Journal of Artificial Intelligence Research 73 (2022) 767-819 Submitted 10/2021; published 02/2022

Survey and Evaluation of Causal Discovery Methods for Time Series

Charles K. Assaad KASSAAD@EASYVISTA.COM
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG,
EasyVista,
38000 Grenoble, France

Emilie Devijver EMILIE.DEVIJVER@UNIV-GRENOBLE-ALPES.FR
Eric Gaussier ERIC.GAUSSIER@IMAG.FR
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG,
38000 Grenoble, France

Abstract

We introduce in this survey the major concepts, models, and algorithms proposed so far to infer causal relations from observational time series, a task usually referred to as *causal discovery in time series*. To do so, after a description of the underlying concepts and modelling assumptions, we present different methods according to the family of approaches they belong to: Granger causality, constraint-based approaches, noise-based approaches, score-based approaches, logic-based approaches, topology-based approaches, and difference-based approaches. We then evaluate several representative methods to illustrate the behaviour of different families of approaches. This illustration is conducted on both artificial and real datasets, with different characteristics. The main conclusions one can draw from this survey is that causal discovery in times series is an active research field in which new methods (in every family of approaches) are regularly proposed, and that no family or method stands out in all situations. Indeed, they all rely on assumptions that may or may not be appropriate for a particular dataset.

1. Introduction

Causality plays a central role in science and has been the subject of many debates among philosophers, biologists, mathematicians and physicists, to name but a few. Causality is implicit in the logic and structure of ordinary language and is embedded in our understanding mechanism that pushes humans to invoke questions. Why is it dark? Why is the sea salty? What is the effect of exercise on heart rate, of a vaccine on a particular disease? What is the effect of industrial pollution on the environment? 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 observations into causal knowledge (Spirtes et al., 2001). Causality is indeed crucial for explanatory purposes since an effect can be explained by its causes, regardless of the correlations it may have with other variables.

The recent decades have seen the development, from philosophers, mathematicians, and computer scientists, of different models and methods to infer causal relations from data and to reason on the basis of these relations (to, *e.g.*, predict the effect of changing a particular medication). If the first studies were dedicated to non temporal data, more and more studies now focus on time series. Indeed, 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). The number of scientific publications dedicated to causality in time series as well as the number of tools developed in this context have steadily increased to a point that it is difficult for non specialists to grasp the most important approaches proposed so far.

The goal of this survey is twofold:

©2022 AI Access Foundation. All rights reserved.

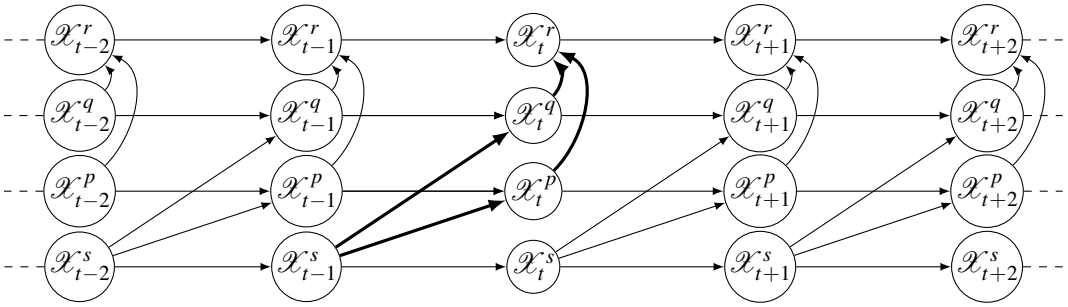


Figure 1: *Running example*: a diamond structure with self causes.

- On the one hand, we want to introduce the major concepts, models, methods, and associated algorithms proposed so far to infer causal relations from observational time series[1], a task usually referred to as *causal discovery*;

- on the other hand, we want to assess how different methods for causal discovery in time series behave in practice.

Several surveys on causal discovery have recently been proposed (Guo, Cheng, Li, Hahn, & Liu, 2020; Nogueira, Gama, & Ferreira, 2021; Glymour, Zhang, & Spirtes, 2019; Vowels, Camgöz, & Bowden, 2021). However, most of them do not discuss time series and when they do, they focus on Granger causality. In contrast, the current survey is dedicated to causal discovery in time series and reviews all families of approaches proposed in this area.

The remainder of this survey is organized as follows. After a description of the underlying concepts and modelling assumptions in Section 2, we present different methods according to the family of approaches they belong to, using the same example, a diamond structure given in Figure 1, for illustration purposes[2]: Granger causality (one of the first approaches proposed) in Section 3, constraint-based approaches (one of the most popular approaches) in Section 4, noise-based approaches (another popular approach) in Section 5, and score-based approaches (with a long history related to Bayesian networks) in Section 6. The main characteristics of representative methods of the above families are summarized in Section 7. We then turn in Section 8 to other approaches (logic-based, topology-based, and difference-based approaches) which differ from the previous ones on several aspects. The behaviour of representative methods of different families[3] is then illustrated in Section 9. This illustration, conducting on both artificial and real datasets, with different characteristics, shows that different families, and within a given family different methods, are adapted to different situations so that there isn't a single family or method that outperforms all the others in all situations. Lastly, Section 10 concludes the survey.

2. Background

We first introduce in this section the basic notions underlying causal inference approaches prior to present the different causal graphs used to represent causal relations within and between time series. The main notations that will be used throughout this survey are summarized in Table 1. The variables first considered here are standard random variables (the extension to time instants of time series is direct).

1. In observational time series, the value of a variable is always determined by its causes, hence it is never set through an intervention. Interventional data are not considered in this survey.
2. Note that, for simplicity, we consider that the method illustrated *works ideally*.
3. All these methods can be used through a Python routine available at `https://github.com/ckassaad/causal_discovery_for_time_series`.

