable for sparse graphs, making RFCI much faster than FCI for sparse graphs.

Temporal extension through window representations and VAR

Entner and Hoyer [2010] adapted FCI to time series by transforming the original time series $X_t = (X_t^1, \cdots, X_t^d)_{1 \leq t \leq N}$ into a sample of random vectors with a sliding window of size τ. This leads to the consideration of $(N - \tau + 1)$ vectors of length τd on which the FCI algorithm can be applied. Additionally, one makes use of temporal priority and consistency throughout time (time invariance) to orient edges and restrict conditioning sets. Unlike FCI, this procedure, called `tsFCI`, neither considers selection variables nor instantaneous relations.

Recently, Malinsky and Spirtes [2018] adapted this idea in a new algorithm called `SVAR-FCI` that is based on FCI for multivariate time series and that allows instantaneous causal relations and arbitrary latent confounding. Stationarity is further used to remove additional edges. The data generation process is a structural vector autoregression (SVAR) model with latent variables.

3.4 Noise-based approaches

We focus now on a class of causal models called functional causal models (FCM) (sometimes also called structural equation models, [Wright, 1921, Pearl, 2000]) which describe a causal system by a set of equations, where each equation explains one variable of the system in terms of its direct causes and some additional noise. For example, if X^p is a cause of X^q, then there exists a function f^q that relates X^p to X^q with some additional noise ξ^q: $X^q = f^q(X^p, \xi^q)$.

Statistical noise is often considered as a nuisance that one has to live with, and is even thought to mask causal relations. However, recent discoveries showed that not only noise does not obscure causal relations, but it can be a valuable source of insight. To understand why noise can be helpful to identify causal relations, let us start with a simple example borrowed from Climenhaga et al. [2019] and let us consider two random variables X^p and X^q such that $X^p \to X^p$ with the underlying relation $X^q = 2X^p + \xi^q$, where ξ^q represents some noise. Given enough observations, one can detect the correlation between X^p and X^q. However, without additional information, it is not possible to distinguish between $X^p \leftarrow X^q$ and $X^p \to X^q$ as the model can either be $X^q = 2X^p + \xi^q$ or $X^p = X^q/2 + \xi^p$. Nevertheless, if one assumes that the noise follows, say, a uniform distribution on $\{-1, 0, 1\}$, then one can decide between those two models. Indeed, by computing the error terms $\xi^q = X^q - 2X^p$ and

$\xi^p = X^p - X^q/2$ over the observations, we can easily check which of the two causal structures is compatible with the distribution assumption we made on the noise, as shown in Table 3.1.

It turns out that similar conclusions can be reached if one replaces the strong assumption on the noise distribution by the assumption of independence of mechanisms, which is more realistic and can be applied in an agnostic scenario, and if one uses additional assumptions on the underlying model.

Principle 1 (Independent Mechanisms [Peters et al., 2017])**.** *The causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other. In the probabilistic case, this means that the conditional distribution of each variable given its causes (i.e., its mechanisms) does not inform or influence the other conditional distributions. In case we have only two variables, this reduces to an independence between the cause distribution and the mechanism producing the effect distribution.*

The consequences of this principle are three-folds:

1. The underlying equations are assumed to be autonomous with respect to any external change in one equation. In other words, changes in the generating process of one variable does not imply changes in the generating process of the other variables.
2. The mechanism generating an effect from its cause contains no information about the mechanism generating the cause (although the effect contains information about its cause). This can also be interpreted as an independence between the cause and the noise of the effect. Back to our example, it is easy to check that $X^p \perp\!\!\!\perp \xi^q$ but $X^q \not\perp\!\!\!\perp \xi^p$ and so the real causal direction is identifiable.
3. Noises associated with different variables are mutually independent.

In the remainder, we focus on FCM models of the form $X^q = f^q(X^p, \xi^q)$ with $X^p \perp\!\!\!\perp \xi^q$.

X^p	X^q	$\xi^q = X^q - 2X^p$	$\xi^p = X^p - X^q/2$
1	2	$0 \in \{-1, 0, 1\}$	$0 \in \{-1, 0, 1\}$
3	6	$0 \in \{-1, 0, 1\}$	$0 \in \{-1, 0, 1\}$
4	9	$1 \in \{-1, 0, 1\}$	$-0.5 \notin \{-1, 0, 1\}$

Table 3.1 – Toy example to illustrate the use of the noise to detect causality. We observe data and compute the two possible noise ξ^p and ξ^q coming from the models $X^q = 2X^p + \xi^q$ and $X^p = X^q/2 + \xi^p$. As we have assumed that the noise's support is $\{-1, 0, 1\}$, only one model is feasible.

It turns out that, in general, one cannot identify the underlying model solely from observations of the joint distribution of the two variables, as stated in the following proposition.

Proposition 1 (Non-uniqueness of graph structures [Peters et al., 2017])**.** *For every joint distribution of two real-valued variables X^p and X^q, there is an FCM $X^q = f^q(X^p, \xi^q)$, where X^p is independent from ξ^q, and where f^q is a measurable function and ξ^q is a real-valued noise variable.*

However, several studies have shown that, with additional assumptions on the models relating causes and effects, one can identify the direction of the causal relation. We review here two such cases which have led to extensions for time series.

3.4.1 Vector autoregressive models

Shimizu et al. [2006] proposed a method for uniquely identifying causal structures based on purely observational, continuous-valued data with the assumptions that the structural equation model is linear, acyclic, with non-Gaussian error terms (LiNGAM). When considering two variables, LiNGAM is of the form:

$$\begin{aligned} X^p &= \xi^p; \\ X^q &= a^{p,q} X^p + \xi^q \qquad \text{with } X^p \perp\!\!\!\perp \xi^q; \end{aligned}$$

where ξ^p and ξ^q are non-Gaussian.

Assuming that there are no hidden confounders and all (or all but one) of the error terms are non-Gaussian, the full generating model can be identified in the limit of an infinite sample (a property known as *asymptotic consistency*).

Theorem 2 (Identifiability of linear non-Gaussian models [Peters et al., 2017])**.** *Assume that the joint distribution of X^p and X^q admits the linear model*

$$X^q = a^{p,q} X^p + \xi^q, \qquad \textit{with } X^p \perp\!\!\!\perp \xi^q,$$

with continuous random variables X^p, ξ^q, and X^q. Then, there exist $a^{q,p} \in \mathbb{R}$ and a random variable X^p such that

$$X^p = a^{q,p} X^q + \xi^p, \qquad \textit{with } X^q \perp\!\!\!\perp \xi^p,$$

if and only if ξ^p and X^q are Gaussian.

To detect causal relations, LiNGAM proceeds as follows. First of all, from the equation $X = \mathbf{A}X + \xi$, one obtains $X = \mathbf{B}\xi$ with $\mathbf{B} = (I - \mathbf{A})^{-1}$. LiNGAM uses a standard independent component analysis algorithm to obtain an estimate of the mixing matrix $\mathbf{B}$ and uses it to compute the matrix $\mathbf{A}$. Furthermore, Shimizu et al. [2011] proposed an algorithmic improvement of their original method that converges to the correct solution in a controlled number of steps depending on the number of variables. The main idea is to find the causal order by constructing a regression model and by checking whether residuals and predictors are independent or not. This step is done recursively by first identifying the predictor that is the most independent from the residuals of its target variables, *i.e.* all variables except the predictor. The same analysis is then performed recursively on those residuals, which ensures to remove the effects of the previously identified predictors. One can then construct a strictly lower triangular matrix $\mathbf{A}$ by following the ordering obtained above. The strength of the connections $A_{i,j}$ are estimated using some conventional covariance-based regression, such as least squares. To get sparse causal models, one can further prune $\mathbf{A}$ by applying Adaptive Lasso [Zou, 2006], which penalizes connections with an ℓ_1 penalty.

We now present an extension of LiNGAM to time series.

Using linear non-gaussian acyclic model

Hyvärinen et al. [2010] introduced a temporal extension of LiNGAM, called `VarLiNGAM`, based on a structural vector autoregressive model of the form:

$$X_t = \sum_{i=0}^{\gamma_{\max}} \mathbf{A}_i X_{t-i} + \mathbf{e}_t, \qquad \text{(SVAR)}$$

where the influences can be either instantaneous ($i = 0$) or lagged, with a maximum time-delay of $\gamma_{\max}$. This model can be rewritten as a vector autoregressive model without instantaneous effect, with $i > 0$:

$$X_t = \sum_{i=1}^{\gamma_{\max}} \mathbf{M}_i X_{t-i} + \mathbf{e}_t. \qquad \text{(VAR)}$$

The basic idea is to use the least-squares estimation of the autoregressive to obtain residuals of the prediction of X_t. Then conduct a LiNGAM analysis on those residuals leading to the estimation of the instantaneous causal model $\mathbf{A}_0$. Finally, $(\mathbf{A}_i)_{i>0}$ are deduced by a reparametrization of $(\mathbf{M}_i)_{i>0}$:

$$\mathbf{A}_i = (I - \mathbf{A}_0)\mathbf{M}_i \quad \text{for all } i \in \{1, \cdots, d\}.$$

3.4.2 Additive noise model

Hoyer et al. [2009] showed that if the underlying causal structural equations are based on an additive noise model (ANM) with nonlinear functions and that if the causal minimality condition holds, then the true causal structure can in general be identified from the probability distribution of the observational data, as stated in Theorem 3. This theorem makes use of the notion of smooth ANM, *i.e.* an ANM of the form:

$$\begin{aligned} X^p &= \xi^p, \\ X^q &= f^q(X^p) + \xi^q \qquad \text{with } X^p \perp\!\!\!\perp \xi^q. \end{aligned}$$

such that ξ^q and X^p have strictly positive three times differentiable densities p_{ξ^q} and p_{X^p}, and f^q is three times differentiable as well.

Theorem 3 (Identifiability of ANMs [Peters et al., 2017, Hoyer et al., 2009])**.** *Assume that the conditional distribution of $X^q \mid X^p$ admits a smooth ANM, and that there exists $x_q \in \mathbb{R}$ such that, for almost all $x^p \in \mathbb{R}$,*

$$(\log p_{\xi^q})''(x^q - f^q(x^p))f^{q'}(x^p) \neq 0.$$

Then, the set of log densities $\log p_X$ for which the obtained joint distribution P_{X^p,X^q} admits a smooth ANM from X^q to X^p is contained in a 3-dimensional affine space.

In the bivariate case, one can regress two models, one of X^q on X^p and another of X^p on X^q, and test the independence with residuals to infer the causal direction. For the multivariate case, one can adopt a pairwise strategy or use an adapted algorithm that can handle more than two variables [Mooij et al., 2009].

We now introduce a well-known method based on ANM for time series.

Times series model with independent noise

A class of restricted FCM called time series models with independent noise `TiMINo` is studied in Peters et al. [2013]. For a multivariate time series X whose finite dimensional distributions are absolutely continuous with respect to a product measure, we say that the time series satisfies a `TiMINo` if there exists $\gamma_{\max} > 0$ such that for all $X^p \in V$ there are sets $\mathrm{Par}(X_0^p) \subseteq V \backslash X^p, \mathrm{Par}(X_k^p) \subseteq X$ for $1 \leq k \leq \gamma_{\max}$ such that for all t:

$$X_t^p = f^p(\mathrm{Par}(X_t^p)_{t-\gamma_{\max}}, \ldots, \mathrm{Par}(X_t^p)_{t-1}, \mathrm{Par}(X_t^p)_t, \xi_t^p), \qquad \text{(TiMINo)}$$

where ξ_t^i are jointly independent over i and t and, for each i, i.i.d. in t. These models include nonlinear and instantaneous effects, but the full time causal graph is required to be acyclic. Under some particular form of f^p (nonlinear function with additive Gaussian noise, linear function with additive non-Gaussian noise, joint distribution faithful with respect to the full time causal graph, and acyclicity of the summary causal graph), the summary causal graph can be recovered from the joint distribution of X. To infer the causal graph in the additive noise model case, statistical tests are conducted to look for independence between residuals and nodes so as to order the variables by parenting relations. Then, spurious links are removed. Note that several fitting methods can be considered (*e.g.*, linear model, generalized additive model and Gaussian process regression are considered in the initial paper) as well as several independence tests (*e.g.*, cross-correlations or HSIC). Note that if the data does not satisfy the model assumption, `TiMINo` falls into an agnostic state instead of drawing wrong causal conclusions. In the case of two time series, an agnostic state can be interpreted as a possible detection of hidden confounders.

3.5 Conclusion

The problem of estimating the causal relations for time series is not solved, but there is progress in understanding how to deal with these problems in various families of methods. The main difficulties are that the generating process may be non-linear, the data acquisition rate may be much slower than the underlying rate of changes, there may be measurement error, the probability distributions of variables conditional on their causes may change, the causal relations may change (known as causal non-stationary), and there may be unmeasured confounding causes.

As we saw, there already exist several approaches to discover causal relations from time series. Some infer directly a summary causal graph and some infer a window causal graph. Some of them treat lagged non linear relations but leave instantaneous relations untreated. Others take into account instantaneous relations but only in the linear case. TCDF and TiMINo, simultaneously solved the problems of non linearity, lagged relations, and instantaneous relations, however, they are as well as most methods presented here, very sensitive to the sampling rate.

In the next chapter, we present a new approach that directly detects the summary causal graph, with non linear, lagged, and instantaneous relations, which can also handle different sampling rate.

Chapter 4

Entropy-based discovery of summary causal graphs in time series

> There is no logical necessity for the existence of a unique direction of total time; whether there is only one time direction, or whether time directions alternate, depends on the shape of the entropy curve plotted by the universe.
>
> Hans Reichenbach

4.1 Introduction

This chapter addresses the central problem of this thesis, namely: given a maximal lag γ_{max} and an observational time series $X^1, \cdots, X^d$ with potentially different sampling rates, infer the underlying summary causal graph corresponding to the true full time causal graph.

An important aspect of real-world time series is that different time series, as they measure different elements, usually have different sampling rates. Despite this, the algorithms that have been developed so far to discover causal structures from temporal observations [Granger, 1969, Hyvärinen et al., 2010, Moneta et al., 2013, Peters et al., 2013, Runge et al., 2019, Nauta et al., 2019] rely on the idealized assumptions that all time series have the same sampling rates with identical timestamps [1].

We introduce in this chapter two causal inference algorithms that can be applied to discrete time series with continuous values and different sampling rates. Both algorithms belong to the constraint-based family of causal algorithms [Spirtes et al., 2000] and the former can be used in situations where all common causes are observed and the latter can be used in situations in which some common causes are unobserved. The skeleton construction (as well as the orientation of instantaneous relations) of the former is similar to the PC algorithm [Spirtes et al., 2000], but adapted to time series; the skeleton construction (as well as the orientation of instantaneous relations) of the latter is similar to the FCI algorithm [Spirtes et al., 2000]. For orienting lagged relations, both algorithms call upon an entropic reduction principle which is inspired by the work of Suppes [1970] (for more details on Suppes' work, see Chapter 2 Section 2.6). At the core of these algorithms lie (in)dependence measures, to detect relevant dependencies, which are based here on an information theoretic approach.

Since their introduction [Shannon, 1948], information theoretic measures have become very popular due to their non parametric nature, their robustness against strictly monotonic transformations, which makes them capable of handling nonlinear distortions in the system, and their good behavior in previous studies on causal discovery [Affeldt and Isambert, 2015]. However, their application to temporal data raises several problems related to the fact that time series may have different sampling rates, be shifted in time; and have strong internal dependencies. Many studies have attempted to re-formalize mutual information for time series. Galka et al. [2006] considered each value of each time series as different random

1. Assuming identical timestamps in the case of identical sampling rates seems reasonable as one can shift time series so that they coincide in time.

variables and proceeded by whitening the data, such that time dependent data will be transformed into independent residuals through a parametric model. However, whitening the data can have severe consequences on causal relations. Schreiber [2000] proposed a reformulation of mutual information, called the transfer entropy (later generalized in Sun et al. [2015]), that represents the information flow from one state to another and thus is asymmetric. Inspired by Kraskov et al. [2004], Frenzel and Pompe [2007] proposed a formulation where time series are represented by vectors, and estimated the mutual information assuming that all vectors are statistically independent. This said, time series are still assumed to have equal sampling rates. Closer to our proposal is the time delayed mutual information proposed in Albers and Hripcsak [2012] that aims at addressing the problem of non uniform sampling rates. The computation of the time delayed mutual information relates single points from a single time series (shifted in time) but does not consider potentially complex relations between time stamps in different time series, as we do through the use of window-based representations and compatible time lags. The time delayed mutual Information can be seen as a special case of the temporal mutual information we introduce in the next section, by considering windows of size 1 and a single time series. The measure we propose is more suited to discover summary causal graph as it can consider potentially complex relations between timestamps in different time series through the use of window-based representations and compatible time lags, and is more general as it can consider different sampling rate.

The remainder of the chapter is organized as follows: Section 4.2 introduces the (conditional) mutual information measures we propose for time series and the entropy reduction principle that our method is based on. Section 4.3 presents the PC-like causal discovery algorithm we have developed on top of these measures. Section 4.4 presents the FCI-like causal discovery algorithm we have developed on top of these measures. Furthermore, we provide respectively in Section 4.5 and Section 4.6, a methodology to use to construct a window causal graph given a summary causal graph, and an adaptation of our method to sequences. Finally, Section 4.7 concludes the chapter.

4.2 Information measures for causal discovery in time series

We present in this section a new mutual information measures that operate on a window-based representation of time series to assess whether time series are (conditionally) dependent or not. We then show how this measure is related to an entropy reduction principle that is a special case of the probabilistic raising principle [Suppes, 1970].

We first assume that all time series are aligned in time, with the same sampling rate, prior to show how our development can be applied to time series with different sampling rates. Without loss of generality, time instants are assumed to be integers. Lastly, as done in previous studies [Schreiber, 2000], we assume that all time series are first-order Markov self-causal (any time instant is caused by its previous instant within the same time series).

4.2.1 Causal temporal mutual information

Let us consider d univariate time series $X^1, \cdots, X^d$, and their observations $(v_1^p, \cdots, v_{N_p}^p)$ $(1 \leq p \leq d)$, where v_t^p $(1 \leq t \leq N_p)$ is the value for the p-th time series at time index t and N_p is the length of X^p. Throughout this section, we will make use of the following example, illustrated in Figure 4.1, to discuss the notions we introduce.

Example 1. *Let us consider the following two time series defined by, for all t,*

$$X_t^p = X_{t-1}^p + \xi_t^p,$$
$$X_t^q = X_{t-1}^q + X_{t-2}^p + X_{t-1}^p + \xi_t^q,$$

with $(\xi_t^p, \xi_t^q) \sim \mathcal{N}(0, 1)$.

One can see in Example 1 that, in order to capture the dependencies between the two time series, one needs to take into account a lag between them, as the true, causal relations are not instantaneous. Several studies have recognized the importance of taking into account lags to measure (conditional) dependencies between time series; for example, Runge et al. [2019] uses pointwise mutual information between time series with lags to assess whether they are dependent or not.

In addition to lags, Example 1 also reveals that a window-based representation may be necessary to fully capture the dependencies between the two time series. Indeed, as X_{t-1}^q and X_t^q are the effects of the same cause (X_{t-2}^p),

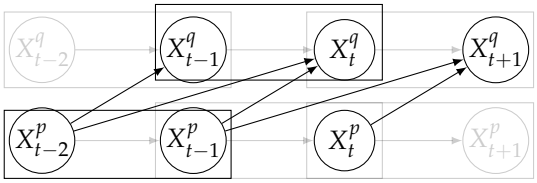


Figure 4.1 – Why do we need windows and lags? An illustration with two time series where X^p causes X^q in two steps (circles correspond to observed points and rectangles to windows). The arrows in black are discussed in the text.

it may be convenient to consider them together when assessing whether the time series are dependent or not. For example, defining (overlapping) windows of size two for X^q and one for X^p with a lag of 1 from X^p to X^q, as in Figure 4.1, allows one to fully represent the causal dependencies between the two time series.

Definition 33. *Let $\gamma_{\max}$ denote the maximum lag between two time series X^p and X^q and let $\lambda_{\max} = 2\gamma_{\max} + 1$. The window-based representation, of size $0 < \lambda_{pq} \leq \lambda_{\max} < N_p$, of the time series X^p with respect to X^q, which will be denoted $X^{(p;\lambda_{pq})}$, simply amounts to considering $(N_p - \lambda_{pq} + 1)$ windows: $w_t^{(p;\lambda_{pq})} = (v_t^p, \cdots, v_{t+\lambda_{pq}-1}^p)$, $1 \leq t \leq N_p - \lambda_{pq} + 1$. The window-based representation, of size $0 < \lambda_{qp} \leq \lambda_{\max} < N_q$, of the time series X^q with respect to X^p is defined in the same way. A temporal lag $\gamma_{pq} \in \mathbb{Z}$* compatible *with λ_{pq} and λ_{qp} relates windows in $X^{(p;\lambda_{pq})}$ and $X^{(q;\lambda_{qp})}$ in such a way that the starting time of the related windows are separated by γ_{pq}. We denote by $\mathcal{C}^{(p,q)}$ the set of window sizes and compatible temporal lags.*

Based on the above elements, we define the *causal temporal mutual information* between two time series as the maximum of the standard mutual information over all possible compatible temporal lags and windows, conditioned by the past of the two series. Indeed, as we are interested in obtaining a summary causal graph, we do not have to consider all the potential dependencies between two time series (which would be necessary for inferring a window causal graph). Using the maximum over all possible associations is a way to summarize all temporal dependencies which ensures that one does not miss a dependency between the two time series. Furthermore, conditioning on the past allows one to eliminate spurious dependencies in the form of auto-correlation, as in transfer entropy [Schreiber, 2000]. We follow this idea here and, as in transfer entropy, consider windows of size 1 and a temporal lag of 1 for conditioning on the

past, which is in line with the first-order Markov self-causal assumption mentioned above.