Notation	Description
X^c, X^p, X^q, X^r, ξ	random variables and noise
X^R	set of random variables
d, N, T	number of time series, of observations and of time points
$\mathscr{X}$, $\mathscr{X}^p$, $\mathscr{X}_t^p$	multivariate time series $\{\mathscr{X}^1, \cdots, \mathscr{X}^d\}$, pth time series $\mathscr{X}^p$, $\mathscr{X}^p$ at time t
$\mathscr{X}^{\mathbf{R}}$, $\mathscr{X}_t^{\mathbf{R}}$, $\mathscr{X}_{\mathbf{T}}^{\mathbf{R}}$	subset of time series from $\mathscr{X}$, at time t, at all time points in $\mathbf{T}$
$\mathscr{L}, \mathscr{S}$	latent variable, hidden selection variable
$\perp\!\!\!\perp, \not\perp\!\!\!\perp$	independent, not independent
$Pr(X = x), \mathbb{E}(X)$	probability of $X = x$ and expectation of X
$I(X^p; X^q)$	mutual information between X^p and X^q
$\mathscr{G}$	causal graph
$\text{Adj}(X^p, \mathscr{G})$, $\text{Adj}(X^{\mathbf{R}}, \mathscr{G})$	variables adjacent to X^p in $\mathscr{G}$, or adjacent to the set $X^{\mathbf{R}}$
$\text{Par}(X^p, \mathscr{G})$	set of parents (set of causes) of X^p in $\mathscr{G}$
$X - Y$	X is a neighbor of Y
$\text{Hom}(X_{t-i}^p, X_t^q, \mathscr{G})$	set of vertex pairs $(X_{k-i}^p, X_k^q)_k$ homologous to (X_{t-i}^p, X_t^q)
$X \to Y$	X is a cause of Y and Y is an effect of X
$X \not\to Y$	X is not a cause of Y
$X \leftrightarrow Y$	X and Y have a common hidden confounder
$Sepset(X^p, X^q)$	seperation set of X^p and X^q
$Dsepset(X^p, X^q)$	d-seperation set of X^p and X^q

Table 1: Main notations used throughout this survey.

2.1 Preliminaries on Causal Inference

One of the goals of causal inference is to build a causal graph from observational data. As we will see, the relations between a probability distribution and its representation as a graph are central to this construction. It is however not always possible to infer a causal graph solely from observational data on which one can only compute correlations and statistical independencies[4]. One needs additional assumptions to do so and the goal of this section is to present the main assumptions and principles behind causal discovery approaches.

Let us first consider the two basic causal structures given in Figure 2. The structure on the left corresponds to a *confounder*, *i.e.* a variable that is a *common cause* of two other variables. The figure on the right represents a *collider*, *i.e.* a variable that is caused by two unrelated variables. If the common cause X^p in the confounder structure was not observed, one would infer a spurious correlation and a causal relation between X^q and X^r as these latter variables are independent only when conditioned on X^p. One way to avoid such spurious correlations is to assume that all common causes are measured.

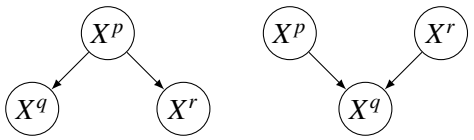


Figure 2: Two basic structures: a confounder (left) and a collider (right)

Definition 1 (Causal Sufficiency, Spirtes et al., 2001) *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.

4. *Correlation is not causation*, as the saying goes, meaning that a correlation does not necessarily correspond to a causal relation.

Under the assumption of causal sufficiency, most causal discovery algorithms assume that the causal structure can be represented by a *Directed Acyclic Graph* (DAG)[5] in which a directed edge represents a relation from a cause to its effect. The absence of an edge between two variables means that the variables are (conditionally) independent. In essence, a DAG represents a factorization of the probability distribution over the variables in which the probability of a variable is conditioned by its parents. For example, the joint probability distribution over X^p, X^q, X^r associated to Figure 2 (left) can be factorized as: $P(X^p, X^q, X^r) = P(X^p)P(X^q|X^p)P(X^r|X^p)$. Whenever a probability distribution can be factorized according to a given DAG, we say that the DAG and the probability distribution are *compatible*.

There is a strong connection, for compatible graphs and probability distributions, between the (conditional) independence/dependence of two variables and the topology of the graph. This connection is based on the concept of d-separation, first introduced in the context of Bayesian networks.

Definition 2 (d-separation, Pearl, 1988) *If $\mathcal{G}$ is a DAG in which X^p and X^q are two vertices and X^R is a set 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 X^p and 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.*

There are two important probabilistic implications of d-separation in a DAG $\mathcal{G}$ (Spirtes et al., 2001):

- If X^R d-separates X^p and X^q, then X^p and X^q are independent given X^R ($X^p \perp\!\!\!\perp X^q \mid X^R$) in all probability distributions compatible with $\mathcal{G}$;
- If X^R d-connects X^p and X^q, then X^p and X^q are dependent given X^R in *almost all* probability distributions compatible with $\mathcal{G}$.

The reverse implication is not true on all compatible distributions as one can tune the parameters to generate independencies along an unblocked path, such a tuning being however unlikely to occur in practice.

Causal inference from observational data consists in first determining all (conditional) independence and dependence relations between variables and then in constructing a graph compatible with these relations. The following theorem states a necessary and sufficient condition for a DAG and a probability distribution to be compatible.

Theorem 1 (Markov Condition, Pearl, 2000) *A necessary and sufficient condition for a probability distribution to be compatible with a DAG $\mathcal{G}$ is that every variable be independent of all its nondescendants (in $\mathcal{G}$), conditional on its parents.*

When the DAG is interpreted causally, then the parents of a variable correspond to its direct causes and one speaks in that case of the *Causal Markov Condition* (Spirtes et al., 2001). As standard in causal discovery studies, we place ourselves in this latter case.

As we will discuss in more detail in Section 4, several DAGs can represent the same set of conditional independencies and be compatible with the same probability distribution. This limits the possibility to infer a causal graph from probabilities alone. Two main additional conditions have thus been introduced so as to restrict the graphs considered from a given probability distribution. The first one is the *minimality condition*, which requires that the graph does not contain dependencies not present in the observational data.

Definition 3 (Minimality Condition, Pearl, 2000) *A DAG $\mathcal{G}$ compatible with a probability distribution P is said to satisfy the* minimality condition *if P is not compatible with any proper subgraph of $\mathcal{G}$.*

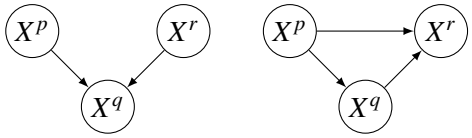


Figure 3: Faithful (left) vs unfaithful (right) graphs.

The minimality condition is however not sufficient to restrict the set of 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 3 are compatible with the probability distribution given above as $P(X^p,X^q,X^r)$ factorizes in both cases as $P(X^p)P(X^r)P(X^q|X^p,X^r)$. 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 on the right states that X^p and X^r are unconditionally dependent whereas the probability distribution states that they are unconditionally independent. This is problematic and we say in such a case that the graph is *unfaithful* according to the following definition.