Definition 34. *Consider two time series X^p and X^q. We define the* causal temporal mutual information *between X^p and X^q as:*

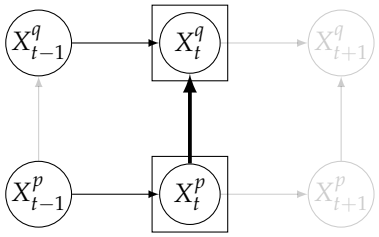
$$\begin{aligned} &CTMI(X^p; X^q) = \\ &\max_{(\lambda_{pq},\lambda_{qp},\gamma_{pq})\in\mathcal{C}^{(p,q)}} I(X_t^{(p;\lambda_{pq})}; X_{t+\gamma_{pq}}^{(q;\lambda_{qp})} | X_{t-1}^{(p;1)}, X_{t+\gamma_{pq}-1}^{(q;1)}) \\ &\triangleq I(X_t^{(p;\bar{\lambda}_{pq})}; X_{t+\bar{\gamma}_{pq}}^{(q;\bar{\lambda}_{qp})} | X_{t-1}^{(p;1)}; X_{t+\bar{\gamma}_{pq}-1}^{(q;1)}), \end{aligned} \tag{4.1}$$

where I represents the mutual information. In case the maximum can be obtained with different values in $\mathcal{C}^{(p,q)}$, we first set $\bar{\gamma}_{pq}$ to its largest possible value. We then set $\bar{\lambda}_{pq}$ to its smallest possible value and finally $\bar{\lambda}_{qp}$ to its smallest possible value. $\bar{\gamma}_{pq}$, $\bar{\lambda}_{pq}$, and $\bar{\lambda}_{qp}$ respectively correspond to the optimal lag and optimal windows.

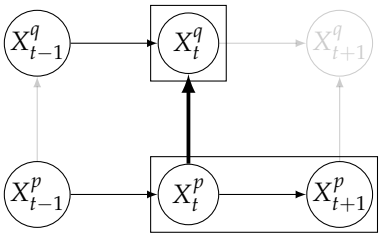
In the context we have retained, in which dependencies are constant over time, CTMI satisfies standard properties of mutual information, namely it is nonnegative, symmetric and equals to 0 *iff* time series are independent. Thus, two time series X^p and X^q such that $CTMI(X^p; X^q) > 0$ are dependent. Setting $\bar{\gamma}_{pq}$ to its largest possible value allows one to get rid of instants that are not crucial in determining the mutual information between two time series. The choice for the window sizes, even though arbitrary on the choice of treating one window size before the other, is based on the same ground, as the mutual information defined above cannot decrease when one increases the size of the windows. Indeed:

$$\begin{aligned} &I(X_t^{(p;\lambda_{pq})}; X_{t+\gamma_{pq}}^{(q;\lambda_{qp})} \mid X_{t-1}^{(p;1)}, X_{t+\gamma_{pq}-1}^{(q;1)}) \\ &= I((X_t^{(p;\lambda_{pq}-1)}, X_{t+\lambda_{pq}-1}^{(p;1)}); X_{t+\gamma_{pq}}^{(q;\lambda_{qp})} \mid X_{t-1}^{(p;1)}, X_{t+\gamma_{pq}-1}^{(q;1)}) \\ &= I(X_t^{(p;\lambda_{pq}-1)}; X_{t+\gamma_{pq}}^{(q;\lambda_{qp})} \mid X_{t-1}^{(p;1)}, X_{t+\gamma_{pq}-1}^{(q;1)}) \\ &\quad + I(X_{t+\lambda_{pq}-1}^{(p;1)}; X_{t+\gamma_{pq}}^{(q;\lambda_{qp})} \mid X_{t-1}^{(p;1)}, X_{t+\gamma_{pq}-1}^{(q;1)}, X_t^{(p;\lambda_{pq}-1)}) \\ &\geq I(X_t^{(p;\lambda_{pq}-1)}; X_{t+\gamma_{pq}}^{(q;\lambda_{qp})} \mid X_{t-1}^{(p;1)}, X_{t+\gamma_{pq}-1}^{(q;1)}). \end{aligned} \tag{4.2}$$

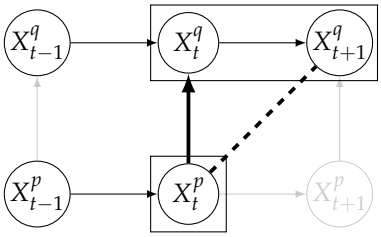
The last inequality is due to the fact that mutual information is positive. It is also interesting to note that CTMI does not necessarily increase symmetrically with respect to the increase of λ_{pq} and λ_{qp}. For an illustration see Figure 4.2.



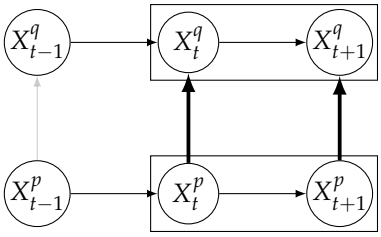
(a) $I(X^p_t, X^q_t \mid X^p_{t-1}, X^q_{t-1}) > 0$



(b) $I(X^{(p;2)}_t, X^q_t \mid X^p_{t-1}, X^q_{t-1}) = I(X^p_t, X^q_t \mid X^p_{t-1}, X^q_{t-1})$



(c) $I(X^p_t, X^{(q;2)}_t \mid X^p_{t-1}, X^q_{t-1}) > I(X^p_t, X^q_t \mid X^p_{t-1}, X^q_{t-1})$



(d) $I(X^{(p;2)}_t, X^{(q;2)}_t \mid X^p_{t-1}, X^q_{t-1}) > I(X^p_t, X^{(q;2)}_t \mid X^p_{t-1}, X^q_{t-1})$

Figure 4.2 – Illustration of the asymmetric increase of CTMI with the increase of the window sizes. The mutual information (conditioned on the past) increases when increasing only the window size of the effect or when increasing simultanesly the window sizes of the effect and the cause (it does not increase when increasing only the window size of the cause). Dashed lines are for correlations which are not causations, and bold arrows correspond to causal relations between the window representations of time series.

Example 2. *Consider the structure described in Example 1, and assume that* $\lambda_{\max} = 3$*. First, we have for the standard mutual information*

$$I(X_t^{(p;1)}; X_t^{(q;1)} \mid X_{t-1}^{(p;1)}, X_{t-1}^{(q;1)}) = 0.$$

We also have that any $\gamma_{pq} < 0$ *have a zero mutual information because conditioning on the past of* $X_t^{(p;1)}; X_t^{(q;1)}$ *(namely* $X_{t-1}^{(p;1)}; X_{t-1}^{(q;1)}$*) is closing all paths from* $X_{t-i}^{(q;1)}$ *to* $X_t^{(p;1)}$ *for all* $i > 0$*. For* $\gamma_{pq} > 0$ *starting by* $\gamma_{pq} = 1$*,*

$$I(X_t^{(p;1)}; X_{t+1}^{(q;1)} \mid X_{t-1}^{(p;1)}, X_t^{(q;1)}) > 0.$$

Similarly for $\gamma_{pq} > 2$*. Now any increase of* λ_{qp} *alone or of* λ_{pq} *and* λ_{qp}*, will generate an increase in the mutual information as long as the difference between the last time point of the window of* X^p *(the cause) and the last time point of the window of* X^q *is less or equal to* γ_{pq} *as*

$$\begin{aligned} &I(X_t^{(p;\lambda_{pq}-1)}; X_{t+\gamma_{pq}}^{(q;\lambda_{pq}-1-\gamma_{pq})} \mid X_{t-1}^{(p;1)}, X_{t+\gamma_{pq}-1}^{(q;1)}) \\ &= I(X_t^{(p;\lambda_{pq})}; X_{t+\gamma_{pq}}^{(q;\lambda_{pq}-1-\gamma_{pq})} \mid X_{t-1}^{(p;1)}, X_{t+\gamma_{pq}-1}^{(q;1)}) \end{aligned}$$

because

$$I(X_{t+\lambda_{pq}-1}^{(p;1)}; X_{t+\gamma_{pq}}^{(q;\lambda_{pq}-1-\gamma_{pq})} \mid X_{t-1}^{(p;1)}, X_{t+\gamma_{pq}-1}^{(q;1)}, X_t^{(p;\lambda_{pq}-1)}) = 0,$$

where $\gamma_{pq} \geq 1$ *(1 is the minimal lag that generates a correlation that cannot be removed by conditioning on the past of* X^p *and* X^q*). For* $\lambda_{\max} = 3$ *the optimal window size* $\bar{\lambda}_{pq}$ *is equal to* 2 *as* X^p *has no other cause than itself;* $\bar{\lambda}_{qp}$ *is equal to* 2 *as* X^p *causes only (except itself)* X_{t+1}^q *and* X_t^q*. Furthermore,* $\bar{\gamma}_{pq} = 1$ *and*

$$\begin{aligned} CTMI(X^p; X^q) &= I((X_t^p, X_{t+1}^p); (X_{t+1}^q, X_{t+2}^q) \mid X_{t-1}^p, X_t^q) \\ &= I(X_t^p; (X_{t+1}^q, X_{t+2}^q) \mid X_{t-1}^p, X_t^q) \\ &\quad + I(X_{t+1}^p; (X_{t+1}^q, X_{t+2}^q) \mid X_{t-1}^p, X_t^q, X_t^p) \\ &= I(X_t^p; X_{t+1}^q \mid X_{t-1}^p, X_t^q) + I(X_t^p; X_{t+2}^q \mid X_{t-1}^p, X_t^q, X_{t+1}^q) \\ &\quad + I(X_{t+1}^p; X_{t+1}^q \mid X_{t-1}^p, X_t^q, X_t^p) \\ &\quad + I(X_{t+1}^p; X_{t+2}^q \mid X_{t-1}^p, X_t^q, X_t^p, X_{t+1}^q) \\ &= 2I(X_t^p; X_{t+1}^q \mid X_{t-1}^p, X_t^q) + I(X_t^p; X_{t+2}^q \mid X_{t-1}^p, X_t^q, X_{t+1}^q) \\ &= 3\log(3)/4. \end{aligned}$$

4.2.2 Entropy reduction principle

Interestingly, CTMI can be related to a version of the probabilistic raising principle (PRP, Suppes [1970]) that states that a cause, here a time series, raises the probability of any of its effects, here another time series, even when the past of the two time series is taken into account, meaning that the relation between the two time series is not negligible compared to the internal dependencies of the time series. In this context, the following definition generalizes to window-based representations of time series the standard definition of *prima facie* causes for discrete variables.

Definition 35 (Prima facie cause for window based time series). *Let X^p and X^q be two time series with window sizes λ_{pq} and λ_{qp} and let $P_{t,t'} = (X_{t-1}^{(p;1)}, X_{t'-1}^{(q;1)})$ represent the past of X^p and X^q for any two instants (t, t'). We say that X^p is a* prima facie cause *of X^q with delay $\gamma_{pq} > 0$ iff there exist Borel sets B_p, B_q and B_P such that one has:*

$$P(X_{t+\gamma_{pq}}^{(q;\lambda_{qp})} \in B_q | X_t^{(p;\lambda_{pq})} \in B_p, P_{t,t+\gamma_{pq}} \in B_P) >$$
$$P(X_{t+\gamma_{pq}}^{(q;\lambda_{qp})} \in B_q | P_{t,t+\gamma_{pq}} \in B_P).$$

We now introduce a slightly different principle based on the causal temporal mutual information which we refer to as the *entropy reduction principle (ERP)*.

Definition 36 (Entropic prima facie cause). *Using the same notations as in Definition 35, we say that X^p is an* entropic prima facie cause *of X^q with delay $\gamma_{pq} > 0$ iff $I(X_t^{(p;\lambda_{pq})}; X_{t+\gamma_{pq}}^{(q;\lambda_{qp})} | P_{t,t+\gamma_{pq}}) > 0$.*

Note that considering that the above mutual information is positive is equivalent to considering that the entropy of X^q when conditioned on the past reduces when one further conditions on X^p. One has the following relation between the ERP and PRP principles.

Property 1. *With the same notations, if X^p is an* entropic prima facie *cause of X^q with delay $\gamma_{pq} > 0$, then X^p is a* prima facie *cause of X^q with delay $\gamma_{pq} > 0$. Furthermore, if $CTMI(X^p; X^q) > 0$ with $\bar{\gamma}_{pq} > 0$ then X^p is an* entropic prima facie *cause of X^q with delay $\bar{\gamma}_{pq}$.*

Proof. Let us assume that X^p is not a *prima facie* cause of X^q for the delay γ_{pq}. Then, for all Borel sets B_p, B_q and B_P one has $P(X_{t+\gamma_{pq}}^{(q;\lambda_{qp})} \in B_q | X_t^{(p;\lambda_{pq})} \in$

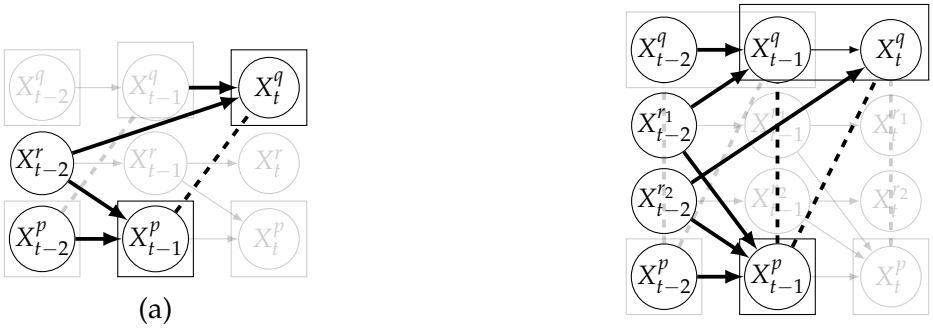


Figure 4.3 – Examples of conditional independence between dependent time series. Dashed lines are for correlations which are not causations, and bold arrows correspond to conditioning variables.

$B_p, P_{t,t+\gamma_{pq}} \in B_P) \leq P(X^{(q;\lambda_{qp})}_{t+\gamma_{pq}} \in B_q | P_{t,t+\gamma_{pq}} \in B_P)$. This translates, in terms of density functions denoted f, as:

$$\forall (x^p_t, x^q_{t+\gamma_{pq}}, p_{t,t+\gamma_{pq}}),\ f(x^q_{t+\gamma_{pq}} | x^p_t, p_{t,t+\gamma_{pq}}) \leq f(x^q_{t+\gamma_{pq}} | p_{t,t+\gamma_{pq}}),$$

which implies that $H(X^{(q;\lambda_{qp})}_{t+\gamma_{pq}} \in B_q | X^{(p;\lambda_{pq})}_t \in B_p, P_{t,t+\gamma_{pq}} \in B_P)$ is greater than $H(X^{(q;\lambda_{qp})}_{t+\gamma_{pq}} \in B_q | P_{t,t+\gamma_{pq}} \in B_P)$ so that X^p is not an *entropic prima facie* cause of X^p with delay γ_{pq}. By contraposition, we conclude the proof of the first statement. The second statement directly derives from the definition of CTMI. □

4.2.3 Conditional causal temporal mutual information

We now extend the causal temporal mutual information by conditioning on a set of variables. In a causal discovery setting, conditioning is used to assess whether two dependent time series can be made independent by conditioning on connected time series, *i.e.* time series which are dependent with at least one of the two times series under consideration. Figure 4.3 illustrates the case where the dependence between X^p and X^q is due to spurious correlations originating from common causes. Conditioning on these common causes should lead to conditional independence of the two time series. Of course, the conditional variables should precede in time the two time series under consideration. This leads us to the following definition of the conditional causal temporal mutual information.

Definition 37. *The* conditional causal temporal mutual information *between two time series X^p and X^q such that $\bar{\gamma}_{pq} \geq 0$, conditioned on a set $X^{\boldsymbol{R}} = \{X^{r_1}, \cdots, X^{r_K}\}$ is given by:*

$$\begin{aligned} &CTMI(X^p; X^q \mid X^{\boldsymbol{R}}) \\ =&I(X_t^{(p;\bar{\lambda}_{pq})}; X_{t+\bar{\gamma}_{pq}}^{(q;\bar{\lambda}_{qp})} | (X_{t-\bar{\Gamma}_k}^{(r_k;\bar{\lambda}_k)})_{1\leq k\leq K}, X_{t-1}^{(p;1)}, X_{t+\bar{\gamma}_{pq}-1}^{(q;1)}), \end{aligned} \tag{4.3}$$

In case the maximum can be obtained with different values, we first set $\bar{\Gamma}_k$ to its largest possible value. We then set $\bar{\lambda}_k$ to its smallest possible value. $(\bar{\Gamma}_1, \ldots, \bar{\Gamma}_K)$ and $(\bar{\lambda}_1, \cdots, \bar{\lambda}_K)$ correspond to the optimal conditional lags and window sizes which minimize, for $\Gamma_1, \ldots, \Gamma_K \geq -\bar{\gamma}_{pq}$:

$$I\left(X_t^{(p;\bar{\lambda}_{pq})}; X_{t+\bar{\gamma}_{pq}}^{(q;\bar{\lambda}_{qp})} | (X_{t-\Gamma_k}^{(r_k;\lambda_k)})_{1\leq k\leq K}, X_{t-1}^{(p;1)}, X_{t+\bar{\gamma}_{pq}-1}^{(q;1)}\right).$$

By considering the minimum over compatible lags and window sizes, one guarantees that if there exist conditioning variables which make the two time series independent, they will be found. Note that the case in which $\bar{\gamma}_{p,q} < 0$ correspond to $\text{CTMI}(X^q; X^p \mid X^{\mathbf{R}})$ where $\bar{\gamma}_{q,p} > 0$.

Figure 4.3 illustrates the above on two different examples. On the left, X_{t-1}^p is correlated to X_t^q as X_{t-2}^r is a common cause with a lag of 1 for X^p and a lag of 2 for X^q. Conditioning on X_{t-2}^r removes the dependency between X^p and X^q. Note that all time series have here a window of size 1. On the right, X^{r_1} and X^{r_2} are common causes of X^p and X^q: X^{r_1} causes X^p and X^q with temporal lag of 1, which renders X^p and X^q correlated at the same time point, while X^{r_2} causes X^p and X^q with temporal lag of 1 and 2 respectively, which renders X^p and X^q correlated at lagged time points. The overall correlation between X^p and X^q is captured by considering a window size of 2 in X^q. All other time series have a window size of 1. By conditioning on both X^{r_1} and X^{r_2}, X^p and X^q become independent.

4.2.4 Estimation and testing

In practice, the success of CTMI approach (and in fact, any entropy-based approaches) depends crucially on reliable estimation of the relevant entropies in question from data. This leads to two practical challenges. The first one is based on the fact that entropies must be estimated from finite time series data. The second is that to detect independence, we need a statistical test to check if the temporal causation entropy is equal to zero.

Here, we rely on the k-nearest neighbor method [Frenzel and Pompe, 2007] for the estimation of CTMI. The distance between two windows considered here is the supremum distance, i.e., the maximum of the absolute difference between any two values in the two windows.

$$d((X_t^{(p;\bar{\lambda}_{pq})}, X_{t+\bar{\gamma}_{p,q}}^{(q;\bar{\lambda}_{pq})})_i, (X_t^{(p;\bar{\lambda}_{pq})}, X_{t+\bar{\gamma}_{pq}}^{(q;\bar{\lambda}_{pq})})_j) = \max_{0\leq\ell<\lambda_p, 0\leq\ell'<\lambda_q} (|(X_t^{(p;\bar{\lambda}_{pq})})_{i+\ell} - (X_t^{(p;\bar{\lambda}_{pq})})_{j+\ell'}|, |(X_t^{(q;\bar{\lambda}_{qp})})_{i+\ell'} - (X_t^{(q;\bar{\lambda}_{qp})})_{j+\ell'}|).$$

In case of the causal temporal mutual information, we denote by $\epsilon_{ik}/2$ the distance from

$$(X_t^{(p;\bar{\lambda}_{pq})}, X_{t+\bar{\gamma}_{pq}}^{(q;\bar{\lambda}_{pq})}, X_{t-1}^p, X_{t+\bar{\gamma}_{pq}-1}^q)$$

to its k-th neighbor, $n_i^{1,3}$, $n_i^{2,3}$ and n_i^3 the numbers of points with distance strictly smaller than $\epsilon_{ik}/2$ in the subspace

$$(X_t^{(p;\bar{\lambda}_{pq})}, X_{t-1}^p, X_{t+\bar{\gamma}_{pq}-1}^q), (X_{t+\bar{\gamma}_{p,q}}^{(q;\bar{\lambda}_{pq})}, X_{t-1}^p, X_{t+\bar{\gamma}_{pq}-1}^q), \text{ and } (X_{t-1}^p, X_{t+\bar{\gamma}_{pq}-1}^q)$$

respectively, and $n_{\gamma_{pq},\gamma_{qp}}$ the number of observations. The estimate of the causal temporal mutual information is then given by:

$$\widehat{CTMI}(X^p; X^q) = \psi(k) + \frac{1}{n_{\gamma_{pq},\gamma_{qp}}} \sum_{i=1}^{n_{\gamma_{pq},\gamma_{qp}}} \psi(n_i^3) - \psi(n_i^{1,3}) - \psi(n_i^{2,3})$$

where ψ denotes the digamma function.

Similarly, for the estimation of the conditional causal temporal mutual information, we denote by $\epsilon_{ik}/2$ the distance from

$$(X_t^{(p;\bar{\lambda}_{pq})}, X_{t+\bar{\gamma}_{pq}}^{(q;\bar{\lambda}_{pq})}, X_{t-1}^p, X_{t+\bar{\gamma}_{pq}-1}^q, (X_{t-\bar{\Gamma}_k}^{(r_k;\bar{\lambda}_k)})_{1\leq k\leq K})$$

to its k-th neighbor, $n_i^{1,3}$, $n_i^{2,3}$ and n_i^3 the numbers of points with distance strictly smaller than $\epsilon_{ik}/2$ in the subspace

$$(X_t^{(p;\bar{\lambda}_{pq})}, X_{t-1}^p, X_{t+\bar{\gamma}_{pq}-1}^q, (X_{t-\bar{\Gamma}_k}^{(r_k;\bar{\lambda}_k)})_{1\leq k\leq K}),$$
$$(X_{t+\bar{\gamma}_{pq}}^{(q;\bar{\lambda}_{pq})}, X_{t-1}^p, X_{t+\bar{\gamma}_{pq}-1}^q, (X_{t-\bar{\Gamma}_k}^{(r_k;\bar{\lambda}_k)})_{1\leq k\leq K}), \text{ and}$$
$$(X_{t-1}^p, X_{t+\bar{\gamma}_{pq}-1}^q, (X_{t-\bar{\Gamma}_k}^{(r_k;\bar{\lambda}_k)})_{1\leq k\leq K})$$

respectively, and $n_{\gamma_{rp},\gamma_{rq}}$ the number of observations. The estimate of the conditional causal temporal mutual information is then given by:

$$\widehat{CTMI}(X^p; X^q \mid X^{\mathbf{R}}) = \psi(k) + \frac{1}{n_{\gamma_{rp},\gamma_{rq}}} \sum_{i=1}^{n_{\gamma_{rp},\gamma_{rq}}} \psi(n_i^3) - \psi(n_i^{1,3}) - \psi(n_i^{2,3})$$

where ψ denotes the digamma function.

To detect independencies through CTMI we rely on the following permutation test:

Definition 38 (Permutation test of CTMI)**.** *Given X^p, X^q and $X^{\boldsymbol{R}}$, the p-value associated to the permutation test of CTMI is given by:*

$$p = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}_{\widehat{CTMI}((X^p)_b; X^q | X^{\boldsymbol{R}}) \geq \widehat{CTMI}(X^p; X^q | X^{\boldsymbol{R}})}, \tag{4.4}$$

where $(X^p)_b$ is a permuted version of X^p and follow the local permutation scheme presented in Runge [2018].

The advantage of the scheme presented in Runge [2018] is that it preserves marginals by drawing as much as possible without replacement and it performs local permutation which insure that by permuting X^p, the dependence between X^p and X^r is not destroyed.

Note that Definition 4.4, is applicable to the causal temporal mutual information (when $\mathbf{R}$ is empty) and to the conditional causal temporal mutual information.

4.2.5 Extension to time series with different sampling rates

The above development readily applies to time series with different sampling rates as one can define window-based representations of the two time series as well as a sequence of joint observations.
Indeed, as one can note, Definition 33 does not rely on the fact that the time series have the same sampling rates. Figure 4.4 displays two time series X^p and X^q with different sampling rates where, while $\lambda_{pq} = 2$ and $\lambda_{qp} = 3$, the time spanned by each window is the same. The joint sequence of observations, relating pairs of windows from X^p and X^q in the form $S = \{(w_{1_p}^{(p;\lambda_{pq})}, w_{1_q}^{(q;\lambda_{qp})}), \cdots, (w_{n_p}^{(p;\lambda_{pq})}, w_{n_q}^{(q;\lambda_{qp})})\}$, should however be such that for all index i of the sequence one has: $s(w_{i_q}^{(q;\lambda_{qp})}) = s(w_{i_p}^{(p;\lambda_{pq})}) + \gamma_{pq}$, where $s(w)$ represents the starting time of the window w, and γ_{pq} is

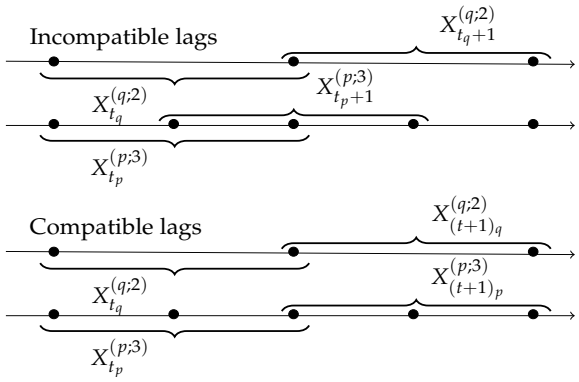


Figure 4.4 – Illustration for constructing sequences of windows for two time series with different sampling rates.

constant over time. This is not the case for the first example, but is true for the second one, which is a relevant sequence of observations.

If the two time series are sufficiently long, there always exists a correct sequence of joint observations. Indeed, if the window sizes λ_{pq} and λ_{qp} are known, let $\gamma_{pq} = s(w_1^{(q;\lambda_{qp})}) - s(w_1^{(p;\lambda_{pq})})$. Furthermore, let N_p and N_q denote the number of observations per time unit [2]. Then, λ_{pq}, λ_{qp} and γ_{pq} are compatible through the set of joint observations S with

$$s(w_{i_p}^{(p;\lambda_{pq})}) = s(w_1^{(p;\lambda_{pq})}) + (i_p - 1)LCM(N_p, N_q)$$

and

$$s(w_{i_q}^{(q;\lambda_{qp})}) = s(w_1^{(q;\lambda_{qp})}) + (i_q - 1)LCM(N_p, N_q),$$

with LCM the lowest common multiple.

4.3 PC based on causal temporal mutual information

We present in this section a new method for causal discovery in time series based on the causal temporal mutual information introduced above

2. Time unit corresponds to the largest (integer) time interval according to the sampling rates of the different time series. For example, if a time series has a sampling rate of 10 per second and another a sampling rate of 3 per 10 minutes, then the time unit is equal to 10 minutes.

to construct the skeleton of the causal graph. This skeleton is then oriented on the basis of the entropy reduction principle and the PC algorithm. Our method assumes both the causal Markov condition and faithfulness of the data distribution, which are classical assumptions for causal discovery within constraint-based methods.

4.3.1 Skeleton construction

We follow the same steps as the ones of the PC algorithm [Spirtes et al., 2000] which assumes that all variables are observed. It aims at building causal graphs by orienting a skeleton obtained, from a complete graph, by removing edges connecting independent variables. The summary causal graphs considered are directed acyclic graphs (DAG) in which self-loops are allowed to represent temporal dependencies within a time series.

Starting with a complete graph that relates all time series, the first step consists in computing CTMI for all pairs of time series and removing edges if the two time series are considered independent. Once this is done, one checks, for the remaining edges, whether the two time series are conditionally independent (the edge is removed) or not (the edge is kept). Starting from a single time series connected to X^p or X^q, the set of conditioning time series is gradually increased untill either the edge between X^p and X^q is removed or all time series connected to X^p and X^q have been considered. We will denote by Sepset(p, q) the separation set of X^p and X^q, which corresponds to the smallest set of time series connected to X^p and X^q such that X^p and X^q are conditionally independent given this set. Note that we make use of the same strategy as the one used in PC-stable [Colombo and Maathuis, 2014b], which consists in sorting time series according to their CTMI scores and, when an independence is detected, removing all other occurrences of the time series. This leads to an order-independent procedure.

The following theorem states that the skeleton obtained by the above procedure is the true one.

Theorem 4. *Let $\mathcal{G} = (V, E)$ be a summary causal graph, and assume that we are given perfect conditional independence information about all pairs of variables (X^p, X^q) in V given subsets $S \subseteq V \backslash \{X^p, X^q\}$. Then the skeleton previously constructed is the skeleton of $\mathcal{G}$.*

Proof. Let us consider two time series X^p and X^q. If they are independent given $X^{\mathbf{R}}$, then CTMI$(X^p; X^q | X^{\mathbf{R}}) = 0$ as otherwise the conditional mutual information between X^p and X^q would be non nul and the two

time series would not be conditionally independent as we are given perfect information. By the causal Markov and faithfulness conditions, there is no edge in this case between X^p and X^q in the corresponding skeleton, as in the true one. Conversely, if CTMI$(X^p; X^q|X^{\mathbf{R}}) = 0$ for any $X^{\mathbf{R}}$, then the two time series cannot be dependent conditioned on $X^{\mathbf{R}}$. Indeed, if this was the case, as we are given perfect conditional information, there would exist a lag γ and two window sizes λ_{pq} and λ_{pq} such that $I(X_t^{(p;\lambda_{pq})}; X_{t+\gamma}^{(q;\lambda_{qp})}|X^{\mathbf{R}}) > 0$ with $0 < \lambda_{pq}, \lambda_{qp} \leq \gamma$. In this case, the two windows of size λ_{max} centered on time point t in both X^p and X^q contain the windows of size λ_{pq} and λ_{pq} separated by a lag γ in X^p and X^q as $\lambda_{max} = 2\gamma_{max} + 1$. Thus, CTMI$(X^p; X^q|X^{\mathbf{R}})$ would be positive as this quantity cannot be less than $I(X_t^{(p;\lambda_{pq})}; X_{t+\gamma}^{(q;\lambda_{qp})}|X^{\mathbf{R}})$, which leads to a contradiction. Finally, as we test all necessary conditioning sets in the construction of the skeleton, we have the guarantee to remove all unnecessary edges. □

4.3.2 Orientation

Once the skeleton has been constructed, one tries to orient as many edges as possible using the standard PC-rules (see Section 3). As we are using here the standard PC rules, we have the following theorem, the proof of which directly derives from results on PC [Spirtes et al., 2000].

Theorem 5 (Theorem 5.1 of Spirtes et al. [2000])**.** *Let the distribution of V be faithful to a DAG $\mathcal{G} = (V, E)$, and assume that we are given perfect conditional independence information about all pairs of variables (X^p, X^q) in V given subsets $X^{\boldsymbol{R}} \subseteq V \backslash \{X^p, X^q\}$. Then the skeleton constructed previously followed by PC-Rules 0, 1, 2, and 3 represents the CPDAG of $\mathcal{G}$.*

In addition to the PC orientation rules, we introduce two new rules which are based on the notion of *possible spurious correlations* and the mutual information we have introduced. The notion of possible spurious correlations captures the fact that two variables may be correlated through relations that do not only correspond to direct causal relations between them. It is formalized as follows:

Definition 39 (Possible spurious correlations)**.** *We say that two nodes $X^p - X^q$ have possible spurious correlations if there exists a path between them that neither contains the edge $X^p - X^q$ nor any collider.*

Interestingly, when two connected variables do not have possible spurious correlations, one can conclude on their orientation using CTMI.

Property 2. *Let us assume that we are given perfect conditional independence information about all pairs of variables (X^p, X^q) in V given subsets $S \subseteq V \backslash \{X^p, X^q\}$. Then every non oriented edge in the CPDAG obtained by the above procedure corresponds to a* prima facie *cause and by, causal sufficiency, to a true causal relation between the related time series. Furthermore, the orientation of an unoriented edge between two nodes X^p and X^q that do not have possible spurious correlations is given by the "direction" of the optimal lag in $CTMI(X^p, X^q)$, assuming that the maximal window size is larger than the longest lag γ_{max} between causes and effects.*

Proof. The first part of Prop. 2 directly derives from Prop. 1. As we assume that we are given perfect conditional information, the skeleton is the true one from Theorem 4. Thus, if two variables do not have possible spurious correlations, the only correlations observed between them correspond to a causal relation. We now need to prove that the optimal lag can be used to orient edges between any pair of variables X^p and X^q.

Without loss of generality, let us assume that X^p causes X_t^q, for any time t, via the K time instants $\{t-\gamma, t-\gamma_1, \cdots, t-\gamma_{K-1}\}$ with $0 < \gamma_{K-1} < \cdots < \gamma_1 < \gamma$. First, let us consider a window of size 1 in X^q, and a window of arbitrary size λ in X^p with a lag γ_{pq} set to $\gamma' \geq 0$. As $\gamma' \geq 0$, there is no cause of X_t^q in the window $X_{t+\gamma'}^{(p;\lambda)}$. Furthermore, the only observed correlations between X^p and X^q correspond to causal relations. We thus have:

$$I(X_t^{(q;1)}; X_{t+\gamma'}^{(p;\lambda)} | X_{t-1}^{(q;1)}, X_{t+\gamma'-1}^{(p;1)}) = 0,$$

as X_t^q and all variables in $X_{t+\gamma'}^{(p;\lambda)}$ are independent of each other. One the contrary, for the same window size in X^p and a lag γ_{pq} set to $-\gamma$ with $\gamma > 0$, one has:

$$I(X_t^{(q;1)}; X_{t-\gamma}^{(p;\gamma)} | X_{t-1}^{(q;1)}, X_{t-\gamma-1}^{(p;1)}) \geq I(X_t^{(q;1)}; X_{t-\gamma}^{(p;1)} | X_{t-1}^{(q;1)}, X_{t-\gamma-1}^{(p;1)}) > 0.$$

The first inequality derives from Inequality 4.2. The second inequality is due to the fact that $X_{t-\gamma}^p$ is a true cause of X_t^q and the fact that we are given perfect information. Thus, when considering a window of size 1 for X^q, the optimal lag given by CTMI will necessarily go from X^p to X^q, which corresponds to the correct orientation.

We now consider the case where we have a window of arbitrary size λ' in X^q. Let us further consider a window of arbitrary size λ in X^p with a lag γ_{pq} set to $\gamma' \geq 0$. If $\lambda' < \gamma' + \gamma_{K-1}$, there is no causal relations between elements in $X_t^{(q;\lambda')}$ and elements in $X_{t+\gamma'}^{(p;\lambda)}$ and the mutual information between these two windows is 0. Otherwise, one can decompose this mutual

information as:

$$
\begin{aligned}
&I(X_t^{(q;\lambda')}; X_{t+\gamma'}^{(p;\lambda)} | X_{t-1}^{(q;1)}, X_{t+\gamma'-1}^{(p;1)}) \\
&= I(X_t^{(q;\gamma'+\gamma_{K-1})}; X_{t+\gamma'}^{(p;\lambda)} | X_{t-1}^{(q;1)}, X_{t+\gamma'-1}^{(p;1)}) \\
&\quad + I(X_{t+\gamma'+\gamma_{K-1}}^{(q;\lambda'-\gamma'-\gamma_{K-1})}; X_{t+\gamma'}^{(p;\lambda)} | X_{t+\gamma'+\gamma_{K-1}-1}^{(q;1)}, X_{t+\gamma'-1}^{(p;1)}),
\end{aligned}
$$

as the conditioning on $X_t^{(q;\gamma'+\gamma_{K-1})}$ and $X_{t-1}^{(q;1)}$ amounts to condition on the instant $X_{t+\gamma'+\gamma_{K-1}-1}^{(q;1)}$ due to the first-order Markov self-causal assumption.

As there are no causal relations between elements in $X_t^{(q;\gamma'+\gamma_{K-1})}$ and elements in $X_{t+\gamma'}^{(p;\lambda)}$ the first term in the right-hand side is 0. Using a similar decomposition in order to exclude elements at the end of $X_{t+\gamma'}^{(p;\lambda)}$ which do not cause any element in $X_t^{(q;\lambda')}$, one obtains:

$$
\begin{aligned}
&I(X_t^{(q;\lambda')}; X_{t+\gamma'}^{(p;\lambda)} | X_{t-1}^{(q;1)}, X_{t+\gamma'-1}^{(p;1)}) \\
&= I(X_{t+\gamma'+\gamma_{K-1}}^{(q;\lambda'-\gamma'-\gamma_{K-1})}; X_{t+\gamma'}^{(p;\min(\lambda,\lambda'-\gamma_{K-1}-\gamma'))} | X_{t+\gamma'+\gamma_{K-1}-1}^{(q;1)}, X_{t+\gamma'-1}^{(p;1)}).
\end{aligned}
$$

Let us now consider the window in X^p of size λ' with a lag γ_{pq} set to $-\gamma_{K-1}$. Using the same reasoning as before, one obtains:

$$
\begin{aligned}
&I(X_t^{(q;\lambda')}; X_{t-\gamma_{K-1}}^{(p;\lambda')} | X_{t-1}^{(q;1)}, X_{t-\gamma_{K-1}-1}^{(p;1)}) \\
&= I(X_t^{(q;\gamma'+\gamma_{K-1})}; X_{t-\gamma_{K-1}}^{(p;\lambda')} | X_{t-1}^{(q;1)}, X_{t-\gamma_{K-1}-1}^{(p;1)}) \\
&\quad + I(X_{t+\gamma'+\gamma_{K-1}}^{(q;\lambda'-\gamma'-\gamma_{K-1})}; X_{t-\gamma_{K-1}}^{(p;\lambda')} | X_{t+\gamma'+\gamma_{K-1}-1}^{(q;1)}, X_{t-\gamma_{K-1}-1}^{(p;1)}),
\end{aligned} \tag{4.5}
$$

with:

$$
\begin{aligned}
&I(X_{t+\gamma'+\gamma_{K-1}}^{(q;\lambda'-\gamma'-\gamma_{K-1})}; X_{t-\gamma_{K-1}}^{(p;\lambda')} | X_{t+\gamma'+\gamma_{K-1}-1}^{(q;1)}, X_{t-\gamma_{K-1}-1}^{(p;1)}) \\
&\geq I(X_{t+\gamma'+\gamma_{K-1}}^{(q;\lambda'-\gamma'-\gamma_{K-1})}; X_{t+\gamma'}^{(p;\min(\lambda,\lambda'-\gamma_{K-1}-\gamma'))} | X_{t+\gamma'+\gamma_{K-1}-1}^{(q;1)}, X_{t+\gamma'-1}^{(p;1)}),
\end{aligned}
$$

as the window $X_{t-\gamma_{K-1}}^{(p;\lambda')}$ contains the window $X_{t+\gamma'}^{(p;\min(\lambda,\lambda'-\gamma_{K-1}-\gamma'))}$. In addition, the first term in the right-hand side of Eq. 4.5 is strictly positive as all the elements in $X_t^{(q;\gamma'+\gamma_{K-1})}$ have causal relations in $X_{t-\gamma_{K-1}}^{(p;\lambda')}$. Thus, the mutual information obtained with the negative lag $-\gamma_{K-1}$ is better than the one obtained with any positive lag,

$$
I(X_t^{(q;\lambda')}; X_{t-\gamma_{K-1}}^{(p;\lambda')} | X_{t-1}^{(q;1)}, X_{t-\gamma_{K-1}-1}^{(p;1)}) > I(X_t^{(q;\lambda')}; X_{t+\gamma'}^{(p;\lambda)} | X_{t-1}^{(q;1)}, X_{t+\gamma'-1}^{(p;1)});
$$

meaning that the optimal lag given by CTMI will necessarily go from X^p to X^q, which corresponds to the correct orientation. □

The following orientation rule is a direct application of the above property.

ER-Rule 0 (Entropy Reduction - γ)**.** *In a pair $X^p - X^q$, such X^p and X^q do not have a possible spurious correlations, if $\bar{\gamma}_{pq} > 0$, then orient the edge as: $X^p \to X^q$.*

Furthermore, we make use of the following rule to orient additional edges when the optimal lag $\bar{\gamma}_{pq}$ is null based on the fact that CTMI increases asymmetrically with respect to the increase of λ_{pq} and λ_{qp} (Figure 4.2). This rule infers the direction of the cause by checking the difference in the window sizes as the window size of the cause cannot be greater than the window size of the effect.

ER-Rule 1 (Entropy Reduction - λ)**.** *In a pair $X^p - X^q$, such X^p and X^q do not have a possible spurious correlations, if $\bar{\gamma}_{pq} = 0$ and $\bar{\lambda}_{pq} < \bar{\lambda}_{qp}$ then orient the edge as: $X^p \to X^q$.*

Algorithm 1 represents the pseudo-code of `PCTMI`. $\text{Adj}(X^q, \mathcal{G})$ represents all adjacent nodes to X^q in the graph $\mathcal{G}$ and $\texttt{sepset}(p, q)$ is the separation set of X^p and X^q.

Finally, given the graph $\mathcal{G}$ inferred with the above procedure, one can verify for each node X^q in $\mathcal{G}$ if it is self causal by checking if for all t, $CTMI(X_t^q; X_{t-1}^q \mid \text{Par}(X_t^q))$ in $\mathcal{G}$.

In practice, we also apply ER-Rule 0 before PC-Rules, because experimentally we found that ER-Rule 0 is more reliable than PC-Rule 0 in detecting lagged unshielded colliders, especially in the case of low sample size.

4.4 FCI based on causal temporal mutual information

When unobserved variables are causing a variable of interest, the PC algorithm is no longer appropriate and one needs to resort to the FCI algorithm introduced in Spirtes et al. [2000] which infers a PAG (partial ancestral graph). We extend here the version of this algorithm presented in Zhang [2008a] and described in Section 3.3, without the selection bias.

From the skeleton obtained in Section 4.3.1, unshielded colliders are detected using the FCI-Rule 0 in Section 3.3. From this, Possible-Dsep sets

Algorithm 1 PCTMI

Require: X a d-dimensional time series of length T, $\gamma_{\max} \in \mathbb{N}$ the maximum number of lags, α a significance threshold
  Form a complete undirected graph $\mathcal{G} = (V, E)$ with d nodes
  n = 0
  while there exists $X^q \in V$ such that $\text{card}(\text{Adj}(X^q, \mathcal{G})) \geq n + 1$ **do**
    $\mathbf{D} = list()$
    for $X^q \in V$ s.t. $\text{card}(\text{Adj}(X^q, \mathcal{G})) \geq n + 1$ **do**
      for $X^p \in \text{Adj}(X^q, \mathcal{G})$ **do**
        for all subsets $X^{\mathbf{R}} \subset \text{Adj}(X^q, \mathcal{G}) \setminus \{X^p\}$ such that $\text{card}(X^{(\mathbf{R})}) = n$ and ($\Gamma_{rp} \geq 0$ or $\Gamma_{rq} \geq 0$) for all $r \in \mathbf{R}$ **do**
          $y_{q,p,\mathbf{R}} = \text{CTMI}(X^p; X^q \mid X^{\mathbf{R}})$
          $\text{append}(\mathbf{D}, \{X^q, X^p, X^{\mathbf{R}}, y_{q,p,\mathbf{R}}\}))$
        end for
      end for
    end for
    Sort **D** by increasing order of y
    while D is not empty **do**
      $\{X^q, X^p, X^{\mathbf{R}}, y\} = \text{pop}(\mathbf{D})$
      if $X^p \in \text{Adj}(X^q, \mathcal{G})$ and $X^{\mathbf{R}} \subset \text{Adj}(X^q, \mathcal{G})$ **then**
        Compute z the p-value of $\text{CTMI}(X^p; X^q \mid X^{\mathbf{R}})$ given by Eq. (**??**)
        if test $z > \alpha$ **then**
          Remove edge $X^p - X^q$ from $\mathcal{G}$
          $\text{Sepset}(p, q) = \text{Sepset}(q, p) = X^{\mathbf{R}}$
        end if
      end if
    end while
    n=n+1
  end while
  for each triple in $\mathcal{G}$ **do** apply PC-Rule 0
  while no more edges can be oriented **do**
    for each triple in $\mathcal{G}$ **do** apply PC-Rules 1, 2, 3
  end while
  for each connected pair in $\mathcal{G}$ **do** apply ER-Rules 0, 1
  Return the summary causal graph $\mathcal{G}$

can be constructed. As elements of Possible-Dsep sets in a PAG play a role similar to the ones of parents in a DAG, additional edges are removed by conditioning on the elements of the Possible-Dsep sets, using the same

strategy as the one given in Section 4.3.1. All edges are then unoriented and the FCI-Rule 0 is again applied as some of the edges of the unshielded colliders originally detected may have been removed by the previous step. Then, as in FCI, FCI-Rules 1, 2, 3, 4, 8 9 and 10 are applied. Note that we have not included FCI-Rules 5, 6 and 7 from Zhang [2008a] as these rules deal with selection bias, a phenomenon that is not present in the datasets we consider. Including these rules in our framework is nevertheless straightforward. Finally as in `PCTMI`, we orient additional edges using the ER-Rules.