Definition 4 (Faithfulness, Spirtes et al., 2001) *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 $\mathcal{G}$.*

Note that the minimality condition is weaker than faithfulness in the sense that faithfulness and Markov conditions together entail minimality, whereas both minimality and Markov conditions do not always entail faithfulness. 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. It is the condition usually adopted in works aiming at inferring causal structures from observational data. If this condition does not hold, then, in the absence of any other assumption, there is no guarantee that the inferred graph is close to the true causal graph.

To conclude this presentation of the basic notions behind causal inference, we consider the case where some causes or effects may be unobserved (in which case the causal sufficiency condition is not satisfied). Figure 4 (left) provides such an example in which the hidden cause L is a hidden confounder (in that case the common cause of X^q and X^r) and the hidden effect S, also called *selection bias* variable, induces a dependence between X^p and X^r. Indeed, as a selection variable is a variable on which all observations are conditioned on, conditioning on S implies conditioning on the collider X^q, which creates a correlation between its causes X^p and L; as L is a cause of X^r, any variable correlated to L is also correlated to X^r.

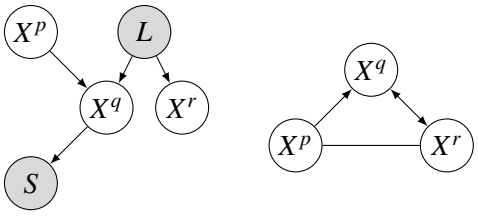


Figure 4: Illustration of hidden confounder (L) and hidden effect (S): on the left when observing all the variables, and on the right the corresponding MAG representation.

L and S being unobserved, they cannot be represented in the final graph. However, one can still try to assess their presence and adapt standard representations. This adaptation goes through *Maximal Ancestral Graphs (MAGS)* (Richardson & Spirtes, 2002) which represent the presence of hidden confounders and selection bias variables through different types of edges: bi-directed edges

5. Note that DAGs can be generalized to *Directed Graphs*, referred to in the following as DiGraphs, where cycles are allowed.

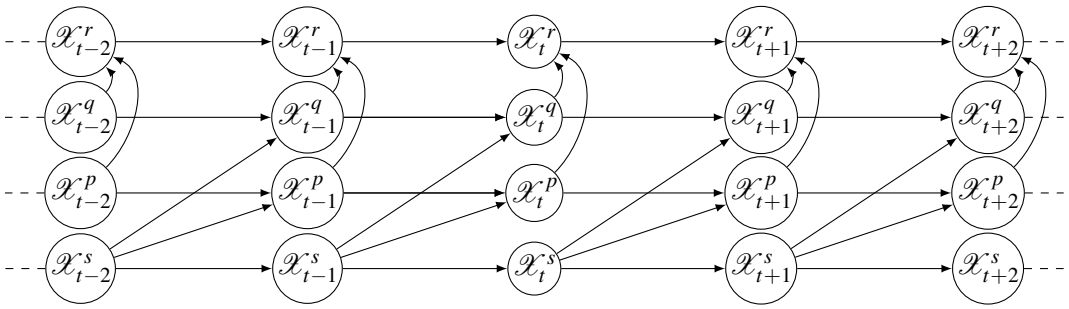
($\longleftrightarrow$) in the graph represent the existence of a hidden confounder whereas undirected edges ($-$) represent unobserved selection bias variables that have been conditioned on rather than marginalized over. MAGs[6] are maximal in the sense that no additional edge may be added to the graph without changing the independence model (Richardson & Spirtes, 2002). The notions introduced before can readily be extended to MAGs.

Lastly, when the variables considered are temporal, then one can rely on the *temporal priority* concept that goes back to Hume (1738) and is described by Rankin and McCormack (2013). In a nutshell, it simply states that a cause precedes its effects.

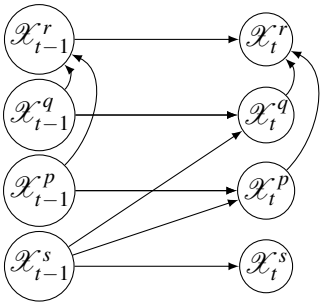
Definition 5 (Temporal Priority) *A causal relation between two variables is said to satisfy the* temporal priority *if it is oriented in such a way that the cause occurred before its effect.*

Temporal priority makes the process of causality asymmetric in time and is useful for orienting a causal relation when one knows that two variables are causally related. That said, the difference in time between two events associated to 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.

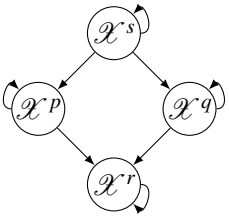
2.2 Causal Graphs for Time Series



Full Time Causal Graph (a)



Window Causal Graph (b)



Summary Causal Graph (c)

Figure 5: Different causal graphs that one can infer from three time series: full time causal graph (5a), window causal graph (5b) and summary causal graph (5c). Note that the first one gives more information but cannot be inferred in practice, the second one is a schematic viewpoint of the full behavior, whereas the last one give an overview and can be deduced from the window causal graph.

Causal discovery in time series aims at discovering, from observational data, causal relations within and between d-variate time series $\mathcal{X}$ where, for a fixed t, each $\mathcal{X}_t$ is a vector $(\mathcal{X}_t^1, \cdots, \mathcal{X}_t^d)$ in which each variable $\mathcal{X}_t^p$ represents a measurement of the p-th time series at time t. There are

6. Similarly to DiGraphs, one can extend MAGs to take into account cycles and self loops.

three ways to represent time series through a causal graph $\mathcal{G} = (V, E)$ with V the set of vertices and E the set of edges. The first is called a *full time causal graph* (also called *infinite dynamic causal graph* by Malinsky & Spirtes, 2018) and represents a complete graph of the dynamic system, as illustrated in Figure 5a.

Definition 6 (Full Time Causal Graph) *Let $\mathcal{X}$ 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 $\mathcal{X}^1, \ldots, \mathcal{X}^d$ at each time $t \in \mathbb{Z}$. The edges E of the graph are defined as follows: variables $\mathcal{X}_{t-i}^p$ and $\mathcal{X}_t^q$ are connected by a lag-specific directed link $\mathcal{X}_{t-i}^p \to \mathcal{X}_t^q$ in $\mathcal{G}$ pointing forward in time if and only if $\mathcal{X}^p$ causes $\mathcal{X}^q$ at time t with a time lag of $i > 0$ for $p = q$ and with a time lag of $i \geq 0$ for $p \neq q$.*

It is usually not possible to infer general full time causal graphs as there usually is a single observation for each time series at each time instant and it is common to rely on the so-called *Consistency Throughout Time* (also referred to as Causal Stationarity by Runge, 2018) assumption.