The overall process, referred to as `FCITMI`, is described in Algorithm 2.

4.5 Extension to window causal graph

In this chapter, we presented our method for the discovery of a summary causal graph. While in many applications the knowledge of summary causal graphs is sufficient, in some particular cases, one may need window causal graphs. Here, we present a procedure, that allows inferring a window causal graph given a summary causal graph. This method can be decomposed into two steps: time-adaptation and temporal skeleton separation. First, information from the summary causal graph is transferred into a window-based graph, which is done by taking all inferred relations in the previous steps and representing them in a time-fashioned graph. The second step captures all confounders of all pairs of all points of time series.

We use the same set of assumptions as in `PCTMI` and assume that the summary causal graph $\mathcal{G}$ is given (obtained using `PCTMI`). We start with constructing a window causal graph $\mathcal{TG}$, where the window have size of $\gamma_{max} + 1$. For each oriented edge $X^p \to X^q$ in the summary causal graph G, we add $X_t^p \to X_{t+k}^q$ in $\mathcal{TG}$ for all the time points k in the defined window. For the non-oriented edges $X^p - X^q$ in G, we add all possible corresponding edges in $\mathcal{TG}$, i.e., we add both $X_t^p \to X_{t+k}^q$ and $X_t^q \to X_{t+k}^p$ in $\mathcal{TG}$ for all the time points k. This process is illustrated in Figure 4.5, with simple summary causal graph $\mathcal{G}$ (a) and derived window causal graph $\mathcal{TG}$ (b). The causal relation $X^r \to X^p$ in $\mathcal{G}$ (Figure 4.5 (a)) is transferred into directed edges from time point X_{t-2}^r to time points $(X_{t-2}^p, X_{t-1}^p, X_t^p)$ and from X_{t-1}^r to time points (X_{t-1}^p, X_t^p) (Figure 4.5 (b)). The non-oriented edge between X^r and X^q in $\mathcal{G}$ is transferred into directed edges from all lagged relations from X^r to X^q and from X^q to X^r, and into undirected edges from the remained instantaneous relations X_{t-2}^r and X_{t-2}^q, X_{t-1}^r and

Figure 4.5 – Time adaptation result for two nodes related by a counfounder for $\gamma_{max} = 2$.

X^q_{t-1}, and X^r_t and X^q_t.

Then, similarly to PC algorithm, non causal relations are removed from graph $\mathcal{TG}$ through conditional independence tests. In particular, for each pair of nodes (X^q_t, X^p_{t+k}), we compute their mutual information $I(X^q_t, X^p_{t+k})$ and test through a permutation test if they are independent. In case of independence we remove the edge between them. Then we check if two nodes (X^q_t, X^p_{t+k}) are conditionally independent by conditioning on all possible sets X^R_T of size 1, which are connected to X^q_t or X^p_{t+k}. We calculate conditional mutual information $I(X^q_t, X^p_{t+1} \mid X^R_T)$ for evaluating conditional independence of nodes (X^q_t, X^p_{t+k}) and remove an edge in case of conditional independence. We repeat this procedure iteratively increasing the size of set $X^{\mathbf{R}}_{\mathbf{T}}$ until edge between (X^q_t, X^p_{t+k}) is removed or all conditioning sets are considered. We use the same procedure as in `PCTMI` algorithm for ranking conditional independence, thus this method is also order-independent.

4.6 Extension to sequences

A time series is a sequence observed at successive equally spaced points in time but sequential data, however, is any kind of data where the order matters and is not necessarily temporal (the time stamp is irrelevant), for example, DNA sequence, text, and trajectories. In this section, we claim that sometimes reducing a time series to a sequence can be beneficial especially in the case of temporal misalignment.

Many real-world time series data sets are subject to temporal misalignment, i.e., the time periods defining the data points are not the same across

different or the same series. Temporal misalignment becomes a problem when multiple irregularly spaced time series are considered together especially for causal discovery algorithms who rely on time to infer the direction of the cause. This phenomenon is illustrated in Figure 4.6 which shows two time series X^p and X^q such that X^p cause X^q with a lag of one. In Figure 4.6b we can see the corresponding data point of the dynamic system is faithful to the real time order. Given these data points, a causal discovery algorithm for time series, namely PCTMI would infer the correct causal relation with $\gamma_{p,q} = 1$. Now suppose that the same time series at the moment of the data collect which produced a misalignment: X^p is shifted back by 1 point and X^q is shifted forward by 1 point as plotted in Figure 4.6c. In this case, PCTMI (and probably every causal discovery algorithm suited for time series) would fail.

Here we provide a simple modification of PCTMI to encounter the problem of misaligned time series, and we conjuncture that this modification can also be used to other types of sequences. First, in the estimation of conditional CTMI, one should drop the conditions on Γ. Formally speaking, in case of misalignment between time series, the conditional causal temporal mutual information between two time series X^p and X^q conditioned on a set $X^{\mathbf{R}} = \{X^{r_1}, \cdots, X^{(r_K)}\}$ is given by:

$$\begin{aligned} &\text{CTMI}(X^p; X^q \mid X^{\mathbf{R}}) \\ =&I(X_t^{(p;\bar{\lambda}_{pq})}; X_{t+\bar{\gamma}_{pq}}^{(q;\bar{\lambda}_{qp})} | (X_{t-\bar{\Gamma}_k}^{(r_k;\bar{\lambda}_k)})_{1\leq k\leq K}, X_{t-1}^{(p;1)}, X_{t+\bar{\gamma}_{pq}-1}^{(q;1)}), \end{aligned} \tag{4.6}$$

where $(\bar{\Gamma}_1, \ldots, \bar{\Gamma}_K)$ and $(\bar{\lambda}_1, \cdots, \bar{\lambda}_K)$ correspond to the optimal conditional lags and window sizes which minimize, for $\Gamma_1, \ldots, \Gamma_K \in \mathbb{Z}$:

$$I\left(X_t^{(p;\bar{\lambda}_{pq})}; X_{t+\bar{\gamma}_{pq}}^{(q;\bar{\lambda}_{qp})} | (X_{t-\Gamma_k}^{(r_k;\lambda_k)})_{1\leq k\leq K}, X_{t-1}^{(p;1)}, X_{t+\bar{\gamma}_{pq}-1}^{(q;1)}\right).$$

Moreover, one should avoid using the ER-Rules, since their main purpose is to find the direction of causation through the direction of associations with respect to time which is now corrupted due to the misalignment. Finally, in the third for loop of the algorithm, while searching for the separation sets, the conditions on γ should be dropped, namely, the condition: $\gamma_{rp} \geq 0$ or $\gamma_{rq} \geq 0$.

Interpretation: If a graph $\mathcal{G}$ is inferred using the adapted version of PCTMI for misaligned time series, and if there exists a causal relation $X^p \to X^q$ in $\mathcal{G}$ such that $\gamma_{pq} < 0$ then one can conclude that time series X^p and time series X^q are misaligned in a way that violates the temporal priority assumption (Definition 8).

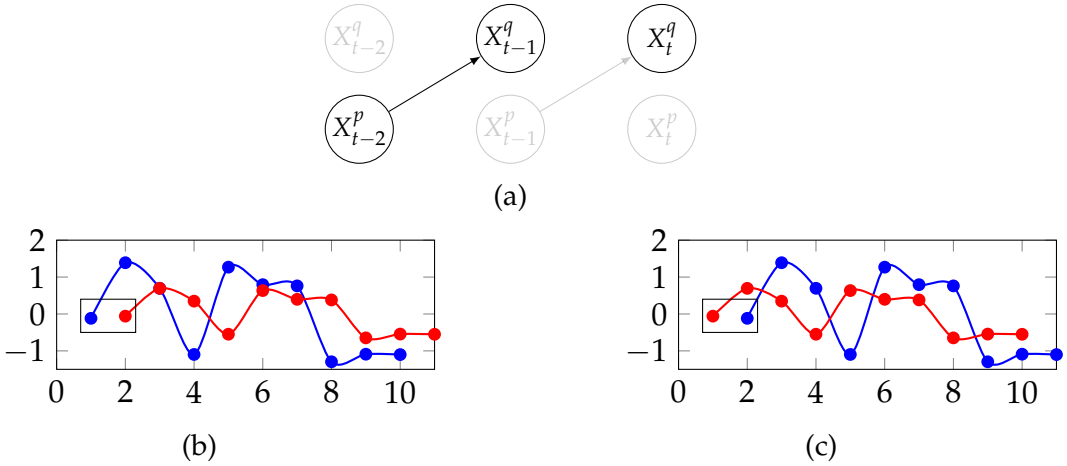


Figure 4.6 – Misaligned time series. X^p_t is sampled from a normal distribution and X^q_{t+1} is X^p_t divided by 2.

4.7 Conclusion

We have addressed in this chapter the problem of learning a summary causal graph on time series with equal or different sampling rates. To do so, we have first proposed a new temporal mutual information measure defined on a window-based representation of time series. We have then shown how this measure relates to an entropy reduction principle that can be seen as a special case of the probabilistic raising principle. We have finally combined these two ingredients in a PC-like algorithm to construct the summary causal graph. Then we extended the main algorithm to handle hidden common causes analogous to the FCI algorithm. Finally, we showed how can a window causal graph be inferred given a summary causal graph and we provided an intuitive adaptation of our main algorithm to sequences. The main limitation of our method is that it is restricted to the Markov equivalence class in case of instantaneous relations. In the next chapter, we will see how we can overcome this issue.

Algorithm 2 `FCITMI`

Require: X a d-dimensional time series of length T, $\gamma_{\max} \in \mathbb{N}$ the maximum number of lags, α a significance threshold
  Form a complete undirected graph $\mathcal{G} = (V, E)$ with d nodes
  n = 0
  while there exists $X^q \in V$ such that $\text{card}(\text{Adj}(X^q, \mathcal{G})) \geq n + 1$ **do**
    $\mathbf{D} = list()$
    for $X^q \in V$ s.t. $\text{card}(\text{Adj}(X^q, \mathcal{G})) \geq n + 1$ **do**
      for $X^p \in \text{Adj}(X^q, \mathcal{G})$ **do**
        for all subsets $X^{\mathbf{R}} \subset \text{Adj}(X^q, \mathcal{G}) \setminus \{X^p\}$ such that $\text{card}(X^{(\mathbf{R})}) = n$ and ($\gamma_{rp} \geq 0$ or $\gamma_{rq} \geq 0$) for all $r \in \mathbf{R}$ **do**
          $y_{q,p,\mathbf{R}} = \text{CTMI}(X^p; X^q \mid X^{\mathbf{R}})$
          $\text{append}(\mathbf{D}, \{X^q, X^p, X^{\mathbf{R}}, y_{q,p,\mathbf{R}}\}))$
        end for
      end for
    end for
    Sort **D** by increasing order of y
    while D is not empty **do**
      $\{X^q, X^p, X^{\mathbf{R}}, y\} = \text{pop}(\mathbf{D})$
      if $X^p \in \text{Adj}(X^q, \mathcal{G})$ and $X^{\mathbf{R}} \subset \text{Adj}(X^q, \mathcal{G})$ **then**
        Compute z the p-value of $\text{CTMI}(X^p; X^q \mid X^{\mathbf{R}})$ given by Eq. (**??**)
        if test $z > \alpha$ **then**
          Remove edge $X^p - X^q$ from $\mathcal{G}$
          $\text{Sepset}(p, q) = \text{Sepset}(q, p) = X^{\mathbf{R}}$
        end if
      end if
    end while
    n=n+1
  end while
  for each triple in $\mathcal{G}$ **do** apply FCI-Rule 0
  using Possible-Dsep sets, remove edges using CTMI
  Reorient all edges as $\circ\!\!-\!\!\circ$ in $\mathcal{G}$
  for each triple in $\mathcal{G}$ **do** apply FCI-Rule 0
  while edges can be oriented **do**
    for each triple in $\mathcal{G}$ apply FCI-Rules 1,2, 3, 4, 7, 9, 10
    for each connected pair in $\mathcal{G}$ **do** apply ER-Rules 0, 1.
  end while
  Return the summary causal graph $\mathcal{G}$

Chapter 5

A mixed noise and entropy based approach to causal inference in time series

> Not only does noise not obscure causal relations, it is an invaluable source of insight regarding them.
>
> Nevin Climenhaga, Lane DesAutels, and Grant Ramsey

5.1 Introduction

In the previous chapter, we introduced a new algorithm, called PCTMI, which is entirely affiliated to the constraint-based family of methods. Like other constraint-based methods, PCTMI relies on conditional independencies and assume *faithfulness*, which states that the joint distribution P over V is faithful to the true causal Directed Acyclic Graph (DAG) $\mathcal{G}$ over V in the sense that every conditional independence statement satisfied by P is entailed by $\mathcal{G}$ [Spirtes et al., 2000]. Moreover, in general, using such approaches, graphs can only be recovered up to Markov equivalence [1] classes. However, since PCTMI uses the notion of time, it can go beyond the Markov equivalence class only for lagged relations [Runge et al., 2019], *i.e.* instantaneous relations are always limited to the Markov equivalence class. Alternatively, noise-based approaches present full indentifiabilty of the causal graphs if all assumptions are met. These methods usually assume causal Markov condition and the minimality condition, which is a weaker assumption than faithfulness, in addition to a restriction to the model class (for more details, see Section 3.4). The main drawbacks are that such methods usually do not scale well [Glymour et al., 2019] and might need a large sample size [Malinsky and Danks, 2018], and in practice, they do not achieve good performance compared to constraint based approaches. To overcome, the limitation of PCTMI that is mentioned above, we propose, a hybrid method, that takes benefit of the two approaches: noise based and constraint based. Thus it is not limited to a Markov equivalence class and provides a specific graph, scales better and needs a smaller sample size.

Our contribution is two-fold. We first use a well known noise-based procedure to infer a causal ordering between the time series which can be interpreted as an oriented graph which contain the true graph and which reduce the space of search compared to a fully non oriented connected graph. Then we introduce a new measure of dependence between two time series called the temporal causation entropy, which is an extension of the standard causation entropy measure [Sun et al., 2015] to time series to handle instantaneous relations and lags bigger than one and it is a asymmetric version of the temporal mutual information introduced in Chapter 4. This new measure is used in a PC-like algorithm to prune unnecessary arrows by looking at possible confounders and therefore end-up with only genuine cause. Remarkably, this is to our knowledge the first

1. Two DAGs are Markov equivalent if and only if they have the same skeleton and the same v-structures [Verma and Pearl, 1991].

algorithm hybrid between constraint-based and noise-based methods for time series.

The remainder of the chapter is organized as follows: Section 5.2 introduces the main causal discovery algorithm, called NBCB. It relies on weak assumptions that are reminded first, and on the temporal causation entropy that is also introduced. Finally, Section 5.4 concludes the chapter.

5.2 Weakening faithfulness and going beyond the Markov equivalence class

The faithfulness assumption is difficult to check in practice, and it has been debated for a long time. It assumes that there are no accidental conditional independence relations in the true distribution, that is, no conditional independence relations unless entailed by the true causal structure. The faithfulness assumption is mainly used in constraint-based methods, where it is used at two different stages (skeleton construction and orientation phase), so it can be decomposed into two assumptions as proposed in Ramsey et al. [2006], namely Adjacency faithfulness (Definition 40) and Orientation faithfulness (Definition 41).

In PC-like algorithm, the skeleton is inferred in the first step, by looking at the adjacency of each pair (X^p, X^q) considering conditional independences. This step relies on the adjacency faithfulness assumption which is defined as follow:

Definition 40 (Adjacency faithfulness [Ramsey et al., 2006])**.** *For every $X^p, X^q \in V$, if X^p and X^q are adjacent in $\mathcal{G}$, then they are not conditionally independent given any subset of $V \backslash \{X^p, X^q\}$.*

In the second step, PC-like algorithm finds all unshielded colliders by considering that every unshielded triple $X^p - X^r - X^q$ is a collider if and only if X^r is not in the separation set of X^p and X^q. This step relies on the Orientation faithfulness assumption which is defined as follow:

Definition 41 (Orientation faithfulness [Ramsey et al., 2006])**.** *For every $X^p, X^q, X^r \in V$ such that $X^p - X^r - X^q$ is an unshielded triple in $\mathcal{G}$:*

- *If the triple X^p, X^r, X^q is a collider, i.e., $X^p \to X^r \leftarrow X^q$ in $\mathcal{G}$, then X^p and X^q are not conditionally independent given any subset of $V \backslash \{X^p, X^q\}$ that includes X^r.*
- *Otherwise, X^p and X^q are not conditionally independent given any subset of $V \backslash \{X^p, X^q\}$ that excludes X^r.*

We follow Ramsey et al. [2006], who relax the standard faithfulness assumption and still have provably correct and informative causal graph discovery procedures. The approach we propose discover causal relations from time series under the causal Markov condition, the causal minimality condition (needed in the causal ordering, when using the noise-based method) and adjacency faithfulness (needed in the pruning step, when using the constraint-based method). We also assume that the summary causal graph is acyclic. Our approach is a hybrid based method which is decomposed into two parts. The first part, a noise-based approach, is described in Algorithm 3. It is similar to Peters et al. [2013], but we extend the theoretical framework. A Gaussian process maps the past of the time series to the present, and a dependency measure between its input and its residuals is used to infer which time series potentially causes the other. The second part, a constraint-based approach, is described in Algorithm 4: considering the set of potential parents, the graph is pruned to remove spurious causes. The two parts are detailed below.

5.2.1 Causal ordering through noise

The first step relies on noise-based approaches, which were initially introduced for i.i.d. data. However, they gained much attention in recent years [Hoyer et al., 2009, Mooij et al., 2009, 2016, Assaad et al., 2019], and have also been extended for time series [Peters et al., 2013].

In this chapter, we focus on Additive Noise Models (ANMs) as defined in the following:

$$X_t^q = f^q(\mathrm{Par}(X_t^q)_{t-\gamma_{max}}, \ldots, \mathrm{Par}(X_t^q)_{t-1}, \mathrm{Par}(X_t^q)_t, \xi_t^q), \tag{5.1}$$

where f^q is a potentially nonlinear function, $\mathrm{Par}(X_0^q) \subseteq X^{V\setminus q}, \mathrm{Par}(X_k^q) \subseteq X$, $(\xi_t^q)_{q,t}$ are jointly independent, for each q, ξ_t^q are identically distributed in t and the finite dimensional distributions for the time series $(X^q)_{1\leq q\leq d}$ are absolutely continuous w.r.t a product measure. Remark that this model allows instantaneous relations. ANMs are identifiable, as they belong to the identifiable functional model class (IFMOC) [Peters et al., 2011], even in case of non-faithful causal models, for which conditional independence-based methods, as constraint-based, usually fail [Peters et al., 2011].

First we focus on the causal ordering. Similarly to the bivariate case [Hoyer et al., 2009, Mooij et al., 2016], independence between signal and residuals allow to detect the most potential cause from a set of variables through the following principle.

Algorithm 3 NBCB Part I: noise-based approach to order causes

Require: X a d-dimensional time series, $\gamma_{\max} \in \mathbb{N}$ the maximum number of lags
  $\mathcal{G}$ an empty graph with nodes $\{X^1, \dots, X^d\}$
  $S = \{1, \dots, d\}$
  while length$(S) > 1$ **do**
    for $q \in S$ **do**
      Learn $\hat{f}^q : \{(X_{t-\gamma_{max}}^{(p;\gamma_{max}+1)})_{p\in S, p\neq q}, (X_{t-\gamma_{max}}^{(q;\gamma_{max})})\} \mapsto X_t^q$
      Deduce $\hat{\xi}_t^q$ and compute c_q from Eq. (5.2)
    end for
    Choose $q^* = \text{argmin}\, c_j$
    $S = S \backslash q^*$
    for $s \in S$ **do**
      Add $X^s \to X^{q^*}$ in $\mathcal{G}$
    end for
  end while
  Return $\mathcal{G}$

Principle 2 (Multivariate additive noise principle)**.** *Suppose we are given a joint distribution $P(X^1, \cdots, X^d)$. If it satisfies an identifiable Additive Noise Model defined in* (5.1) *such that $\{(X_{t-\gamma_{max}}^{(p;\gamma_{max}+1)})_{1\leq p\neq q\leq d}, (X_{t-\gamma_{max}}^{(q;\gamma_{max})})\} \to X_t^q$, then it is likely that $\{(X_{t-\gamma_{max}}^{(p;\gamma_{max}+1)})_{1\leq p\neq q\leq d}, (X_{t-\gamma_{max}}^{(q;\gamma_{max})})\}$ precedes X_t^q in the causal order.*

Similarly to Mooij et al. [2016], when considering a suitable regression estimator and a suitable dependency estimator, the true causal order will be inferred. If we consider the fully connected graph given by this causal ordering (an edge between each node and its parents), it yields to a graph that contains the real graph: all true causal relations are in the inferred graph.