Definition 7 (Consistency Throughout Time) *A causal graph $\mathcal{G} = (V, E)$ for a multivariate time series $\mathcal{X}$ is said to be* consistent throughout time *if all the causal relationships remain constant in direction throughout time.*

When assuming consistency throughout time, the full time causal graph can be contracted to give a finite graph which we call *window causal graph*. 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.

Definition 8 (Window Causal Graph) *Let $\mathcal{X}$ 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 $\mathcal{X}^1, \ldots, \mathcal{X}^d$ at each time $t, \ldots, t + \tau$. The edges E of the graph are defined as follows: variables $\mathcal{X}_{t-i}^p$ and $\mathcal{X}_t^q$ are connected by a lag-specific directed link $\mathcal{X}_{t-i}^p \to \mathcal{X}_t^q$ in $\mathcal{G}$ pointing forward in time if and only if $\mathcal{X}^p$ causes $\mathcal{X}^q$ at time t with a time lag of $0 \leq i \leq \tau$ for $p \neq q$ and with a time lag of $0 < i \leq \tau$ for $p = q$.*

Figure 5b illustrates a window causal graph corresponding to the full time causal graph given in Figure 5a with consistency throughout time. This graph encodes the following causal relations: $\mathcal{X}^s$ causes itself with a lag equal to 1, causes $\mathcal{X}^p$ with a lag equal to 1 and causes $\mathcal{X}^q$ with a lag equal to 1; $\mathcal{X}^p$ causes itself with a lag equal to 1, $\mathcal{X}^q$ causes itself with a lag equal to 1, and $\mathcal{X}^p$ and $\mathcal{X}^q$ cause $\mathcal{X}^r$ with a lags equal to 0; $\mathcal{X}^r$ causes itself with a lag equal to 1. Note that the full time causal graph and the window causal graph are equivalent when assuming consistency throughout time. When this assumption is not made, the only representation one can use is the full time causal graph. Lastly, the window causal graph can be summarized into a summary causal graph (see below), at the cost of losing information on the particular instants in the past at which the causes occured.

In practice, it is often sufficient to know the causal relations between time series as a whole, without knowing precisely the relations between time instants. In that case, one can further compress the causal graph in a *summary graph* (also called *unit graph* by Chu & Glymour, 2008) that represents causal relations within and between time series without any time information. An example of such a graph is given in Figure 5c. Note that since a summary causal graph is a summary of the full time causal graph, it can contain cycles.

Definition 9 (Summary Causal Graph) *Let $\mathcal{X}$ 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 time series $\mathcal{X}^1, \ldots, \mathcal{X}^d$. The edges E of the graph are defined as follows: variables $\mathcal{X}^p$ and $\mathcal{X}^q$ are connected if and only if there exists some time t and some time lag i such that $\mathcal{X}_{t-i}^p$ causes $\mathcal{X}_t^q$ at time t with a time lag of $0 \leq i$ for $p \neq q$ and with a time lag of $0 < i$ for $p = q$.*

Summary graphs are in general less sensitive to possible variations in time and errors in estimating time lags compared to full time and window causal graphs.

2.3 When is a Method *Truly Causal*?

Most methods reviewed in this survey fit within the following, general form for the functional model of any potential effect $\mathscr{X}^q$ (this model is compatible with *temporal priority* (Def. 5) and *Consistency Throughout Time* (Def. 7)):

$$\forall t,\ \mathscr{X}_t^q = f(\mathscr{C}_t^q(\mathscr{X}^{r_1}),\cdots,\mathscr{C}_t^q(\mathscr{X}^{r_q}),\xi_t^q), \tag{1}$$

where f denotes any real-valued multivariate function and ξ_t^q represents some noise independent from all the causes of $\mathscr{X}_t^q$. $\mathscr{C}^q = \{\mathscr{X}^{r_1},\cdots,\mathscr{X}^{r_q}\}$ is the set of time series which are causes of $\mathscr{X}^q$. $\mathscr{C}_t^q(\mathscr{X}^r)$, for $\mathscr{X}^r \in \mathscr{C}_q$, represents the past instants (*i.e.*, time instants before t) of $\mathscr{X}^r$ which are causes of $\mathscr{X}_t^q$. It can be written as:

$$\mathscr{C}_t^q(\mathscr{X}^r) = \{\mathscr{X}_{t-\gamma_1}^r,\cdots,\mathscr{X}_{t-\gamma_{K_r}}^r\}, \tag{2}$$

where $K_r \in \mathbb{Z}^+$ and $\gamma_1,\cdots,\gamma_{K_r}$ are integers such that $\gamma_1 > \cdots > \gamma_{K_r} \geq 0$. As past instants of a time series can (and usually do) participate to the definition of the current instant, $\mathscr{X}^q$ can of course be a cause of itself. Methods usually differ on the assumptions made on f, $\mathscr{C}_t^q()$ and the observational data. Note however that few methods, as topology-based and difference-based methods (Section 8), rely on a different modelling, based on differential equations (see also Blom et al., 2019).

Not all methods are deemed to recover true causal relations even though the distinction between those which are and those which aren't is not always clearcut. With the model above, a time series correlated with $\mathscr{X}^q$ and not in $\mathscr{C}^q$ corresponds to a spurious correlation. At the finer grained level of time instants, for a time series $\mathscr{X}^r \in \mathscr{C}^q$, any past instant of $\mathscr{X}^r$ correlated with $\mathscr{X}_t^q$ and not in $\mathscr{C}_t^q(\mathscr{X}^r)$ also corresponds to a spurious correlation. Methods aiming at discovering summary causal graphs may be interested in just the first type of correlations (between time series), whereas methods aiming at discovering window causal graphs are usually interested in both types (between time series and between time instants). In this survey, we say that a method is truly causal when it aims at distinguishing spurious correlations from causal relations, be it at the level of time series or at the finer grained level of time instants. In that case, we will use the standard vocabulary of causality and say, for example, that a variable *causes* another variable. The pairwise Granger method (Section 3), for example, is not truly causal and we say in that case that a variable *Granger-causes* another variable. This is also the case for CCM causality and PAI causality (Definitions 15 and 16, Section 8.2) which do not clearly distinguish correlations and causal relations but mainly focus on specific correlations. Multivariate and recent extensions of the Granger method however aim at distinguishing spurious correlations and causal relations and are considered truly causal. This is also the case for constraint-based, noise-based, score-based and logic-based approaches.

We now turn to the different approaches used to infer causal graphs between time series.

3. 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 causes and effects, a cause preceding its effects[7]. 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:

7. Originally, Granger causality was introduced for continuous time series. It has however been extended to temporal point processes (Kim, Putrino, Ghosh, & Brown, 2011; Casile, Faghih, & Brown, 2021).

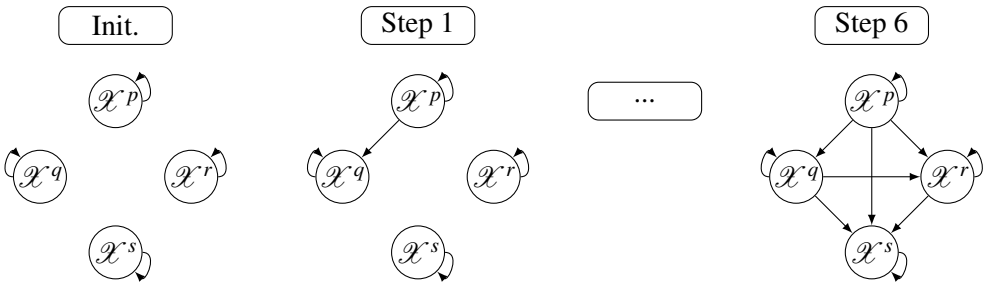


Figure 6: Running example: structure inferred by the pairwise Granger method (an arbitrary order has been chosen for the example).

Definition 10 (Granger Causality, Granger, 1980) *A time series $\mathscr{X}^p$ Granger-causes $\mathscr{X}^q$ if past values of $\mathscr{X}^p$ provide unique, statistically significant information about future values of $\mathscr{X}^q$.*

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. In addition, the temporal precedence constraints it relies on prevents one from inferring the direction of "instantaneous" relations. Indeed, modifying Granger causality by regressing $\mathscr{X}_t^q$ using the past values of $\mathscr{X}^q$ and $\mathscr{X}^p$, as well as $\mathscr{X}_t^p$ 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). Moreover, Granger Causality can be problematic in dynamic systems with weak to moderate coupling, because separability (information about causes is not contained in effects) is not always met (Granger, 1969; Sugihara, May, Ye, hao Hsieh, Deyle, Fogarty, & Munch, 2012).

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 & 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.

3.1 Standard Pairwise Granger Causality

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

$$\mathscr{X}_t^q = a_{q,0} + \sum_{i=1}^{\tau} a_{q,i} \mathscr{X}_{t-i}^q + \xi_t^q, \tag{Mres}$$

and its augmented version:

$$\mathscr{X}_t^q = a_{q,0} + \sum_{i=1}^{\tau} a_{q,i} \mathscr{X}_{t-i}^q + \sum_{i=1}^{\tau} a_{p,i} \mathscr{X}_{t-i}^p + \xi_t^q, \tag{Mfull}$$

where $(\xi_t^q)_t$ are uncorrelated random variables with zero mean and variance σ^2, $(a_{q,i})_{1\leq i\leq\tau}$ and $(a_{p,i})_{1\leq i\leq\tau}$ are real coefficients, and τ 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 $\mathscr{X}^q$ 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 $\mathscr{X}^q$ and $\mathscr{X}^p$ to predict the current value of $\mathscr{X}^q$. If the full model is significantly more accurate than the restricted model, one can conclude that $\mathscr{X}^p$ Granger-causes $\mathscr{X}^q$. 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 $\mathscr{X}^p$ does not Granger-cause $\mathscr{X}^q$. In practice, the optimal lag τ can be estimated using any information criterion, as the Akaike or Schwartz information criteria.

Algorithm 1 PWGC

Require: $\mathscr{X}$ a d-dimensional time series of length T, $\tau_{\max} \in \mathbb{N}$ the maximum number of lags
  Form an empty graph $\mathscr{G}$ with d nodes V
  Standardize the data and check if it is covariance stationary
  Find the optimal lag value $\tau \in \{1, \ldots, \tau_{\max}\}$
  for $\mathscr{X}^q \in V$ **do**
    Fit Mres: $(\mathscr{X}^q_{t-i})_{1 \leq i \leq \tau} \mapsto \mathscr{X}^q_t$ and compute its residuals
    for $\mathscr{X}^p \in V \setminus \{\mathscr{X}^q\}$ **do**
      Fit Mfull: $(\mathscr{X}^p_{t-i}, \mathscr{X}^q_{t-i})_{1 \leq i \leq \tau} \mapsto \mathscr{X}^q_t$ and compute its residuals
      z = test to compare (Mfull) and (Mres)
      if $z < \alpha$ **then** add edge $\mathscr{X}^p \to \mathscr{X}^q$ to $\mathscr{G}$
  Return the Summary DiGraph $\mathscr{G}$

Figure 6 illustrates the behaviour of this method which infers a summary causal graph. Starting from an empty graph (with self causes), all relations between pairs of variables are iteratively tested.

In a multivariate setting, a pairwise analysis can be performed using the bivariate approach summarized in Algorithm 1. 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 Multivariate Granger Causality

To overcome the problem of common confounders, all relevant information needs to be included in the analysis. Let $\mathscr{X} = (\mathscr{X}^1, \mathscr{X}^2, \cdots, \mathscr{X}^d)$ 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:

$$\mathscr{X}^q_t = a_{q,0} + \sum_{\substack{r=1 \\ r \neq p}}^{d} \sum_{i=1}^{\tau} a_{r,i} \mathscr{X}^p_{t-i} + \xi^q_t, \qquad \text{(mvMres)}$$

$$\mathscr{X}^q_t = a_{q,0} + \sum_{r=1}^{d} \sum_{i=1}^{\tau} a_{r,i} \mathscr{X}^r_{t-i} + \xi^q_t, \qquad \text{(mvMfull)}$$

where $(\xi^q_t)_t$ are uncorrelated random variables with zero mean and variance σ^2, $a_{q,0}$ and $(a_{r,i})_{1 \leq r \leq d, 1 \leq i \leq \tau}$ are real coefficients, and τ 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 $\mathscr{X}^p$. Analogously to the bivariate case, and as shown in Algorithm 2, if the full model is significantly more accurate than the restricted model (through a statistical test), one concludes that $\mathscr{X}^p$ Granger-causes $\mathscr{X}^q$. This version is sound and usually yields better results; however, its computation overload is such that in practice many studies rely on the pairwise version.