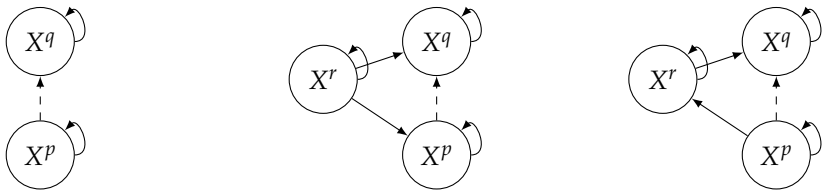
In practice, we first estimate for all $q \in \{1, \dots, d\}$,

$$\hat{f}^q : \{(X_{t-\gamma_{max}}^{(p;\gamma_{max}+1)})_{1\leq p\neq q\leq d}, (X_{t-\gamma_{max}}^{(q;\gamma_{max})})\} \mapsto X_t^q$$

by a Gaussian Process and deduce the residuals

$$\hat{\xi}_t^q = X_t^q - \hat{f}^q\{(X_{t-\gamma_{max}}^{(p;\gamma_{max}+1)})_{1\leq p\neq q\leq d}, (X_{t-\gamma_{max}}^{(q;\gamma_{max})})\}.$$

The last place in the causal ordering (which belongs to the most probable effect of all other time series) is given to the time series which yields the



(a) (b) (c)

Figure 5.1 – Wrong causal relations potentially inferred in the first step of our algorithm. Dashed lines represents wrong causal relations. On the left, we show a spurious cause, whereas on the middle and on the right, we provide two indirect causes.

residuals that are more independent to the other time series. The dependency between the residuals and the input is estimated with

$$c_q = C\left((X_{t-\gamma_{max}}^{(p;\gamma_{max}+1)})_{1\leq p\neq q\leq d}, (X_{t-\gamma_{max}}^{(q;\gamma_{max})}), \hat{\xi}_t^q\right), \tag{5.2}$$

where C is a dependence measure [2]. We then compare values of dependencies (or associated p-value if one wants to use statistical test).

However, this method is not capable of detecting independence between two time series, and thus it is susceptible to treat indirect causes as direct causes. To remove indirect causes or detect independencies, we complement this procedure with a second step that prunes spurious relations from the graph. It necessitates an exact estimation of the lag between two time series whereas for now we were content in using all possible lags.

5.2.2 Pruning using temporal causation entropy

Knowing the list of potential parents of each time series, as detected in the previous step, one way to prune the causes that are not genuine is to conduct conditional independence tests between time series. Indeed, suppose X^p is a potential cause of X^q but X^p and X^q are conditionally independent, then we can conclude that X^p is not a cause of X^q. This is illustrated in Figure 5.1.

In order to capture the dependencies (and conditional dependencies) between two time series, one needs to take into account the lag between them, as the true causal relations might be not instantaneous. Several studies have acknowledged the importance of taking into account lags to measure (conditional) dependencies between time series [Granger, 2004, Sun

2. As motivated in Peters et al. [2013], we use the partial correlation to measure the dependence, but one can generalize our procedure with any measure.

et al., 2015]. Causation entropy, introduced in Sun et al. [2015], is an asymmetric measure that detects the uncertainty reduction of the future states of X^q as a result of knowing the past states of X^p given that the past of $X^{\mathbf{R}}$ is already known, where $\mathbf{R}$ is a subset of $\{1, \cdots, d\}$. However, it only considers causation with lag of size one, whereas it can take any values in practice.

In addition to lags, a window-based representation may be necessary to fully capture the dependencies between the two time series. So it may be convenient to consider them together when assessing whether the time series are dependent or not. We thus introduce the temporal causation entropy, that extends the causation entropy to general lags and window representation of time series.

Definition 42 (Temporal causation entropy)**.** *We first define the optimal lag $\bar{\gamma}_{pq}$ between time series X^p and X^q and ($\bar{\lambda}_{pq}$, $\bar{\lambda}_{qp}$) the optimal windows of time series X^p regarding X^q and of time series X^q regarding X^p respectively as:*

$$\bar{\gamma}_{pq}, \bar{\lambda}_{pq}, \bar{\lambda}_{qp} = \underset{\gamma_{pq} \geq 0, \lambda_{pq}, \lambda_{qp}}{\operatorname{argmax}} \; h(X_t^{(q;\lambda_{qp})} \mid X_{t-\gamma_{pq}-1}^{(p;1)}, X_{t-1}^{(q;1)}) - h(X_t^{(q;\lambda_{qp})} \mid X_{t-\gamma_{pq}-1}^{(p;\lambda_{pq}+1)}, X_{t-1}^{(q;1)}),$$

where h denotes the entropy. The temporal causation entropy *from time series X^p to time series X^q conditioned on a set $X^{\boldsymbol{R}} = \{X^{r_1}, \cdots, X^{r_K}\}$ is given by:*

$$\begin{aligned} TCE(X^p \to X^q \mid X^{\boldsymbol{R}}) = & \min_{\Gamma_{r_i} \geq 0, 1 \leq i \leq K} h(X_t^{(q;\bar{\lambda}_{qp})} \mid (X_{t-\Gamma_{r_i}}^{r_i})_{1 \leq i \leq K}, X_{t-\bar{\gamma}_{pq}-1}^{(p;1)}, X_{t-1}^{(q;1)}) \\ & - h(X_t^{(q;\bar{\lambda}_{qp})} \mid (X_{t-\Gamma_{r_i}}^{r_i})_{1 \leq i \leq K}, X_{t-\bar{\gamma}_{pq}-1}^{(p;\bar{\lambda}_{pq}+1)}, X_{t-1}^{(q;1)}) \\ \triangleq & \; h(X_t^{(q;\bar{\lambda}_{qp})} \mid (X_{t-\bar{\Gamma}_{r_i}}^{r_i})_{1 \leq i \leq K}, X_{t-\bar{\gamma}_{pq}-1}^{(p;1)}, X_{t-1}^{(q;1)}) \\ & - h(X_t^{(q;\bar{\lambda}_{qp})} \mid (X_{t-\bar{\Gamma}_{r_i}}^{r_i})_{1 \leq i \leq K}, X_{t-\bar{\gamma}_{pq}-1}^{(p;\bar{\lambda}_{pq}+1)}, X_{t-1}^{(q;1)}), \end{aligned}$$

where $(\bar{\Gamma}_{r_1}, \cdots, \bar{\Gamma}_{r_K})$ are the optimal lags between $X^{\boldsymbol{R}}$ and X^q.

First, the lag between X^p and X^q is detected by maximizing the dependency between X^p and X^q. As we measure the amount of information brought by the observations of one variable on the observations of another variable, taking the maximum ensures that one does not miss any possible information contributing to relating the two time series. In a second step, we find the lags between (X^p, X^q) and $X^{\mathbf{R}}$ that minimize the conditional dependency between X^p and X^q conditioned on $X^{\mathbf{R}}$. Taking the minimum ensures that we search for the lags that break the maximal dependence.

Algorithm 4 NBCB part II: constraint-based approach for pruning

Require: X a d-dimensional time series, $\gamma_{\max} \in \mathbb{N}$ the maximum number of lags, α a significance threshold, $\mathcal{G}$ a causal graph
n = 0
while there exists $X^q \in V$ such that $\text{card}(\text{Par}(X^q, \mathcal{G})) \geq n + 1$ **do**
  $\mathbf{D} = list()$
  for $X^q \in V$ such that $\text{card}(\text{Par}(X^q, \mathcal{G})) \geq n + 1$ **do**
    for $X^p \in \text{Par}(X^q, \mathcal{G})$, $X^{\mathbf{R}} \subset \text{Par}(X^q, \mathcal{G}) \setminus \{X^p\}$ with $\text{card}(X^{\mathbf{R}}) = n$ **do**
      $y_{q,p,\mathbf{R}} = \text{TCE}(X^p; X^q \mid X^{\mathbf{R}})$
      $\text{append}(\mathbf{D}, \{X^q, X^p, X^{\mathbf{R}}\}))$
    end for
  end for
  Sort **D** by increasing order of y
  while D is not empty **do**
    $\{X^q, X^p, X^{\mathbf{R}}\} = \text{pop}(\mathbf{D})$
    if $X^p \in \text{Par}(X^q, \mathcal{G})$ and $X^{\mathbf{R}} \subset \text{Par}(X^q, \mathcal{G})$ **then**
      Compute z the p-value of $\text{TCE}(X^p; X^q \mid X^{\mathbf{R}})$ given by Eq. (5.3)
      if $z > \alpha$ **then**
        Remove edge $X^p \to X^q$ from $\mathcal{G}$
      end if
    end if
  end while
  n=n+1
end while
Return $\mathcal{G}$

Following the temporal priority principle, which states that causes precede their effects in time, we also ensure while finding only nonnegative lags that X^p as well as the conditional variables should precede in time X^q. If $\bar{\gamma}_{pq} = 1$ and $\bar{\lambda}_{pq} = \bar{\lambda}_{qp} = 1$, then the temporal causation entropy is equivalent to causation entropy.

As for CTMI (Section 4.2 of Chapter 4), to estimate TCE, we rely on the knn estimator introduced in Frenzel and Pompe [2007]. We denote by $\epsilon_{ik}/2$ the distance from

$$(X_{t-\bar{\gamma}_{pq}-1}^{(p;\bar{\lambda}_{pq}+1)}, X_{t-1}^{(q;\bar{\lambda}_{qp}+1)}, (X_{t-\bar{\Gamma}_{r_i}}^{r_i})_{1\leq i\leq K})$$

to its k-th neighbor, $n_i^{1,3}$, $n_i^{2,3}$ and n_i^3 the numbers of points with distance

strictly smaller than $\epsilon_{ik}/2$ in the subspace

$$(X_{t-\bar{\gamma}_{pq}-1}^{(p;\bar{\lambda}_{pq}+1)}, X_{t-1}^{(q;1)}, (X_{t-\bar{\Gamma}_{r_i}}^{r_i})_{1\leq i\leq K}),$$

$$(X_{t-\bar{\gamma}_{pq}-1}^{(p;1)}, X_{t-1}^{(q;\bar{\lambda}_{qp}+1)}, (X_{t-\bar{\Gamma}_{r_i}}^{r_i})_{1\leq i\leq K})$$

and

$$(X_{t-\bar{\gamma}_{pq}-1}^{(p;1)}, X_{t-1}^{(q;1)}, (X_{t-\bar{\Gamma}_{r_i}}^{r_i})_{1\leq i\leq K})$$

respectively, and $n_{\gamma_{rp},\gamma_{rq}}$ the number of observations. The estimate of the temporal causation entropy is then given by:

$$\widehat{TCE}(X^p \to X^q \mid X^{\mathbf{R}}) = \psi(k) + \frac{1}{n_{\gamma_{rp},\gamma_{rq}}} \sum_{i=1}^{n_{\gamma_{rp},\gamma_{rq}}} \psi(n_i^3) - \psi(n_i^{1,3}) - \psi(n_i^{2,3})$$

where ψ denotes the digamma function.

Again, as for CTMI (Section 4.2 of Chapter 4), to test independencies through TCE, we rely on the following permutation test (with the permutation scheme presented by Runge [2018]):

Definition 43 (Permutation test of TCE)**.** *Given X^p, X^q and $X^{\mathbf{R}}$, the p-value associated to the permutation test of TCE is given by:*

$$p = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}_{\widehat{TCE}((X^p)_b \to X^q | X^{\mathbf{R}}) \geq \widehat{TCE}(X^p \to X^q | X^{\mathbf{R}})}, \tag{5.3}$$

where $(X^p)_b$ is a permuted version of X^p and follow the local permutation scheme presented in Runge [2018].

The method is the following, detailed in Algorithm 4. Starting with a fully directed graph (with one sided edges coming from a causal ordering), the first step consists of removing arrow between nodes that are unconditionally independent: for each pair of nodes, a test of TCE is computed and the edge is removed if their dependency is insignificant given the threshold α. Once this is done, the algorithm checks, for the remaining oriented edges, whether two time series are conditionally independent or not given a set of parents of the arrow side node: in the first iteration the set of parents is of size one and then it gradually increases until either the edge between X^p and X^q is removed or all subsets of parents of X^q have been considered. Note that we make use of the same strategy as the one used in PC-stable [Colombo and Maathuis, 2014b], which consists in sorting time series according to their TCE scores and, when an independence

is detected, removing all other occurrences of the time series. This leads to an order-independent procedure.

The following theorem states that the graph obtained by the above procedure is the true one.

Theorem 6. *Given the true ordering of the causal process, Algorithm 4 is complete.*

Proof. Similarly to PC, Algorithm 4 prunes all unnecessary edges by removing edges that are conditionally independent given a subset $X^{\mathbf{S}}$. Thanks to the causal order, the possible subsets space is reduced. By removing all links that are conditionally independent, by causal Markov condition, adjacency faithfulness and causal sufficiency we are left with links that are directly causal and which are orientated with respect to causal ordering. □

Finally, similarly to `PCTMI`, given the graph $\mathcal{G}$ inferred with the above procedure, one can verify for each node X^q in $\mathcal{G}$ if it is self causal by checking if there exists a $\gamma > 0$ such that for all t, $X_t^q \not\perp\!\!\!\perp X_{t-\gamma}^q \mid \mathrm{Par}(X^q)$ in $\mathcal{G}$.

5.3 Toward a pairwise strategy

As presented in Section 5.2, a multivariate approach can be used to detect causal ordering which was already proven to indentify the right order [Peters et al., 2013] and which would serve well our purpose. But since the presented procedure uses a regression function estimator, it is subject to the curse of dimensionality when d is too large compared to N. So we also consider a pairwise version of the procedure which consists on estimating for each pair of time series (X^q, X^p) two regression functions

$$f^q : \{(X_{t-\gamma_{max}}^{(p;\gamma_{max}+1)}), (X_{t-\gamma_{max}}^{(q;\gamma_{max})})\} \mapsto X_t^q$$
$$f^p : \{(X_{t-\gamma_{max}}^{(q;\gamma_{max}+1)}), (X_{t-\gamma_{max}}^{(p;\gamma_{max})})\} \mapsto X_t^p.$$

Then we compare the dependency of the residuals of those two functions with their inputs, and as before the potential cause is the one that is mapped by the function that yields the higher dependency, i.e., we choose the causal direction that satisfies the most a bivariate ANM. We do not guarantee that the theory presented in Section 5.2 hold for this approximation, and do not prove that the inferred graph contains the real one. Nevertheless, we conjecture that such strategy might be useful in practice.

5.3.1 Time complexity reduction through multitask learning and denoising

In what follows, we focus on the pairwise strategy. Another issue rises from using regression function: they might be computationally expensive. So here we present a way to reduce the number of functions that have to be learned. First, we present a version of our proposal in an i.i.d. setting and then we apply it for time series.

In case of i.i.d. data

Here, X represents a set of random variables.

As shown in Section 3.4, in case of i.i.d. data, ANMs simply consider, in the bivariate case, that the effect is a function of its cause *plus* a noise term independent of the cause. No further assumption is made regarding the function relating the cause to the effect. The corresponding structural causal model is given by:

$$\begin{aligned} X^p &:= \xi^p \\ X^q &:= f^q(X^p) + \xi^q, \qquad X^p \perp\!\!\!\perp \xi^q. \end{aligned}$$

An important property of ANMs is that they are usually identifiable except for some specific distributions contained in a 3-dimensional affine space [Hoyer et al., 2009]. Recall that within ANMs, the direction of the causal relation is determined according to the lowest dependence between the potential cause and its residual when predicting the potential effect. In the case where, we have many causal relations to find and we rely on a pairwise strategy, ANM computes regression functions between all pairs of variables, which is of course problematic when the number of variables is important but also when the number of observations is important as each regression function will take more time to be estimated in this case. We specifically address these problems here and introduce a procedure that dispenses with training many regression functions. Intuitively, one can use an autoencoder to estimate the relations between all variables and *mask* (in a sense described below) some of the inputs and outputs of this autoencoder to obtain regressors between subsets of variables. By doing so, one dispenses with computing many different regressors. In addition, the regressors obtained are simple and scale well wrt the number of variables and observations.

Let us assume two bivariate data sets, $\mathcal{D}_n := (x_i^p, x_i^q)_{i=1}^n$, and $\mathcal{D}_n^{'} := (x_i^{p'}, x_i^{q'})_{i=1}^n$, both consisting of i.i.d. observations from P_{X^p,X^q} and let $\underline{x}^p$

denote the set of values $(x_1^p, \cdots, x_n^p)$ ($\underline{x}^q$, $\underline{x}^{p'}$, ... are defined in the same way). The **causal ordering procedure** [Mooij et al., 2016] for identifying bivariate causal graphs in ANMs can be summarized as follows:

1. Using $\mathcal{D}_n$, learn $\hat{f}^q$ (resp. $\hat{f}^p$), an estimator of the regression function which maps x^p (resp. x^q) to $\mathbb{E}(X^q|X^p = x^p)$ (resp. $\mathbb{E}(X^p|X^q = x^q)$);
2. On $\mathcal{D}_n'$, compute residuals $\hat{\underline{e}}^{q'} = \underline{x}^{q'} - \hat{f}^q(\underline{x}^{p'})$ and $\hat{\underline{e}}^{p'} = \underline{x}^{p'} - \hat{f}^p(\underline{x}^{q'})$;
3. Output $X^p \to X^q$ if $\hat{C}(\underline{x}^{p'}, \hat{\underline{e}}^{q'}) < \hat{C}(\underline{x}^{q'}, \hat{\underline{e}}^{p'})$ and $X^q \to X^p$ if $\hat{C}(\underline{x}^{q'}, \hat{\underline{e}}^{p'}) < \hat{C}(\underline{x}^{p'}, \hat{\underline{e}}^{q'})$, where $\hat{C}$ is an estimator of the dependence between the two variables (as measured through sets of values).

If the regression functions $\hat{f}^q$ and $\hat{f}^p$ are *suitable* (i.e. the mean squared error between true and predicted residuals vanishes asymptotically in expectation) and if the score estimator $\hat{C}$ is consistent, then the above inference procedure is consistent.

As mentioned before, we want to use an autoencoder to estimate the relations between variables and then mask some of its inputs and outputs to obtain regressors between subsets of variables. The autoencoders we consider in this study are based on multilayer perceptrons (MLP) with only one hidden layer. Assuming a linear function at the output layer and a non-linear, squashing function σ at the input layer [3], the class of such MLPs takes the form:

$$\mathcal{F}_n = \left\{ \sum_{i=1}^{k_n} c_{i,j}\sigma(\mathbf{a_i}^T\mathbf{u} + b_i) + c_{0,j} : 1 \leq j \leq d', k_n \in \mathbb{N}, \right. \\ \left. (\mathbf{a_i}, \mathbf{u}) \in \mathbb{R}^d, b_i \in \mathbb{R}, \sum_{i=1}^{k_n}\sum_{j=1}^{d'} |c_{i,j}| \leq \beta_n \right\} \tag{5.4}$$

with d (resp. d'), k_n and β_n corresponding respectively to the dimension of the input (resp. output) of the MLP, to the number of hidden units and to a constraint on output weights. This class of function is weakly universally consistent:

Theorem 7 (extension of Theorem 16.1 of Györfi et al. [2002] for $d' > 1$)**.** *Let $\mathcal{F}_n$ be the class of neural networks defined in* (5.4)*, $\hat{f}_{mlp}(.; \mathcal{D}_n)$ be the network that minimizes the empirical L_2 risk in $\mathcal{F}_n$. If k_n and β_n satisfy, for $n \to +\infty$: $k_n \to +\infty, \beta_n \to +\infty$, and $k_n\beta_n^4 \log(k_n\beta_n^2)/n \to 0$, then $\hat{f}_{mlp}(.;., \mathcal{D}_n)$ is weakly universally consistent for all distributions of input and output variables*

3. In practice, we consider a more general class of functions.

$(\mathbf{U}, \mathbf{V})$ *with, for all* $1 \leq j \leq d'$, $\mathbb{E}(V_j^2) < \infty$*:*

$$\lim_{n\to\infty} \mathbb{E} \int ||\hat{f}_{mlp}(\mathbf{u}; \mathcal{D}_n) - \mathbb{E}(\mathbf{V}|\mathbf{U} = \mathbf{u})||_2^2 d\mathbf{u} = 0.$$

Therefore, by Lemma 19 of Mooij et al. [2016], $\mathbf{u} \mapsto \hat{f}_{mlp}(\mathbf{u}; \mathcal{D}_n)$ is a suitable function. Let us now consider the case where $\mathbf{U} = \mathbf{V} = (X^p \ X^q)^T$ and where the MLP considered is a *denoising* autoencoder [Vincent et al., 2008] that will be denoted by $\hat{f}_{ae}(.; \mathcal{D}_n)$. In our denoising autoencoder, one variable, randomly chosen, is arbitrarily set to 0 in the input, but not in the output, at each iteration during training, which enables to reconstruct a corrupted version of the data. One thus considers different types of inputs, corresponding to whether or not a variable has been set to 0. We further denote by $\hat{f}_{ae}^q$ (resp. $\hat{f}_{ae}^p$) the value predicted by the autoencoder for the output corresponding to X^q (resp. X^p). Then, from Theorem 7, as all expectations are positive, one has:

$$\lim_{n\to\infty} \mathbb{E} \int (\hat{f}_{ae}^q(\mathbf{u}; \mathcal{D}_n) - \mathbb{E}(\mathbf{X}^q|\mathbf{U} = \mathbf{u}))^2 d\mathbf{u} = 0, \tag{5.5}$$

and similarly for $\hat{f}_{ae}^p$.
Focusing first on variable X^q, we denote by $\mathbf{u}_{|x^q=0}$ the situation in which the input variable X^q has been set to 0 and by $\mathbf{u}_{|x^q\neq 0}$ the situation in which it has not been changed. One can decompose the expectation in Eq. 5.5 according to these two cases:

$$\begin{aligned}\int (\hat{f}_{ae}^q(\mathbf{u}; \mathcal{D}_n) - \mathbb{E}(\mathbf{X}^q|\mathbf{U} = \mathbf{u}))^2 d\mathbf{u} \\ = \int (\hat{f}_{ae}^q(\mathbf{u}_{|x^q=0}; \mathcal{D}_n) - \mathbb{E}(\mathbf{X}^q|\mathbf{U} = \mathbf{u}_{|x^q=0}))^2 d\mathbf{u}_{|x^q=0} \\ + \int (\hat{f}_{ae}^q(\mathbf{u}_{|x^q\neq 0}; \mathcal{D}_n) - \mathbb{E}(\mathbf{X}^q|\mathbf{U} = \mathbf{u}_{|x^q\neq 0}))^2 d\mathbf{u}_{|x^q\neq 0}.\end{aligned} \tag{5.6}$$

Hence, exploiting again the fact that all quantities are positive in the right-hand side of Eq. (5.6) and that the left-hand side of Eq. (5.5) is equal to zero for $n \to \infty$, one obtains:

$$\lim_{n\to\infty} \mathbb{E} \int (\hat{f}_{ae}^q(\mathbf{u}_{|x^q=0}; \mathcal{D}_n) - \mathbb{E}(\mathbf{X}^q|\mathbf{U} = \mathbf{u}_{|x^q=0}))^2 d\mathbf{u}_{|x^q=0} = 0,$$

and similarly for $\hat{f}_{ae}^p$ and $\mathbf{u}_{|x^p=0}$.