3.3 Extensions

In its original version, Granger causality cannot deal with non-stationary processes. A linear regression learning process with weighted distribution shifts, called linear WDS, was recently introduced by Luo et al. (2015) to overcome this problem. In linear WDS, distribution shifts are detected by analyzing the mean and standard deviation of the preceding points: a distribution shift is identified

Algorithm 2 MVGC

Require: $\mathscr{X}$ a d-dimensional time series of length T, $\tau_{\max} \in \mathbb{N}$ the maximum number of lags
  Form an empty graph $\mathscr{G}$ with d nodes V
  Standardize the data and check if it is covariance stationary
  Find the optimal lag value $\tau \in \{1, \dots, \tau_{\max}\}$
  for $\mathscr{X}^q \in V$ **do**
    Fit mvMfull: $(\mathscr{X}_{t-i})_{1 \leq i \leq \tau} \mapsto \mathscr{X}_t^q$ and compute its residuals
    for $\mathscr{X}^p \in V \setminus \{\mathscr{X}^q\}$ **do**
      Fit mvMres: $(\mathscr{X}_{t-i} \setminus \{\mathscr{X}_{t-i}^p\})_{1 \leq i \leq \tau} \mapsto \mathscr{X}_t^q$ and compute its residuals
      z = test to compare (mvMfull) and (mvMres)
      if $z < \alpha$ **then** add edge $\mathscr{X}^p \to \mathscr{X}^q$ to $\mathscr{G}$
  Return the Summary DiGraph $\mathscr{G}$

at t if $\mathscr{X}_t \notin [\mu - k\sigma, \mu + k\sigma]$, where k is a parameter that controls the strength of the detection, and μ and σ are respectively the mean and the standard deviation computed over a sliding window of past values. Samples are then divided into two subgroups, corresponding to normal samples and samples with local distribution shifts. The cost function finally considered corresponds to a weighted quadratic mean of these two subgroups.

Another drawback of Granger causality is related to its underlying linear assumption as associations are highly likely to be non-linear on real datasets. To overcome this, several extensions have been proposed.

For example, Hiemstra and Jones (1994) state that time series $\mathscr{X}^p$ does not Granger-cause times series $\mathscr{X}^q$ if for given values of $a \geq 1, b \geq 1, m \geq 1$ and $\varepsilon > 0$ one has:

$$\begin{aligned}&\Pr\left(\|\mathscr{X}_{t:m}^q - \mathscr{X}_{s:m}^q\|_\infty < \varepsilon \mid \|\mathscr{X}_{t-a:a}^q - \mathscr{X}_{s-a:a}^q\|_\infty < \varepsilon, \|\mathscr{X}_{t-b:b}^p - \mathscr{X}_{s-b:b}^p\|_\infty < \varepsilon\right)\\ &= \Pr\left(\|\mathscr{X}_{t:m}^q - \mathscr{X}_{s:m}^q\|_\infty < \varepsilon \mid \|\mathscr{X}_{t-a:a}^q - \mathscr{X}_{s-a:a}^q\|_\infty < \varepsilon\right).\end{aligned}$$

with $\mathscr{X}_{t:m} = (\mathscr{X}_t, \cdots, \mathscr{X}_{t+m-1})$; the infinite norm $\|.\|_\infty$ corresponds to the maximal component of the vector. A test with correlation-integral estimators is used to determine whether the above equality holds or not.

An analogous causality testing procedure between univariate time series has been developed by Bell et al. (1996). On top of a non parametric regression, an additive modeling framework is used where the restricted and the full models are as follows:

$$\mathscr{X}_t^2 = \sum_{k=1}^{\tau} f_2(\mathscr{X}_{t-k}^2) + \xi_t^{\mathscr{X}^2}, \tag{nlMres}$$

$$\mathscr{X}_t^2 = \sum_{k=1}^{\tau} f_2(\mathscr{X}_{t-k}^2) + \sum_{k=1}^{\tau} f_1(\mathscr{X}_{t-k}^1) + \xi_t^{\mathscr{X}^2|\mathscr{X}^1}, \tag{nlMfull}$$

where $(\xi_t)_t$ are uncorrelated random variables with zero mean and a variance σ^2.

Following slightly different directions, Ancona et al. (2004) proposed to use radial basis functions in the restricted and full models, whereas Chen et al. (2004) proposed a method, called extended Granger causality, which relies on local linear functions corresponding to the standard restricted and full models applied on the points of the same neighborhood. The extended Granger causality is defined as the average of those local Granger causality models. There is a trade-off in this approach between considering large neighborhoods, which ensures representative estimates, and considering small neighborhoods, for which the linearization is more valid.

In Marinazzo et al. (2008), the authors proposed to use kernel approximations of the nonlinear models. Similarly, Sun (2008) used a kernel framework to infer the causality between multivariate time series. Faes et al. (2008) introduced a nonlinear exogenous autoregressive (NARX) model, the parameters of which are estimated through an optimal parameter search. This method is however

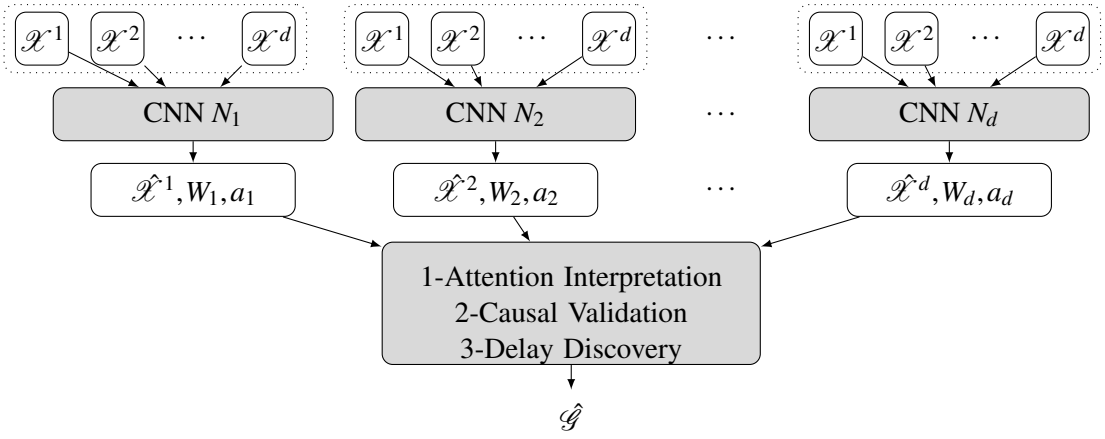


Figure 7: Neural network associated to TCDF: d independent CNNs $(N_q)_{1\leq q\leq d}$, all having time series $\mathscr{X}^1 \cdots \mathscr{X}^d$ of length T as input. For $1 \leq q \leq d$, the network N_q predicts $\mathscr{X}^q$ by $\hat{\mathscr{X}}^q$, and also outputs the kernel weights $(W_{q,p,k})_{1\leq p\leq d, 1\leq k\leq K}$ (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.