Thus, the function $\mathbf{u} \mapsto \hat{f}_{ae}^q(\mathbf{u}_{|x^q=0}; \mathcal{D}_n)$, regressing X^q on X^p and obtained by setting the input X^q of the denoising autoencoder considered

above to 0, is weakly universally consistent. By Lemma 19 of Mooij et al. [2016], this function is also suitable, and so is the function $\mathbf{u} \mapsto \hat{f}^p_{ae}(\mathbf{u}_{|x=0}; \mathcal{D}_n)$ regressing X^p on X^q.

This leads us to the following consistency result:

Theorem 8. *Let X^p, X^q be two real-valued random variables with joint distribution P_{X^p,X^q} that either satisfies an ANM $X^p \to X^q$, or $X^q \to X^p$, but not both. Suppose we are given a training data set $\mathcal{D}_n$ and a test data set $\mathcal{D}'_n$ in the data splitting scenario. Let $k : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be a bounded non-negative Lipschitz-continuous kernel. Then, the **causal ordering procedure** in which $\hat{C}$ is consistent, $\hat{f}^q(x^p) = \hat{f}^q_{ae}(\mathbf{u}_{|x^q=0}; \mathcal{D}_n)$ and $\hat{f}^p(x^q) = \hat{f}^p_{ae}(\mathbf{u}_{|x^p=0}; \mathcal{D}_n)$ is a consistent procedure for estimating the direction of the ANM.*

The proof of Theorem 8 directly parallels the proof of Corollary 21 of Mooij et al. [2016] and exploits the consistency of $\hat{C}$ and the suitability of the regression functions considered.

In case of time series

Returning to time series, we reduce the number of functions to consider, as in the case of i.i.d., however, here we do not need to fully rely on an autoencoder, since only a partial amount of the input needs to be encoded and decoded. When considering time series, we propose to map the past and the present $(t - \gamma_{max}, \cdots, t)$ of the two time series to their present (t). Then masking one of the two points at time t of its inputs (the points that correspond to the partial amount that needs to be encoded and decoded), for example X^p_t (resp. X^p_t), and one of its outputs, for example X^p_t (resp. X^q_t), to obtain specialized regressors *i.e.* an approximation of the regression that maps $(X^{(p;\gamma_{max}+1)}_{t-\gamma_{max}}, X^{(q;\gamma_{max})}_{t-\gamma_{max}})$ to X^q_t and an approximation of the regression that maps $(X^{(p;\gamma_{max})}_{t-\gamma_{max}}, X^{(q;\gamma_{max}+1)}_{t-\gamma_{max}})$ to X^p_t. Formally, given two time series X^p and X^q, we estimate:

$$\hat{f}^{pq} : (X^{(p;\gamma_{max}+1)}_{t-\gamma_{max}}, X^{(q;\gamma_{max}+1)}_{t-\gamma_{max}}) \mapsto \mathbb{E}(X^p_t, X^q_t, |X^{(p;\gamma_{max}+1)}_{t-\gamma_{max}}, X^{(q;\gamma_{max}+1)}_{t-\gamma_{max}}),$$

where one of X^p_t and X^p_t of the input (but not in the output) is set arbitrary to 0 during training. Once the regression function is learned, we compute the residuals: for $l \in \{p, q\}$,

$$\xi^l = \hat{X}^l_t - \hat{f}^{pq}((X^{(p;\gamma_{max}+1)}_{t-\gamma_{max}}, X^{(q;\gamma_{max}+1)}_{t-\gamma_{max}})_{|x^l_t=0})_l. \tag{5.7}$$

Algorithm 5 NBCBk Part I: Noise based approach to order causes with knock in kock out

Require: X a d-dimensional time series, $\gamma_{\max} \in \mathbb{N}$ the maximum number of lags
  $\mathcal{G}$ an empty graph with nodes $=\{X^1, \ldots, X^d\}$
  for each pair of nodes (X^p, X^q) in $\mathcal{G}$ **do**
    Learn $\hat{f}^q : \{(X_{t-\gamma_{max}}^{(p;\gamma_{max}+1)}), (X_{t-\gamma_{max}}^{(q;\gamma_{max}+1)})\} \mapsto (X_t^p, X_t^q)$
    for $l \in \{p, q\}$ **do**
      Compute ξ^l from Eq. (5.7)
    end for
    Compute c_p from Eq. (5.8)
    Compute c_q from Eq. (5.9)
    if $c_p > c_q$ **then**
      Add $X^p \to X^q$ to $\mathcal{G}$
    end if
  end for
  Return $\mathcal{G}$

Then, to estimate the dependency between the residuals and the input, we compute

$$c_p = \hat{C}((X_{t-\gamma_{max}}^{(p;\gamma_{max})}, X_{t-\gamma_{max}}^{(q;\gamma_{max}+1)}), \xi_t^p) \tag{5.8}$$

$$c_q = \hat{C}((X_{t-\gamma_{max}}^{(p;\gamma_{max}+1)}, X_{t-\gamma_{max}}^{(q;\gamma_{max})}), \xi_t^q) \tag{5.9}$$

where $\hat{C}$ is an estimator of a dependence measure. Finally, the inference rule is the following: if $c_p > c_q$ then $X^p \to X^q$ and if $c_2 > c_1$ then $X^q \to X^p$.

5.4 Conclusion

We have addressed in this study the problem of learning a summary causal graph on time series without being restricted to the Markov equivalent class even in the case of instantaneous relations. To do so, we followed a hybrid strategy. First we used a noise-based method to find the causal ordering between the time series under the assumption of additive noise models. Second, we used a constraint-based method to prune unnecessary parents and therefore ending up with an oriented causal graph. The second step heavily relies on a new temporal causation entropy measure that generalizes the causation entropy by removing the restriction of one

time lag. Finally we made a pairwise extension of our algorithm which involved an introduction of a regression technique that permits in a bivariate case to use one regressor instead of two.

Chapter 6

Experiments

> Development of Western Science is based on two great achievements, the invention of the formal logical system (in Euclidean geometry) by the Greek philosophers, and the discovery of the possibility to find out causal relationships by systematic experiment (Renaissance).
>
> Albert Einstein

In this chapter, the causal discovery methods introduced in this thesis is studied experimentally on several datasets. We propose first an extensive analysis on simulated data, generated from basic causal structures; then we perform an analysis on real world datasets. First, we describe the evaluation measures, the different settings of methods we compare with, the datasets, and then we describe the results.

6.1 Evaluation measures

To assess the quality of causal inference, we use three different measures of accuracy: the F-score regarding adjacencies in the graph (F1), the F-score regarding directed edges in the graph ($\overrightarrow{\text{F1}}$), and the F-score regarding self loops in the graph ($\mathring{\text{F}}1$). Firstly, F1 regarding adjacencies in the graph is computed as follows:

$$\text{F1} = 2\text{TA}/(2\text{TA} + \text{FA} + \text{FNA}),$$

where TA is the true adjacency, which refers to the correct inference of the existing edges between nodes. FA and FNA are respectively discovery of edges between nodes that does not exist (false adjacency) and false inference of the absence of the edges between the nodes, when the nodes are adjacent (false non adjacency). The F-score regarding directed edges in the graph is defined as:

$$\overrightarrow{\text{F1}} = 2\text{TC}/(2\text{TC} + \text{FC} + \text{FNC}),$$

where TC, FC and FNC respectively refers to correct inference of the true direct cause, the false inference of the direct cause in a case of absence of the direct cause between the nodes, and the false inference of the absence of the direct cause, when the direct cause is present. The scores F1 and $\overrightarrow{\text{F1}}$ do not consider self loops, while the following metric $\mathring{\text{F}}1$ is introduced to measure the accuracy regarding self loops. This metric is defined as:

$$\mathring{\text{F}}1 = 2\text{TS}/(2\text{TS} + \text{FS} + \text{FNS}),$$

where TS is a related to correct inference of the self loop, FS refers to false detection of self loop, FNS relates to false inference of absence of self loop in case when self loop exists.

Those criteria are computed for the summary causal graph. We distinguish the case of self caused variables for the summary causal graph, as some algorithms are assuming self causes, some are assuming no self

causes, and some are estimating also those self causes, and it can drastically change the performance. Finally, we also have a look at the adjacency of the summary causal graph, focusing on the edges, to validate the skeleton.

6.2 Methods and their use

In PCTMI and FCITMI introduced in Chapter 4, we fix the number of nearest neighbor to $k = 10$ for the causal temporal mutual information. NBCB introduced in Chapter 5 and its pairwise version denoted pwNBCB are fitting a Gaussian Process with zero mean and squared exponential covariance function. The hyper-parameters are automatically chosen by marginal likelihood optimization. For pwNBCBk, we consider a neural network composed of 5 hidden layers: the first two contain 10 neurons each with linear activation functions, the third is a sequential convolution that uses a Relu as an activation function with kernel size $K = 5$ and a padding $P = 2$. The last two hidden layers are similar to the first two. Adam optimizer is used with a learning rate 0.01 and 1000 epochs. The partial denoising sub neural network is set to denoise an observation with a probability 0.5. In case of denoising, one chooses one variable at random and forces its value to 0, while the others are left untouched. Here also we fix the number of nearest neighbor to $k = 10$ for the temporal causation entropy. The Python code of all our methods is available at https://github.com/kassaad/causal_discovery_for_time_series.

From the Granger family, we compare the pairwise implementation with the multivariate one (respectively GCPW and GCMV). Statistically, the full model is compared to the restricted model using a F-test. We implement the pairwise version, and use for GCMV the code available there: http://www.sussex.ac.uk/sackler/mvgc/. We also use TCDF through the implementation available at https://github.com/M-Nauta/TCDF. Some hyper parameters have to be defined: we use a kernel of size 4, a dilation coefficient 4, one hidden layer, a learning rate of 0.01, and 5000 epochs.

From the constraint-based family, we run PCMCI using either the partial correlation (PCMCI-PC) or the mutual information (PCMCI-MI) to measure the dependence, both provided in the implementation available at https://github.com/jakobrunge/tigramite. We also use oCSE, which we implement. In all those methods, the mutual information is estimated using k-nearest neighbour [Runge, 2018] which we also fix the number of nearest neighbor to $k = 10$. Since the output of those measures are necessarily positive given finite sample size and finite numerical preci-

sion, we uses a significance permutation test. Finally, we compare FCITMI with tsFCI, provided at https://sites.google.com/site/dorisentner/publications/tsfci, where independence or conditional independence are tested respectively with tests of zero correlation or zero partial correlation.

Among the noise-based approaches, we run VarLiNGAM and TiMINo, which are respectively available at https://github.com/cdt15/lingam and http://web.math.ku.dk/~peters/code.html. For VarLiNGAM, the regularization parameter in the adaptive Lasso is selected using BIC, and no statistical test is performed as we use the value of the statistic. TiMINo uses a test based on cross-correlation that can be derived from [Brockwell and Davis, 1986, Thm 11.2.3.].

For all the methods, we use $\gamma_{\max} = 5$ and when doing a statistical test, we use a significance level of 0.05.

A Python routine to use all the methods introduced here is available at https://github.com/kassaad/causal_discovery_for_time_series.

6.3 Dataset

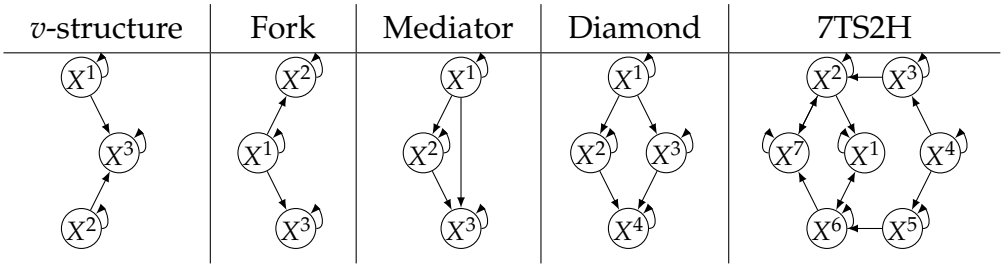
To illustrate the behavior of the causal inference algorithms we rely on both artificial and real-world datasets.

6.3.1 Simulated data

We first test our causal sufficient methods (PCTMI, NBCB, pwNBCB, pwNBCBk) on simulated data generated from four different causal structures: fork, v-structure, mediator and diamond. To evaluate the performance of the inference with respect to the length of time series, we consider several time stamps, from 125, 250, 500, and 1000. To test FCITMI we consider the structure 7TS2H which represents a nine nodes structure introduced in Spirtes et al. [2000]; seven nodes correspond to observational time series and two to hidden common causes, represented by double arrows). The structures are presented in Table 6.1. For each benchmark and for each length of time series, we generate randomly 10 data set available at https://dataverse.harvard.edu/dataverse/basic_causal_structures_additive_noise. The data generating process is the following: for all q, $X_0^q = 0$ and for all $t > 0$,

$$X_t^q = a_{t-1}^{qq} X_{t-1}^q + \sum_{\substack{(p,\gamma) \\ X_{t-\gamma}^p \in Par(X_t^q)}} a_{t-\gamma}^{pq} f(X_{t-\gamma}^p) + 0.1\zeta_t^q,$$

Table 6.1 – Structures of simulated data. $A \to B$ means that A causes B and $A \longleftrightarrow$ B represents the existence of a hidden common cause between A and B.



where $\gamma \geq 0$, a_t^{jq} are random coefficients chosen uniformly in $\mathcal{U}([-1;-0.1] \cup [0.1;1])$ for all $1 \leq j \leq d$, $\xi_t^q \sim \mathcal{N}(0, \sqrt{15})$ and f is a non linear function chosen at random uniformly between absolute value, tanh, sine, cosine.

To highlight the limitations of constraint-based methods including `PCTMI` and `FCITMI` we provide a more complicated configuration of the simulation of 10 data sets with 1000 time stamps of the structures: fork, mediator, and diamond. In this setting we consider a fork structure which is not unique in its Markov equivalence class, as all relations are instantaneous, so time reference is not useful to differentiate between common cause and intermediate cause. In the mediator and diamond structures we violate the assumption of faithfulness by considering linear relations and fixing coefficients in such a way that different causal path eliminate each other. Both structures are without self cause, and all relations are instantaneous. For the mediator structure, we consider $a^{13} = -a^{12}a^{23}$ and following Zhalama et al. [2016], for diamond structure we set the coefficient $a^{34} = -a^{12}a^{23}/a^{13}$.

6.3.2 Real data

Three different real datasets are considered in this study. We detail the performance of each method in the following paragraphs, but the results are summarized in Table 6.5.

Temperature

This bivariate time series available at `https://webdav.tuebingen.mpg.de/cause-effect/`, of length 168 is about indoor X^{in} and outdoor X^{out} measurements. We expect that there is the following causal link: $X^{\text{out}} \to X^{\text{in}}$.

Diary

This dataset available at `http://future.aae.wisc.edu`, provides 10 years (from 09/2008 to 12/2018) of monthly prices for milk X^m, butter X^b and cheddar cheese X^c, so the three time series are of length 124. We expect that the price of milk is a common cause of the price of butter and the price of cheddar cheese.

BOLD FMRI

The last real-world dataset benchmark is about FMRI (Functional Magnetic Resonance Imaging) that contains BOLD (Blood-oxygen-level dependent) datasets [Smith et al., 2011] for 28 different underlying brain networks. It measures the neural activity of different regions of interest in the brain based on the change of blood flow. There are 50 regions in total, each with its own associated time series. Since not all existing methods can handle 50 time series, datasets with more than 10 time series are excluded. In total we are left with 26 datasets containing between 5 and 10 brain regions. The original data is available at `https://www.fmrib.ox.ac.uk/datasets/netsim/index.html`, and a preprocessed version is available at `https://github.com/M-Nauta/TCDF/tree/master/data/fMRI`.

6.4 Numerical results

6.4.1 Simulated data

With causal sufficiency

We provide in Figure 6.1 the performance of the methods on simulated data. We compare summary causal graphs for 12 methods on the causal sufficient structures (first four structures in Table 6.1) using F1,$\overrightarrow{\text{F1}}$, and $\mathring{\text{F}}1$ metrics.

First, in the left column of Figure 6.1, we provide the F1 score of the adjacency matrix of the summary causal graph, so we focus on the skeleton. Overall, for all tested structures the performance is high and comparable for all the methods, except `VarLiNGAM` and `TCDF` that have lower performance. We know that `VarLiNGAM` is not adapted to this dataset as it infers linear relations, whereas the generation process is not linear for two different time series and linear for self caused time series. Moreover, we remark that results for v-structure and fork are very variable, with a high variance. Particularly for fork (and in some extent for v-structure), results are

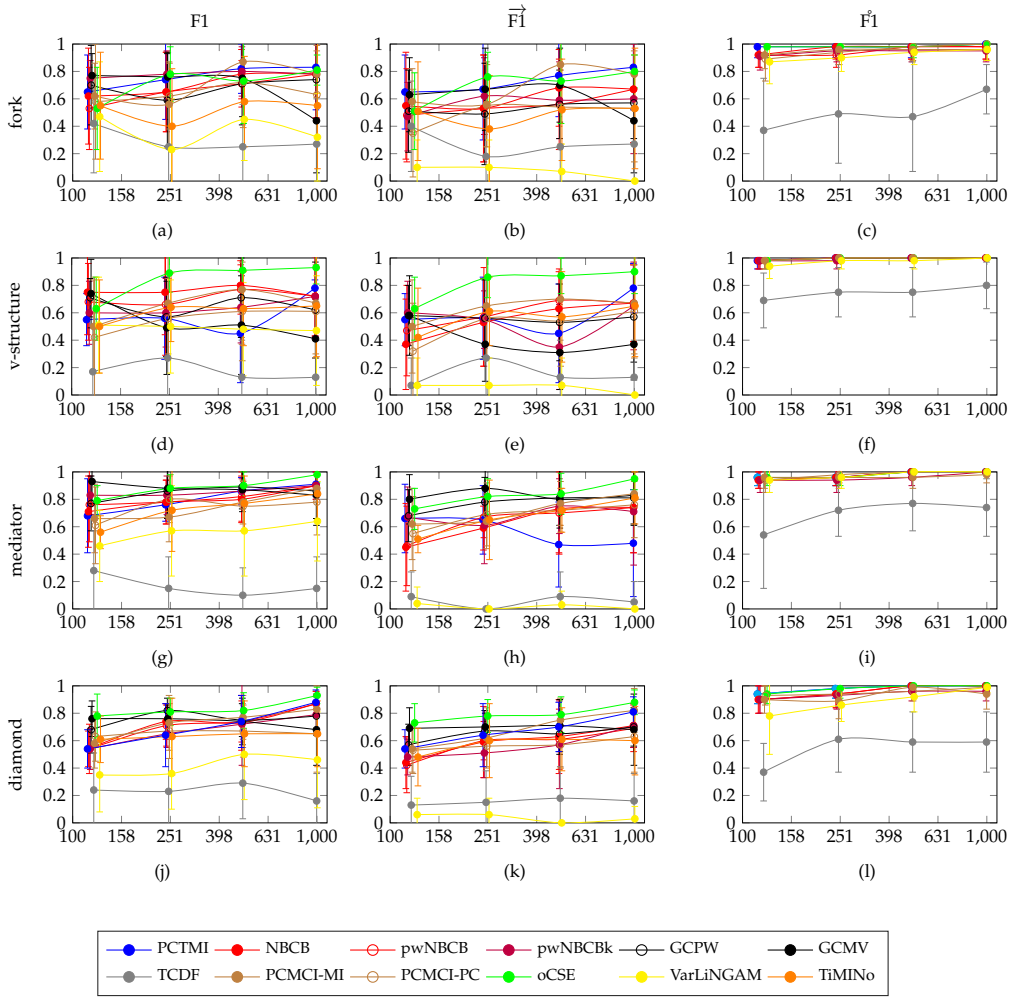


Figure 6.1 – Adjacency (F1), external causation ($\overrightarrow{F1}$) and self causation ($\mathring{F1}$) in the summary causation graph for all the methods on 4 simulated datasets (mean $\pm$ standard deviation). Results are computed for various time grid sizes, from 125 to 1000. A log-scale is used in abscissa.

also erratic, with no convergence for large time point observations. For all the tested structures constraint-based methods (PCMCT-MI, PCMCI-PC, and oCSE) and our methods (PCTMI, NBCB, pwNBCB, pwNBCBk) have similarly high performance, except the mediator structure, where the performance is high for most methods. All our methods have good results, and they are outperformed only by oCSE method. In particular, among our methods NBCB has the best results. Algorithm oCSE consistently performs better than the others.

Secondly, in the middle column of Figure 6.1 we consider the $\overrightarrow{\text{F1}}$ score to detect direct causal relations external to the time series. In comparison to the left column, the overall performance for all methods is slightly lower and more variable. As before VarLiNGAM and TCDF have lowest performance among other methods for all structures and also VarLiNGAM performs worse than in skeleton case. We can note similar erratic behavior, with no convergence for large time point observations for fork and v-structure. Same as in the left column, constraint-based methods (PCMCT-MI, PCMCI-PC, and oCSE) and our methods (PCTMI, NBCB, pwNBCB, pwNBCBk) have high performance, except for the mediator structure, where NBCB methods have relatively low performance compared to constraint based methods and to TiMINo, and the performance of PCTMI deteriorate with the increase of the size of the time series. The behavior of PCTMI on mediator is expected since in theory PCTMI cannot orient any edge in a mediator structure. In practice the few orientation that are made are due to some mistakes in the skeleton construction phase. In general, our methods have good results when good results is expected, and they are outperformed only by oCSE method. In contrast to the NBCB in adjacency case, the best method for external causation among our methods is PCTMI and it outperforms oCSE for fork structure. Algorithm oCSE has the best performance, for all structures, except fork.

Finally, we provide a distinction between external causations and self causations (when it is estimated) and present results for self causation in the right column of Figure 6.1. Some methods always consider that there is self causation, whereas some of them are estimating those links: TCDF, PCMCI, oCSE, VarLiNGAM and all our methods (PCTMI, NBCB, pwNBCB, and pwNBCBk). Performance for self causation for these methods are good, except for TCDF. Overall good performance is not surprising, because the relations in self causation case are linear.

Results illustrated in Figure 6.1 confirm the good performance of our methods. Focusing on results for external causality, which is more relevant to the problem statement of this thesis, our method PCTMI has better results than the others for fork structure and outperformed only by oCSE

for other structures (except for mediator). High performance of oCSE is related to strong assumptions used by this method namely that a cause relations is necessary 1-order Markov. In general constraint based approaches perform best for these simulated data, because the assumptions in which they are build upon are met, however, it is not the case for unfaithful data.

Unfaithful

Table 6.2 – Results obtained on the unfaithful simulated data for the different structures with 1000 observations. We report the mean and the standard deviation of the F1 score. The best results are in bold.

	Fork		unfaith. Mediator		unfaith. Diamond	
	F1	$\overrightarrow{\text{F1}}$	F1	$\overrightarrow{\text{F1}}$	F1	$\overrightarrow{\text{F1}}$
PCTMI	0.83 ± 0.16	0.07 ± 0.19	0.77 ± 0.11	0.27 ± 0.18	0.91 ± 0.36	0.34 ± 0.18
NBCB	0.87 ± 0.15	0.45 ± 0.34	0.8 ± 0.0	0.56 ± 0.26	0.93 ± 0.08	$\mathbf{0.5} \pm 0.31$
pwNBCB	$\mathbf{0.91} \pm 0.15$	$\mathbf{0.49} \pm 0.31$	0.84 ± 0.08	0.46 ± 0.23	0.91 ± 0.09	0.39 ± 0.22
pwNBCBk	0.85 ± 0.15	0.41 ± 0.29	0.82 ± 0.06	0.4 ± 0.29	0.92 ± 0.06	0.45 ± 0.25
GCPW	0.12 ± 0.24	0.05 ± 0.15	0.28 ± 0.37	0.12 ± 0.27	0.32 ± 0.28	0.14 ± 0.23
GCMV	0.15 ± 0.3	0.1 ± 0.3	0.33 ± 0.21	0.16 ± 0.28	0.32 ± 0.14	0.16 ± 0.28
TCDF	0.39 ± 0.42	0.34 ± 0.37	0.74 ± 0.12	0.4 ± 0.22	0.48 ± 0.21	0.33 ± 0.17
PCMCI-MI	0.28 ± 0.29	0.07 ± 0.019	0.27 ± 0.29	0.05 ± 0.15	0.41 ± 0.25	0.20 ± 0.22
PCMCI-PC	0.41 ± 0.36	0.31 ± 0.27	0.44 ± 0.31	0.21 ± 0.21	0.25 ± 0.22	0.11 ± 0.18
oCSE	0.18 ± 0.28	0.12 ± 0.24	0.05 ± 0.15	0.05 ± 0.15	0.12 ± 0.18	0.08 ± 0.16
VarLiNGAM	0.6 ± 0.42	0.05 ± 0.15	$\mathbf{0.98} \pm 0.06$	0.0 ± 0.0	$\mathbf{0.94} \pm 0.04$	0.02 ± 0.06
TiMINo	0.67 ± 0.23	0.45 ± 0.15	0.95 ± 0.15	$\mathbf{0.64} \pm 0.08$	0.78 ± 0.06	0.49 ± 0.03

Table 6.2 shows the results of applying different methods on data with a fork structure, that is not unique in its Markov equivalent class and on the unfaithful simulated data (see description in Section 6.3). We remind that in these simulations we do not have self causes (self loops), so we consider only F1 and $\overrightarrow{\text{F1}}$ metrics.

For the fork structure in the left column of Table 6.2 we can see that the results for the different methods significantly differ from each other and overall their performance is better for F1 than for $\overrightarrow{\text{F1}}$. Most of the constraint-based approaches have low performance for the adjacency F1-score. Interestingly, oCSE, which was the best for faithful data, shows almost the worst result here. However, our method PCTMI, which is too related to the constraint based family, has good performance. But the best results are achieved by our hybrid methods NCBC, pwNBCB, pwNBCBk. For $\overrightarrow{\text{F1}}$

metric, constraint-based approaches, including PCTMI, show poor performance. Our methods NBCB, pwNBCB and pwNBCBk perform better in comparison to all other methods for the $\overrightarrow{\text{F1}}$ metrics, along with the noise-based algorithm TiMiNo. Interestingly, pwNBCB yields the best accuracy.

For both unfaithful structures (the middle and the right columns of Table 6.2) we see the same pattern we saw with the fork structure: the results are highly variable. NBCB, pwNBCB, pwNBCB in addition to TiMiNo perform best, and constraint-based approaches perform poorly. For the mediator structure TiMiNo worked best, while on the diamond structure, NBCB has the best accuracy.

Among all our methods NBCB is the best method for both the mediator and the diamond structures, while pwNBCB is the best for the fork structure. We also see that TCDF has relatively good results while having moderate performance on the fork structure and low performance for faithful data. Good results for noise-based method TiMiNo and our hybrid methods NBCB, pwNBCB, and pwNBCBk are not surprising because these methods are not restricted to the Markov equivalence class nor to orientation faithfulness which is violated in the last two structures. At the same time, another noise-based method VarLiNGAM has a high F1 score, but fails to detect direct causes. Another important observation from Table 6.2 is that PCTMI manages to keep a good F1 score, which means that even though it did not detect the true causal relations, it succeeded in estimating the structure correctly.

Different sampling rate

Table 6.3 – Results obtained by PCTMI with different sampling rates on the four structures: fork, v-structure, mediator, and diamond. We report the mean of the F1 score and the standard deviation for the two measures.

	v-structure	Fork	Mediator	Diamond
F1	0.63 ± 0.23	0.80 ± 0.31	0.80 ± 0.29	0.71 ± 0.26
$\overrightarrow{\text{F1}}$	0.56 ± 0.30	0.80 ± 0.31	0.58 ± 0.31	0.66 ± 0.24
$\mathring{\text{F}}1$	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0

We also assessed the behavior of PCTMI when the time series have different sampling rates. We present here results only for PCTMI because other methods are not applicable to the data with different sampling rates.

The results for faithful data with four structures: fork, v-structure, mediator, and diamond are presented in Table 6.3. As one can see, its perfor-

mance is close to the ones obtained with equal sampling rates (Figure 6.1), the degradation being not really surprising as one has less data to rely on. This method has the best performance for the mediator structure and the worst one for the v-structure.

Without causal sufficiency

Table 6.4 – Results obtained by FCITMI and tsFCI on 7TS2H with 1000 observations. We report the mean of the F1 score and the standard deviation. The best results are in bold.

	F1	$\overrightarrow{\text{F1}}$	$\mathring{\text{F1}}$
FCITMI	**0.90** $\pm$ 0.05	**0.44** $\pm$ 0.11	1.0 $\pm$ 0.0
tsFCI	0.77 $\pm$ 0.06	0.37 $\pm$ 0.09	1.0 $\pm$ 0.0

The inference of causal graph in the presence of hidden confounders is a difficult but an important problem, seldom exploited. Our FCITMI algorithm that addresses this problem is compared with tsFCI on data with hidden common causes. From Table 6.4 we can note that both methods have reasonable performance, while FCITMI performs remarkably better for all metrics.

6.4.2 Real data

We provide in Table 6.5 the results for Temperature, Diary and FMRI for all the methods. Since we do not know if self causation exist in the causal mechanisms that's behind these datasets, we only report the F1 and the $\overrightarrow{\text{F1}}$ metrics.

Temperature

VarLiNGAM and TCDF wrongly infers no causal relation, GCPW and GCMV infer a bi-directed arrow and TiMINo remains undecided. PCMCI-PC, PCMCI-MI, oCSE, PCTMI, NBCB, pwNBCB and pwNBCBk correctly infer $X^{\text{out}} \to X^{\text{in}}$.

Diary

VarLiNGAM wrongly infers X^b as common cause of X^m and X^c, GCPW and GCMV wrongly infers $X^m \leftrightarrow X^b \to X^c \to X^m$ and TiMINo only infers one wrong causal relation $X^c \to X^m$. TCDF infers no causal relation. PCMCI-PC

and PCMCI-MI wrongly infer the causal chain $X^m \to X^c$. PCTMI infers one correct causal relation $X^c \to X^m \to X^b$ oCSE, NBCB, pwNBCB, and pwNBCBk correctly infer the causal relations but also add a wrong causal $X^c \to X^b$.

BOLD FMRI

PCTMI, pwNBCBk and VarLiNGAM clearly outperforms other methods. All other methods are comparable, except TCDF which performs very poorly. Interestingly, PCMCI-PC performs better than PCMCI-MI, and VarLiNGAM outperforms TiMINo which suggests the possibility of existence of linear causal relations.

Table 6.5 – Results for real datasets. We report the mean and the standard deviation of the F1 score. The best results are in bold.

	Temperature		Diary		FMRI	
	F1	$\overrightarrow{\text{F1}}$	F1	$\overrightarrow{\text{F1}}$	F1	$\overrightarrow{\text{F1}}$
PCTMI	**1**	**1**	0.67	0.67	0.47 ± 0.31	0.32 ± 0.17
NBCB	**1**	**1**	0.8	**0.8**	0.76 ± 0.16	0.40 ± 0.21
pwNBCB	**1**	**1**	0.8	**0.8**	0.78 ± 0.13	0.39 ± 0.21
pwNBCBk	**1**	**1**	0.8	**0.8**	$\mathbf{0.85} \pm 0.06$	0.44 ± 0.15
GCPW	**1**	0.66	0.8	0.28	0.47 ± 0.24	0.31 ± 0.17
GCMV	**1**	0.66	0.8	0.33	0.56 ± 0.18	0.24 ± 0.18
TCDF	0	0	0	0.0	0.13 ± 0.21	0.07 ± 0.13
PCMCI-MI	**1**	**1**	**1**	0.5	0.38 ± 0.23	0.22 ± 0.18
PCMCI-PC	**1**	**1**	**1**	0.5	0.44 ± 0.22	0.29 ± 0.19
oCSE	**1**	**1**	0.8	**0.8**	0.25 ± 0.26	0.16 ± 0.20
VarLiNGAM	0	0	0.5	0.0	0.74 ± 0.27	$\mathbf{0.49} \pm 0.28$
TiMINo	0	0	0.67	0.0	0.55 ± 0.21	0.32 ± 0.11

6.5 Complexity analysis

Our proposed methods benefits from a smaller number of tests compared to constraint-based methods that infer the full temporal graph. In the worst case, the complexity of PC in a temporal graph is bounded by:

$$\frac{(d\gamma_{max})^2(d\gamma_{max}-1)^{k-1}}{(k-1)!},$$

where k represents the maximal degree of any vertex and γ_{max} is the maximum number of lags. Each operation consists in conducting a significance

test on a conditional independence measure. Algorithms adapted to time series, as PCMCI [Runge et al., 2019], rely on time information to reduce the number of tests. Indeed, with this information, the complexity can be divided by 2 (when instantaneous relations are not taken into account). PCTMI and NBCB infer a summary causal graph, which limits the number of decisions that need to be taken. Indeed, PCTMI's complexity in the worst case (when all relations are instantaneous) is bounded by:

$$\frac{d^2(d-1)^{k-1}}{(k-1)!},$$

whereas NBCB's complexity in the worst case is bounded by:

$$d^2 f(n,d) + \frac{d^2(d-1)^{k-1}}{(k-1)!},$$

where $f(n,d)$ is the complexity of the user-specific regression method.

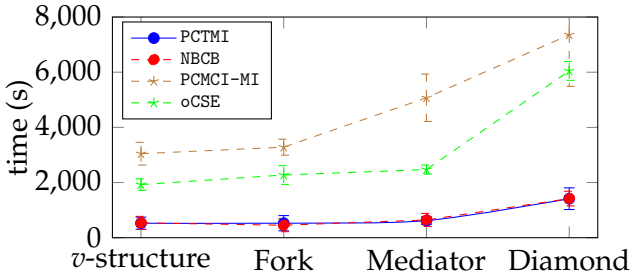


Figure 6.2 – Time computation (in seconds) for PCTMI, NBCB, PCMCI-MI and oCSE.

Figure 6.2 provides an empirical illustration of the difference in complexity of the two approaches on the four structures (v-structure, fork, mediator, diamond), sorted according to their number of nodes, their maximal out-degree and their maximal in-degree. The time is given in seconds. As one can note, both PCTMI and NBCB are always faster than PCMCI-MI and oCSE, the difference being more important when the structure to be inferred is complex.

6.6 Conclusion

In this chapter, we validated our algorithms experimentally on simulated and real datasets. Most of the obtained results are expected. On the

faithful data constraint-based methods have the best results, including our PCTMI algorithm. Importantly, proposed hybrid methods (NBCB, pwNBCB, pwNBCBk) perform sometimes on par with the best performing methods. On the unfaithful simulated data and the fork structure, that is not unique in its Markov equivalent class, we can see the drop in performance for constraint-based methods, while the noise-based methods work better. In this case our hybrid methods have high performance and in some experiments outperform the noise-base methods. On the real data, our methods have the best overall performance. In particular, on the FMRI dataset, PCTMI have the second best accuracy after the VarLinGAM. At the same time VarLinGAM algorithm have poor performance on other real datasets and on both types of simulation datasets. For the temperature data, all our methods inferred the correct causal relation. Finally, on the diary dataset our hybrid methods (NBCB, pwNBCB, pwNBCBk) along with oCSE have the best accuracy.

For some specific data types, we showed that our methods perform well and could be used in practice. In particular, our results of the PCTMI algorithm on time series with different sampling rate, confirm the practical value of the method. In case of non causal sufficient data our proposed method FCITMI performs better then state-of-the-art method.

To conclude, PCTMI proved to work well when some specific assumptions are held and with small time complexity in comparison with similar approaches that use mutual information. The above experiments demonstrated that for each type of data, NBCB, pwNBCB, and pwNBCBk performed almost as well as the best method specific for that type of data, which makes them the best performing overall.

Chapter 7

Conclusion

> Why? Causal inference is all about taking this question seriously.
>
> Judea Pearl

Conclusion and future works In this thesis, we focused on discovering causal relations from observational time series. We have shown how summary causal graphs can be inferred without truly building a full-time causal graph nor a window causal graph. As a first contribution, we considered the problem of inferring a summary causal graph in which only instantaneous relations are restricted to the Markov equivalent class and while relaxing the assumption that different time series should have an equal sampling rate. Our proposition relies on the constraint-based approaches for causal discovery and the entropy reduction principle as well as on the new causal temporal mutual information measure which can be used to assess the mutual information between time series. Taking advantage of the additive noise principle, we also tackled the problem of instantaneous causal relations that do not belong to the Markov equivalent class. The elaboration of this method necessitated the introduction of the temporal causal entropy which can be used to asymmetrically assess the mutual information. We highlight that we extended the base approach into a pairwise direction and we systematically introduced a method that uses multitask learning and a denoising technique to accelerate the estimation of two regression functions. Our findings regarding the pairwise extension in time series should be regarded as initial results rather than a full theoretical answer.

The correctness of our base algorithms was proved theoretically and illustrated experimentally. The experiments conducted on different simulated and real benchmark datasets, showed both the efficacy and efficiency of our approaches compared to other methods. Table 7.1 summarizes the characteristics of all algorithms presented in this thesis with respect to the problems we wanted to solve. The columns in the Table show whether the algorithm can infer a summary causal graph without necessarily going through a time consuming inference of a window causal graph and if it can discover instantaneous and time delayed causal relations as well as self causes. The table also tells us which algorithm can handle observable counfounder and hidden counfounder. In addition, we can see which algorithms are not restricted to the Markov equivalence class, do not assume orientation faithfulness and can handle different sampling rate. Furthermore, some methods can handle no-linear relations but others do not. Regardless of their strengths and weaknesses, we stress that all these algorithms including the ones presented in this thesis should be used with care. They can only infer causal relations relative to the set of observable V and to a set of assumptions. Any change in the former will lead to different causal structure and any change in the latter might lead to substantial decrease in performance. So in other words, these algorithms are not tools

to answer *why* questions, instead they are tools that assist the expert or the researcher in his quest of taking the question *why* seriously.

Future works Many different adaptations, tests, and experiments have been left for the future (i.e. the experiments with real data are usually very time consuming). Among others, future work will concern deeper analysis of particular extensions of base methods. Fo instance, we have provided a reference to a motivation for the question estimating a graph from sequences, namely, misaligned time series which should be regarded as intuitive guidance for future research. Although it seems trivial to adapt the `FCITMI` algorithm for selection bias (by using Rules 5, 6 and 7), an investigation and an experimental validation is needed to check whether or not the these rules work well in time series framework.

Concerning the imperfectness of time series in real application, aside of different sampling rate, the field of causal discovery between time series still face tremendous challenges such as missing data and non-stationarity. In addition, sometimes the informative part of a time series is negligible compared to the size of the time series, in such cases most current causal discovery algorithms are bound to fail. So further research is needed to investigate the possibility of searching for informative parts of time series and maybe conducting local discoveries.

Obviously, in this thesis we only focused on one type of time series: time series with continuous values and discrete time, in other words on quantitative data. However, causal mechanisms can be embodied in the form of events as much as they can be embodied in the form of processes. Some causal discovery algorithms (for example, the constraint-based approaches including `PCTMI`) can be easily adapted to qualitative data. In the case of `PCTMI`, we only need to find another estimator of the mutual information or the conditional mutual information that handle qualitative data. However, multivariate complex systems rarely contain data of the same type. For example the net flow constantly causes the CPU usage and the CPU usage causes the system to slow down only after reaching a certain threshold. In such a system we would want a discovery algorithm that supports causal mechanisms based on both events and processes, quantitative and qualitative variables.

Finally, we also think that future works should also consider collaborative causal discovery algorithms; collaborative in the sense that the algorithm should collaborate with the expert or the researcher in order to infer underlying causal graph of a given system.

Table 7.1 – Summary of the methods introduced in this thesis, and their main assumptions.

	Algorithm	Infer summary causal graph directly	Handle instantaneous relations	Can detect self causation	Confounders	Handle hidden Confounder	Not restricted to Markov equivalence class	Do not assume orientation Faithfulness	Handle different sampling rate	Handle non-linear relations
New methods	PCTMI	✓	✓	✓	✓	✗	✗	✗	✓	✓
	FCITMI	✓	✓	✓	✓	✓	✗	✗	✓	✓
	NBCB	✓	✓	✓	✓	✗	✓	✓	✗	✓
	pwNBCB	✓	✓	✓	✓	✗	✓	✓	✗	✓
	pwNBCBk	✓	✓	✓	✓	✗	✓	✓	✗	✓
Granger	GCPW	✓	✗	✗	✗	✗	✗	✓	✗	✗
	GCMV	✓	✗	✗	✓	✗	✗	✓	✗	✗
	TCDF	✗	✓	✓	✓	✓	✓	✓	✗	✓
Constraint-based	PCMCI-MI	✗	✓	✓	✓	✗	✗	✗	✗	✓
	PCMCI-PC	✗	✓	✓	✓	✗	✗	✗	✗	✗
	oCSE	✓	✗	✓	✓	✗	✗	✗	✗	✓
	tsFCI	✗	✓	✓	✓	✓	✗	✗	✗	✓
Noise-based	VarLiNGAM	✗	✓	✓	✓	✗	✓	✓	✗	✗
	TiMINo	✓	✓	✗	✓	✗	✓	✓	✗	✓

Bibliography

Séverine Affeldt and Hervé Isambert. Robust reconstruction of causal graphical models based on conditional 2-point and 3-point information. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, UAI'15, 2015. ISBN 978-0-9966431-0-8.

David J. Albers and George Hripcsak. Estimation of time-delayed mutual information and bias for irregularly and sparsely sampled time-series. *Chaos, Solitons & Fractals*, 45(6):853 – 860, 2012.

Ayesha R. Ali, Thomas S. Richardson, Peter Spirtes, and Jiji Zhang. Towards characterizing markov equivalence classes for directed acyclic graphs with latent variables. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, UAI'05, pages 10–17, Arlington, Virginia, USA, 2005. AUAI Press. ISBN 0974903914.

Steen A. Andersson, David Madigan, and Michael D. Perlman. A characterization of markov equivalence classes for acyclic digraphs. *Annals of Statistics*, 25(2):505–541, 1997. doi: 10.1214/aos/1031833662.

Gertrude Elizabeth Margaret Anscombe. Causality and determination. In E. Sosa M. Tooley, editor, *Causation*, pages 88–104. Oxford Up, 1993.

Aristotle and C. D. C. Reeve. *Physics*. Hackett Publishing Company, 2018.

Karim Assaad. A brief history of causality's principle. submitted.

Karim Assaad, Emilie Devijver, and Eric Gaussier. A survey on causal discovery for time series. submitted, a.

Karim Assaad, Emilie Devijver, Eric Gaussier, and Ali Aït-Bachir. Entropy-based discovery of summary causal graphs in time series. submitted, b.

Karim Assaad, Emilie Devijver, Eric Gaussier, and Ali Aït-Bachir. Scaling causal inference in additive noise models. In Thuc Duy Le, Jiuyong

Li, Kun Zhang, Emre Kıcıman Peng Cui, and Aapo Hyvärinen, editors, *Proceedings of Machine Learning Research*, volume 104 of *Proceedings of Machine Learning Research*, pages 22–33, Anchorage, Alaska, USA, 05 Aug 2019. PMLR.

Karim Assaad, Emilie Devijver, Eric Gaussier, and Ali Ait-Bachir. A mixed noise and constraint-based approach to causal inference in time series. In Nuria Oliver, Fernando Pérez-Cruz, Stefan Kramer, Jesse Read, and Jose A. Lozano, editors, *Machine Learning and Knowledge Discovery in Databases. Research Track*, pages 453–468, Cham, 2021. Springer International Publishing. ISBN 978-3-030-86486-6. doi: 10.1007/978-3-030-86486-6_28.