parametric, only applicable to bivariate interactions and only appropriate for nonlinearities up to the third order as the number of model parameters that need to be estimated becomes computationally intractable for higher orders. More recently, and still within the bivariate setting, Jiao et al. (2013) proposed a universal estimation of directed information, and detailed how it can be used to infer causal influences within the Granger causality framework. Even more recently, Nicolaou and Constandinou (2016) proposed a method based on Non-Parametric Multiplicative Regression (NPMR), that detects causal relationships by using the error variances obtained from the NPMR model. Papagiannopoulou et al. (2017) presented a direct extension of the standard method by replacing the linear models in the restricted and full models with non-linear models based on random forests. Copulas have also been used to model nonlinear relations between values of time series as by Hu and Liang (2014) and Kim et al. (2019). Lastly, and not surprisingly, several researchers have 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.

3.4 A Deep Learning Extension for Causal Discovery

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 with an attention mechanism within dilated depthwise[8] 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 $\mathscr{X}^q$ as illustrated in Figure 7. Each neural network outputs its prediction, attentions scores and kernel weights which allow a causal interpretation of the results: a high attention on a time series $\mathscr{X}^p$ while forecasting a time series $\mathscr{X}^q$ indicates that the former contains information that helps better forecasting the latter.

Thus, 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 $\mathscr{X}^q$. To interpret the attention scores causally, the Softmax function σ is applied, followed by a semi-binarization step that filters out all attention scores that fall below a threshold s_q. To determine s_q, TCDF starts by ranking the attention

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

Algorithm 3 TCDF

Require: $\mathscr{X}$ a d-dimensional time series of length T, number of hidden layers L, kernel size K, dilation coefficient c, number of epochs, loss function and learning rate
  $\tau_{max} = 1 + (K-1)\sum_{l=0}^{L} c^l$
  Form an empty graph $\mathscr{G}$ with $d\tau_{max}$ nodes V
  for $q \in \{1, \cdots, d\}$ **do**
    Fit $N_q : (\mathscr{X}_{t-i})_{1 \leq i \leq \tau} \mapsto \mathscr{X}_t^q$
    Compute the attention scores a_q and the kernel weights W_q
    Sort the attention scores a_q into b with decreasing order
    Compute the biggest attention score s_q associated to the largest gap in b
    for $p \in \{1, \cdots, d\}$ **do**
      if $\sigma(a_{q,p}) > s_q$ **then**
        $i = \text{argmax}(W_{q,p,.})$
        Add edge $\mathscr{X}_{t-i}^p \to \mathscr{X}_t^q$ to $\mathscr{G}$
        for $(\mathscr{X}_{j-i}^p, \mathscr{X}_j^q) \in \text{Hom}(\mathscr{X}_{t-i}^p, \mathscr{X}_t^q, \mathscr{G})$ **do** add edge $\mathscr{X}_{j-i}^p \to \mathscr{X}_j^q$ to $\mathscr{G}$
    for $\mathscr{X}_{t-i}^p \in \text{Par}(\mathscr{X}_t^q, \mathscr{G})$ **do**
      Compute the loss of N_q on $\mathscr{X}$ where $\mathscr{X}_{t-i}^p$ is permuted
      if the loss increases significantly **then**
        Remove edge $\mathscr{X}_{t-i}^p \to \mathscr{X}_t^q$ from $\mathscr{G}$
        for $(\mathscr{X}_{j-i}^p, \mathscr{X}_j^q) \in \text{Hom}(\mathscr{X}_{t-i}^p, \mathscr{X}_t^q, \mathscr{G})$ **do** remove edge $\mathscr{X}_{j-i}^p \to \mathscr{X}_j^q$ from $\mathscr{G}$
  Return the Window MAG $\mathscr{G}$

scores from high to low and then searches for the largest gap[9] between two adjacent attention scores. The threshold s_q is then equal to the biggest attention score associated to that gap. To distinguish causality-based from correlation-based attention, a causal validation step is applied: potential 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 $\mathscr{X}^q$, 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.

Figure 8: How TCDF deals with hidden confounders. A red edge (left) indicates a wrong causal relation discovered by TCDF, whereas a red double edge (right) indicates that a true causal relation is discovered. Numbers correspond to delays.

As can be seen in Figure 7, TCDF can learn self-causation since it includes the past of $\mathscr{X}^q$ 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 that bidirectional causal relations cannot be instantaneous. For example, TCDF is able to detect the presence of a hidden confounder in Figure 8 (right) but not in Figure 8 (left). A sketch of TCDF is presented in Algorithm 3.

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 tune 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.

9. Additional constraints can be added; for more details see Nauta et al. (2019).

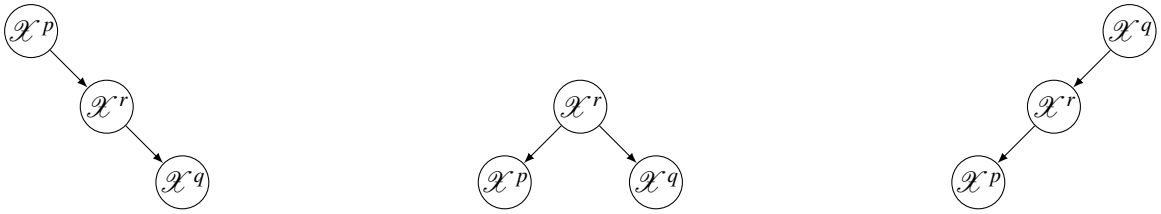


Figure 9: Three equivalent structures

4. 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 (right), page 769). We first cover here the main algorithms assuming causal sufficiency, corresponding to situations when all possible common causes are observed, prior to dealing with situations without causal sufficiency, *i.e.*, with hidden causes.

4.1 With Causal Sufficiency

The goal here is to exploit conditional independencies[10] , obtained from observational data, to construct the underlying causal graph which is typically represented by a directed acyclic graph (DAG) 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 9, borrowed from Verma and Pearl (1991), all represent the same independence relation "$\mathscr{X}^p$ is independent from $\mathscr{X}^q$ given $\mathscr{X}^r$": $\mathscr{X}^p \perp\!\!\!\perp \mathscr{X}^q | \mathscr{X}^r$. 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. They can be represented by partially directed acyclic graphs (PDAGs), in which some edges are not oriented, which can be useful when dealing with situations in which it is difficult, or even impossible, to decide on an orientation. Given an equivalence class of DAGs, Andersson et al. (1997), 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.