Adam B. Barrett, Lionel C. Barnett, and Anil K. Seth. Multivariate Granger causality and generalized variance. *Physical review E*, 81:041907, 2010. doi: https://doi.org/10.1103/PhysRevE.81.041907.

Helen Beebee. Causation and observation. In Helen Beebee, Christopher Hitchcock, and Peter Menzies, editors, *The Oxford Handbook of Causation*. Oxford University Press, 2009.

Patrick Bloebaum, Dominik Janzing, Takashi Washio, Shohei Shimizu, and Bernhard Schölkopf. Cause-effect inference by comparing regression errors. 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 900–909, Playa Blanca, Lanzarote, Canary Islands, 2018. PMLR. URL `http://proceedings.mlr.press/v84/bloebaum18a.html`.

Peter J Brockwell and Richard A Davis. *Time Series: Theory and Methods*. Springer-Verlag, Berlin, Heidelberg, 1986. ISBN 0387964061.

Andrea Brovelli, Mingzhou Ding, Anders Ledberg, Yonghong Chen, Richard Nakamura, and Steven L. Bressler. Beta oscillations in a large-scale sensorimotor cortical network: Directional influences revealed by granger causality. *Proceedings of the National Academy of Sciences*, 101(26): 9849–9854, 2004. ISSN 0027-8424. doi: 10.1073/pnas.0308538101.

Edwin Arthur Burtt. *The Metaphysical Foundations of Modern Physical Science*. Garden City, N.Y., Doubleday, 1926.

Rudolf Carnap. Meaning and necessity: A study in semantics and modal logic. *Mind*, 58(230):228–238, 1949.

Nancy Cartwright. Causal laws and effective strategies. *Noûs*, 13(4):419–437, 1979.

Nancy Cartwright. *Nature's Capacities and Their Measurement*. Oxford University Press, 1989.

Yonghong Chen, Govindan Rangarajan, Jianfeng Feng, and Mingzhou Ding. Analyzing multiple nonlinear time series with extended granger causality. *Physics Letters A*, 324:26–35, 2004. doi: 10.1016/j.physleta.2004.02.032.

David Maxwell Chickering. Learning equivalence classes of bayesian-network structures. *Journal of Machine Learning Research*, 2:445–498, 2002. ISSN 1532-4435. doi: 10.1162/153244302760200696.

Tianjiao Chu and Clark Glymour. Search for additive nonlinear time series causal models. *Journal of Machine Learning Research*, 9:967–991, 2008.

Nevin Climenhaga, Lane DesAutels, and Grant Ramsey. Causal inference from noise. *Noûs*, 2019. doi: 10.1111/nous.12300.

Diego Colombo and Marloes H. Maathuis. Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, 15 (116):3921–3962, 2014a.

Diego Colombo and Marloes H. Maathuis. Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, 15 (116):3921–3962, 2014b.

Diego Colombo, Marloes H. Maathuis, Markus Kalisch, and Thomas S. Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *Annals of Statistics*, 40(1):294–321, 2012. doi: 10.1214/11-AOS940.

Edwin Curley. Behind the geometrical method: A reading of spinoza's ethics. *Noûs*, 26(3):371–373, 1992. doi: 10.2307/2215962.

Adnan Darwiche. Human-level intelligence or animal-like abilities? *Commun. ACM*, 61(10):56–67, September 2018. ISSN 0001-0782. doi: 10.1145/3271625. URL `https://doi.org/10.1145/3271625`.

Stanislas Dehaene. *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts*. Penguin Publishing Group, 2014. ISBN 9780698151406. URL `https://books.google.fr/books?id=CWw2AAAAQBAJ`.

René Descartes, Valentine Rodger Miller, and Reese P. Miller. *Principles of philosophy*. Dordrecht, Holland ; Boston, U.S.A. : Reidel ; Hingham, Mass. : Distributed by Kluwer Boston, 1983. ISBN 9027714517. Translation of: Principia philosophiae. 1644.

René Descartes. *The Philosophical Writings of Descartes*, volume 1. Cambridge University Press, 1985. doi: 10.1017/CBO9780511805042.

Mingzhou Ding, Yonghong Chen, and Steven Bressler. Granger causality: Basic theory and application to neuroscience. *Handbook of Time Series Analysis*, 2006. doi: 10.1002/9783527609970.ch17.

Isabelle Drouet. *Causalité et probabilités : réseaux bayésiens, propensionnisme*. Theses, Université Panthéon-Sorbonne - Paris I, December 2007. URL `https://tel.archives-ouvertes.fr/tel-00265287`.

Ellery Eells. *Probabilistic Causality*. Cambridge Studies in Probability, Induction and Decision Theory. Cambridge University Press, 1991. doi: 10.1017/CBO9780511570667.

Michael Eichler. Causal inference from time series: What can be learned from granger causality? *Proceedings from the 13th International Congress of Logic, Methodology and Philosophy of Science*, 2008.

Doris Entner and Patrik Hoyer. On causal discovery from time series data using fci. *Proceedings of the 5th European Workshop on Probabilistic Graphical Models, PGM 2010*, 2010.

K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song. Robust physical-world attacks on deep learning visual classification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1625–1634, 2018.

David Freedman. From association to causation: some remarks on the history of statistics. *Statist. Sci.*, 14(3):243–258, 08 1999. doi: 10.1214/ss/1009212409. URL `https://doi.org/10.1214/ss/1009212409`.

Stefan Frenzel and Bernd Pompe. Partial mutual information for coupling analysis of multivariate time series. *Physical review letters*, 99:204101, 2007.

Andreas Galka, Tohru Ozaki, Jorge Bosch Bayard, and Okito Yamashita. Whitening as a tool for estimating mutual information in spatiotemporal data sets. *Journal of Statistical Physics*, 124(5):1275–1315, 2006.

Tobias Gerstenberg, Noah D Goodman, David Lagnado, and Joshua B Tenenbaum. From counterfactual simulation to causal judgment. 01 2014. doi: 10.13140/2.1.3144.2887.

John Geweke. Measurement of linear dependence and feedback between multiple time series. *Journal of the American Statistical Association*, 77 (378):304–313, 1982. ISSN 01621459. URL http://www.jstor.org/stable/2287238.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10:524, 2019. ISSN 1664-8021. doi: 10.3389/fgene.2019.00524. URL https://www.frontiersin.org/article/10.3389/fgene.2019.00524.

Irving J. Good. A causal calculus (i). *The British Journal for the Philosophy of Science*, 11(44):305–318, 1961. doi: 10.1093/bjps/XI.44.305.

Clive Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–38, 1969.

Clive Granger. Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and Control*, 2:329–352, 1980. doi: 10.1016/0165-1889(80)90069-X.

Clive Granger. Some recent development in a concept of causality. *Journal of Econometrics*, 39(1-2):199–211, 1988. URL https://EconPapers.repec.org/RePEc:eee:econom:v:39:y:1988:i:1-2:p:199-211.

Clive W. J. Granger. Time series analysis, cointegration, and applications. *The American Economic Review*, 94(3):421–425, 2004. ISSN 00028282.

László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer series in statistics. Springer, 2002.

Joseph Y. Halpern. *Actual Causality*. The MIT Press, 2016. ISBN 0262035022, 9780262035026.

Yuval Noah Harari. *Sapiens: A Brief History of Humankind*. Harper, 2015. ISBN 9780062316103. URL https://books.google.fr/books?id=FmyBAwAAQBAJ.

Craig Hiemstra and Jonathan D. Jones. Testing for linear and nonlinear granger causality in the stock price-volume relation. *The Journal of Finance*, 49(5):1639–1664, 1994. doi: 10.1111/j.1540-6261.1994.tb04776.x.

Christopher Hitchcock. Probabilistic Causation. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2018 edition, 2018.

Thomas Hobbes. *Elements of philosophy: the first section, concerning body / written in Latine by Thomas Hobbes of Malmesbury ; and now translated into English ; to which are added Six lessons to the professors of mathematicks of the Institution of Sr. Henry Savile, in the University of Oxford*. R. and W. Leybourn for Andrew Crooke, 1656.

Carl Hoefer. Causal determinism. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2016 edition, 2016.

Patrik O. Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 689–696. Curran Associates, Inc., 2009. URL `http://papers.nips.cc/paper/3548-nonlinear-causal-discovery-with-additive-noise-models.pdf`.

David Hume. *A Treatise of Human Nature*. Oxford University Press, 1738.

David Hume. *An Enquiry Concerning Human Understanding*. Oxford University Press, 1748.

Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O. Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *J. Mach. Learn. Res.*, 11:1709–1731, 2010. ISSN 1532-4435.

I. Bernard Cohen Isaac Newton, Anne Whitman, and Julia Budenz. *The Principia: Mathematical Principles of Natural Philosophy*. University of California Press, 1 edition, 1999. ISBN 9780520088160. URL `http://www.jstor.org/stable/10.1525/j.ctt9qh28z`.

Vladimir Ivancevic and Tijana Ivancevic. *Computational Mind: A Complex Dynamics Perspective*, volume 60. 01 2007. doi: 10.1007/978-3-540-71561-0.

Diviyan Kalainathan, Olivier Goudet, Isabelle Guyon, David Lopez-Paz, and Michèle Sebag. Sam: Structural agnostic model, causal discovery and penalized adversarial learning. *arXiv*, 2018.

Markus Kalisch and Peter Bühlmann. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8:613–636, 2007. ISSN 1532-4435.

Immanuel Kant. *Kant: Prolegomena to Any Future Metaphysics: With Selections from the Critique of Pure Reason*. Cambridge Texts in the History of Philosophy. Cambridge University Press, 1997. doi: 10.1017/CBO9781139164061.

Harri Kiiveri and T. P. Speed. Structural analysis of multivariate data: A review. *Sociological Methodology*, 13:209–289, 1982. ISSN 00811750, 14679531. URL `http://www.jstor.org/stable/270722`.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 69 6 Pt 2:066138, 2004.

Pierre Simon Laplace. *Essai philosophique sur les probabilités;*. Mme. Ve. Courcier, 1814.

David Lewis. Causation. *Journal of Philosophy*, 70(17):556–567, 1973.

David Lewis. Postscripts to 'causation'. In David Lewis, editor, *Philosophical Papers Vol. Ii*. Oxford University Press, 1986.

David Lewis. Causation as influence. *Journal of Philosophy*, 97(4):182–197, 2000.

David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Schölkopf, and Léon Bottou. Discovering causal signals in images. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 58–66, 2017. doi: 10.1109/CVPR.2017.14.

Ardon Lyon. Causality. *The British Journal for the Philosophy of Science*, 18 (1):1–20, 1967. doi: 10.1093/bjps/18.1.1. URL `http://dx.doi.org/10.1093/bjps/18.1.1`.

John Leslie Mackie. *The Cement of the Universe: A Study of Causation*. Clarendon Press, 1980.

Daniel Malinsky and David Danks. Causal discovery algorithms: A practical guide. *Philosophy Compass*, 13(1), 2018.

Daniel Malinsky and Peter Spirtes. Causal structure learning from multivariate time series in settings with unmeasured confounding. In *Proceedings of 2018 ACM SIGKDD Workshop on Causal Disocvery*, volume 92 of *Proceedings of Machine Learning Research*, pages 23–47, London, UK, 2018. PMLR.

Christopher Meek. Causal inference and causal explanation with background knowledge. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, page 403–410, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. ISBN 1558603859.

Stijn Meganck, Philippe Leray, and Bernard Manderick. Causal graphical models with latent variables: Learning and inference. In Khaled Mellouli, editor, *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pages 5–16, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. ISBN 978-3-540-75256-1.

Peter Menzies and Huw Price. Causation as a Secondary Quality. *The British Journal for the Philosophy of Science*, 44(2):187–203, 06 1993. ISSN 0007-0882. doi: 10.1093/bjps/44.2.187. URL `https://doi.org/10.1093/bjps/44.2.187`.

John Stuart Mill. *A System of Logic, Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation*. Number vol. 1 in A System of Logic, Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation. John W. Parker, 1843.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1 – 38, 2019. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2018.07.007. URL `http://www.sciencedirect.com/science/article/pii/S0004370218305988`.

Alessio Moneta, Doris Entner, Patrik O. Hoyer, and Alex Coad. Causal inference by independent component analysis: Theory and applications. *Oxford Bulletin of Economics and Statistics*, 75(5):705–730, 2013.

Joris Mooij, Dominik Janzing, Jonas Peters, and Bernhard Schölkopf. Regression by dependence minimization and its application to causal inference in additive noise models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 745–752, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-516-1. doi: 10.1145/1553374.1553470.

Joris M. Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: Methods and benchmarks. *Journal of Machine Learning Research*, 17(1):1103–1204, 2016.

Stephen Mumford and Rani Lill Anjum. *Causation: A Very Short Introduction*. Oxford University Press, 2013.

Meike Nauta, Doina Bucur, and Christin Seifert. Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction*, 1(1):312–340, 2019. ISSN 2504-4990. doi: 10.3390/make1010019.

Vivian Nutton. G. E. R. Lloyd, magic, reason and experience. studies in the origins and development of greek science. *Medical History*, 24(4):477–477, 1980. doi: 10.1017/S0025727300040680.

Christina Papagiannopoulou, Diego G. Miralles, Stijn Decubber, Matthias Demuzere, Niko E. C. Verhoest, Wouter A. Dorigo, and Willem Waegeman. A non-linear granger-causality framework to investigate climate-vegetation dynamics. *Geoscientific Model Development*, 10(5):1945–1960, 2017. doi: 10.5194/gmd-10-1945-2017.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988. ISBN 1558604790.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, NY, USA, 2000. ISBN 0-521-77362-8.

Judea Pearl. The causal foundations of structural equation modeling. *Handbook of Structural Equation Modeling*, 12 2010.

Judea Pearl and Elias Bareinboim. Transportability of causal and statistical relations: A formal approach. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, AAAI'11, page 247–254. AAAI Press, 2011.

Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, 2018.

Karl Pearson. *The Grammar of Science*. A. and C. Black, 1900. URL `https://books.google.fr/books?id=kXoKAAAAIAAJ`.

Karl Pearson. *The Grammar of Science*. Number vol. 1 in The Grammar of Science. A.and C. Black, 1911. URL https://books.google.fr/books?id=5nsuAAAAIAAJ.

Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. Identifiability of causal graphs using functional models. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, page 589–598, Arlington, Virginia, USA, 2011. AUAI Press. ISBN 9780974903972.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Causal inference on time series using restricted structural equation models. In *Advances in Neural Information Processing Systems 26*, pages 154–162, 2013.

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. doi: https://doi.org/10.1111/rssb.12167. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12167.

Jonas Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, USA, 2017.

Arthur C. Pigou. Alcoholism and Heredity. *International Journal of Epidemiology*, 12 2017. ISSN 0300-5771. doi: 10.1093/ije/dyw340. dyw340.

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, page 401–408, Arlington, Virginia, USA, 2006. AUAI Press. ISBN 0974903922.

Hans Reichenbach. *The Direction of Time*. Dover Publications, 1956.

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, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939778. URL https://doi.org/10.1145/2939672.2939778.

Thomas Richardson. A discovery algorithm for directed cyclic graphs. In *Proceedings of the Twelfth International Conference on Uncertainty in Artificial Intelligence*, UAI'96, pages 454–461, San Francisco, CA, USA, 1996. Morgan Kaufmann Publishers Inc. ISBN 1-55860-412-X. URL http://dl.acm.org/citation.cfm?id=2074284.2074338.

Thomas Richardson and Peter Spirtes. Ancestral graph markov models. *Ann. Statist.*, 30(4):962–1030, 08 2002.

James M. Robins, Richard Scheines, Peter Spirtes, and Larry Wasserman. Uniform consistency in causal inference. *Biometrika*, 90(3):491–515, 2003. doi: 10.1093/biomet/90.3.491. URL http://dx.doi.org/10.1093/biomet/90.3.491.

Kenneth J. Rothman. Causes. *American Journal of Epidemiology*, 104(6):587–592, 12 1976. ISSN 0002-9262. doi: 10.1093/oxfordjournals.aje.a112335. URL https://doi.org/10.1093/oxfordjournals.aje.a112335.

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

Donald B. Rubin. [on the application of probability theory to agricultural experiments. essay on principles. section 9.] comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, 5(4):472–480, November 1990.

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, Playa Blanca, Lanzarote, Canary Islands, 09–11 Apr 2018. PMLR.

Jakob Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In Jonas Peters and David Sontag, editors, *Proceedings of Machine Learning Research*, volume 124, pages 1388–1397. PMLR, 2020. URL http://proceedings.mlr.press/v124/runge20a.html.

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.

Bertrand Russell. On the notion of cause. *Proceedings of the Aristotelian Society*, 7:1–26, 1912.

Robert Rynasiewicz. Newton's views on space, time, and motion. In Edward N. Zalta, editor, *Stanford Encyclopedia of Philosophy*, pages 8–12. The Metaphysics Research Lab, 2008.

Wesley C. Salmon. *Scientific Explanation and the Causal Structure of the World*. Princeton University Press, 1984.

Ruben Sanchez-Romero, Joseph D. Ramsey, Kun Zhang, Madelyn R. K. Glymour, Biwei Huang, and Clark Glymour. Estimating feedforward and feedback effective connections from fmri time series: Assessments of statistical methods. *Network Neuroscience*, 3(2):274–306, 2019. doi: 10.1162/netn_a_00061.

Jonathan Schaffer. Trumping preemption. *Journal of Philosophy*, 97(4):165, 2000. doi: 10.2307/2678388.

Arthur Schopenhauer. *Two Essays by Arthur Schopenhauer: I. on the Fourfold Root of the Principle of Sufficient Reason, II. on the Will in Nature: a Literal Translation*. London : G. Bell, 1889.

Thomas Schreiber. Measuring information transfer. *Physical review letters*, 85:461–4, 2000.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.

Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006. ISSN 1532-4435. URL `http://dl.acm.org/citation.cfm?id=1248547.1248619`.

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O. Hoyer, and Kenneth Bollen. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research*, 12:1225–1248, 2011. ISSN 1532-4435. URL `http://dl.acm.org/citation.cfm?id=1953048.2021040`.

Stephen M. Smith, Karla L. Miller, Gholamreza Salimi Khorshidi, Matthew A. Webster, Christian F. Beckmann, Thomas E. Nichols, Joseph

Ramsey, and Mark W. Woolrich. Network modelling methods for fmri. *NeuroImage*, 54:875–891, 2011.

J. Snow. *On the Mode of Communication of Cholera*. John Churchill, 1855. URL https://books.google.fr/books?id=-N0_AAAAcAAJ.

Baruch Spinoza. The ethics. 1677. doi: 10.2307/2215962.

Peter Spirtes and Kun Zhang. Causal discovery and inference: concepts and recent methodological advances. *Applied Informatics*, 3(1):3, 2016. ISSN 2196-0089. doi: 10.1186/s40535-016-0018-x.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT press, 1st edition, 1990.

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

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. ISSN 08834237.

S. Spreeuwenberg, P. Henao, and K. Hiroi. *AIX: Artificial Intelligence Needs EXplanation: Why and how Transparency Increases the Success of AI Solutions*. CB, 2019. ISBN 9789081556842. URL https://books.google.fr/books?id=KeMhzAEACAAJ.

Galen Strawson. *The Secret Connexion: Causation, Realism, and David Hume: Revised Edition*. Oxford University Press UK, 2014.

Jie Sun, Dane Taylor, and Erik Bollt. Causal network inference by optimal causation entropy. *SIAM Journal on Applied Dynamical Systems*, 14(1): 73–106, 2015. doi: 10.1137/140956166.

Patrick Suppes. *A Probabilistic Theory of Causality*. Amsterdam: North-Holland Pub. Co., 1970.

C.C.W. Taylor. Leucippus (5th century bc). 1998. doi: 10.4324/9780415249126-A064-1. URL https://www.rep.routledge.com/articles/biographical/leucippus-5th-century-bc/v-1.

Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, UAI '90, pages 255–270, New York, NY, USA, 1991. Elsevier Science Inc. ISBN 0-444-89264-8.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 1096–1103, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-205-4. doi: 10.1145/1390156.1390294. URL http://doi.acm.org/10.1145/1390156.1390294.

J. von Kügelgen, L. Gresele, and B. Schölkopf. Simpson's paradox in covid-19 case fatality rates: a mediation analysis of age-related causal effects, 2020.

Abraham Wald and Center for Naval Analyses (U.S.). *A reprint of "A method of estimating plane vulnerability based on damage of survivors" by Abraham Wald*. Alexandria, Va. : Center for Naval Analyses, 1980.

James Woodward. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, 2003.

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

George U. Yule. Karl Pearson, 1857-1936. *Obituary Notices of Fellows of the Royal Society*, 2(5):72–110, 1936. ISSN 1479-571X. doi: 10.1098/rsbm.1936.0007. URL http://rsbm.royalsocietypublishing.org/content/2/5/72.

Aleš Završnik. Algorithmic justice: Algorithms and big data in criminal justice settings. *European Journal of Criminology*, 0(0):1477370819876762, 2019. doi: 10.1177/1477370819876762.

Zhalama, J. Zhang, and W. Mayer. Weakening faithfulness: some heuristic causal discovery algorithms. *International Journal of Data Science and Analytics*, 3:93–104, 2016.

David D. Zhang, Harry F. Lee, Cong Wang, Baosheng Li, Qing Pei, Jane Zhang, and Yulun An. The causality analysis of climate change and large-scale human crisis. *Proceedings of the National Academy of Sciences*, 108(42):17296–17301, 2011. ISSN 0027-8424. doi: 10.1073/pnas.1104268108.

Jiji Zhang. A characterization of markov equivalence classes for directed acyclic graphs with latent variables. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, UAI'07, pages 450–457, Arlington, Virginia, USA, 2007. AUAI Press. ISBN 0974903930.

Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16):1873 – 1896, 2008a. ISSN 0004-3702.

Jiji Zhang. Causal reasoning with ancestral graphs. *J. Mach. Learn. Res.*, 9: 1437–1474, June 2008b. ISSN 1532-4435.

Jiji Zhang and Peter Spirtes. Strong faithfulness and uniform consistency in causal inference. In *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*, UAI'03, pages 632–639, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc. ISBN 0127056645.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006. doi: 10.1198/016214506000000735.