For non temporal data, one of the oldest constraint-based algorithm is the SGS algorithm (Spirtes et al., 1990), which has been proved to be consistent under *independently, identically discributed (i.i.d)* observations assuming causal sufficiency. SGS starts with a full undirected graph connecting all variables. In a second step, 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)) and removes the edge between them if it is the case. It then orients undirected edges by subsequently employing orientation rules to derive causal relations. The second step of SGS makes it unusable in practice as the number of conditional independencies that needs to be tested in a fully connected graph grows exponentially with the number of variables, while

10. Conditional independencies can be estimated in a parametric or nonparametric way. We provide here a general explanation of the methods and postpone specific details on the statistical tests used in Section 9.2.

it is known that conditional independencies are difficult to compute (Shah & Peters, 2020). The Peter-Clark (PC) algorithm was introduced (Spirtes et al., 2001) to address this issue.

4.1.1 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 $\text{Sepset}(p,q)$. 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., 2001; Colombo & Maathuis, 2014), 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 paper.

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 sound and complete (Meek, 1995; Andersson et al., 1997) in the set of Markov equivalence graphs, where "sound" means that all causal relations detected by the rules are correct, and "complete" that all possible causal relations in the Markov equivalence class are detected by the algorithm.

Theorem 2 (Theorem 5.1 by Spirtes et al., 2001) *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 $S \subseteq V \setminus \{X^p, X^q\}$. Then, the output of the PC-algorithm is the CPDAG that represents $\mathcal{G}$.*

Algorithm 4 PCMCI

Require: $\mathscr{X}$ a d-dimensional time series of length T, $\tau_{\max} \in \mathbb{N}$ the maximum number of lags, α a significance threshold

Form an oriented graph $\mathscr{G}$ with $d\tau_{max}$ nodes V such that $\mathscr{X}_{t-i}^p \to \mathscr{X}_t^q$ for all $\mathscr{X}_{t-i}^p, \mathscr{X}_t^q \in V, i \in \{1, \cdots, \tau_{max}\}$

for $\mathscr{X}_t^q \in V$ **do**

  $n = 0$

  while $\text{card}(\text{Par}(\mathscr{X}_t^q, \mathscr{G})) \geq n+1$ **do**

    for $\mathscr{X}_{t-i}^p \in \text{Par}(\mathscr{X}_t^q, \mathscr{G})$ s.t. $\text{card}(\text{Par}(\mathscr{X}_t^q, \mathscr{G}) \setminus \mathscr{X}_{t-i}^p) = n$ **do**

      $\mathscr{X}_{\mathbf{I}}^{\mathbf{R}} =$ first n variables of $\text{Par}(\mathscr{X}_t^q, \mathscr{G}) \setminus \{\mathscr{X}_{t-i}^p\}$

      Compute $y_{q,p}$ the statistics that corresponds to the test $\mathscr{X}_{t-i}^p \perp\!\!\!\perp \mathscr{X}_t^q \mid \mathscr{X}_{\mathbf{I}}^{\mathbf{R}}$ and its p-value z

      if $z > \alpha$ **then**

        Remove edge $\mathscr{X}_{t-i}^p \to \mathscr{X}_t^q$ from $\mathscr{G}$

        for $(\mathscr{X}_{j-i}^p, \mathscr{X}_j^q) \in \text{Hom}(\mathscr{X}_{t-i}^p, \mathscr{X}_t^q, \mathscr{G})$ **do** remove edge $\mathscr{X}_{j-i}^p \to \mathscr{X}_j^q$ from $\mathscr{G}$

    Sort $\text{Par}(X_t^q, \mathscr{G})$ by decreasing order of the statistics $(y_{q,p})_p$

    $n = n+1$

for $\mathscr{X}_t^q \in V$ **do**

  for $\mathscr{X}_{t-i}^p \in \text{Par}(X_t^q, \mathscr{G})$ s.t. $\text{card}(\text{Par}(\mathscr{X}_{t-i}^p, \mathscr{G})) > 0$ **do**

    Compute z the p-value that corresponds to the test $\mathscr{X}_t^q \perp\!\!\!\perp \mathscr{X}_{t-i}^p \mid \text{Par}(\mathscr{X}_t^q, \mathscr{G}) \setminus \{\mathscr{X}_{t-i}^p\} \cup \text{Par}(\mathscr{X}_{t-i}^p, \mathscr{G})$

    if $z > \alpha$ **then**

      Remove edge $\mathscr{X}_{t-i}^p \to \mathscr{X}_t^q$

      for $(\mathscr{X}_{j-i}^p, \mathscr{X}_j^q) \in \text{Hom}(\mathscr{X}_{t-i}^p, \mathscr{X}_t^q, \mathscr{G})$ **do** remove edge $\mathscr{X}_{j-i}^p \to \mathscr{X}_j^q$ from $\mathscr{G}$

Return the window DAG $\mathscr{G}$

Consistency of the PC algorithm has been discussed by Spirtes et al. (2001), 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 (2014) 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 the 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.

4.1.2 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 Figure 5c page 772). The method is divided into three steps. First, a partially connected graph $\mathscr{G}$ is constructed, such that all pairs of nodes $(\mathscr{X}_{t-i}^p, \mathscr{X}_t^q)$ are directed as $\mathscr{X}_{t-i}^p \to \mathscr{X}_t^q$ 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 throug time to remove homologous edges: for each edge $\mathscr{X}_{t-i}^p \to \mathscr{X}_t^q$ removed, all edges included in $\text{Hom}(\mathscr{X}_{t-i}^p, \mathscr{X}_t^q, \mathscr{G})$ are removed as well, where $\text{Hom}(\mathscr{X}_{t-i}^p, \mathscr{X}_t^q, \mathscr{G})$ represents the set of instants homologous to $\mathscr{X}_{t-i}^p$ and $\mathscr{X}_t^q$, *i.e.*, instants in $\mathscr{X}^p$ and $\mathscr{X}^q$ shifted by a lag of i from p to q (see Section 2, Table 1). As the conditioning is based only on the parents of $\mathscr{X}_t^q$, one cannot control false positives with large autocorrelations in $\mathscr{X}_{t-i}^p$. The third step deals with these autocorrelations by using the Momentary Conditional Independence test (MCI). MCI conditions on the parents of $\mathscr{X}_t^q$ and the parents of $\mathscr{X}_{t-i}^p$

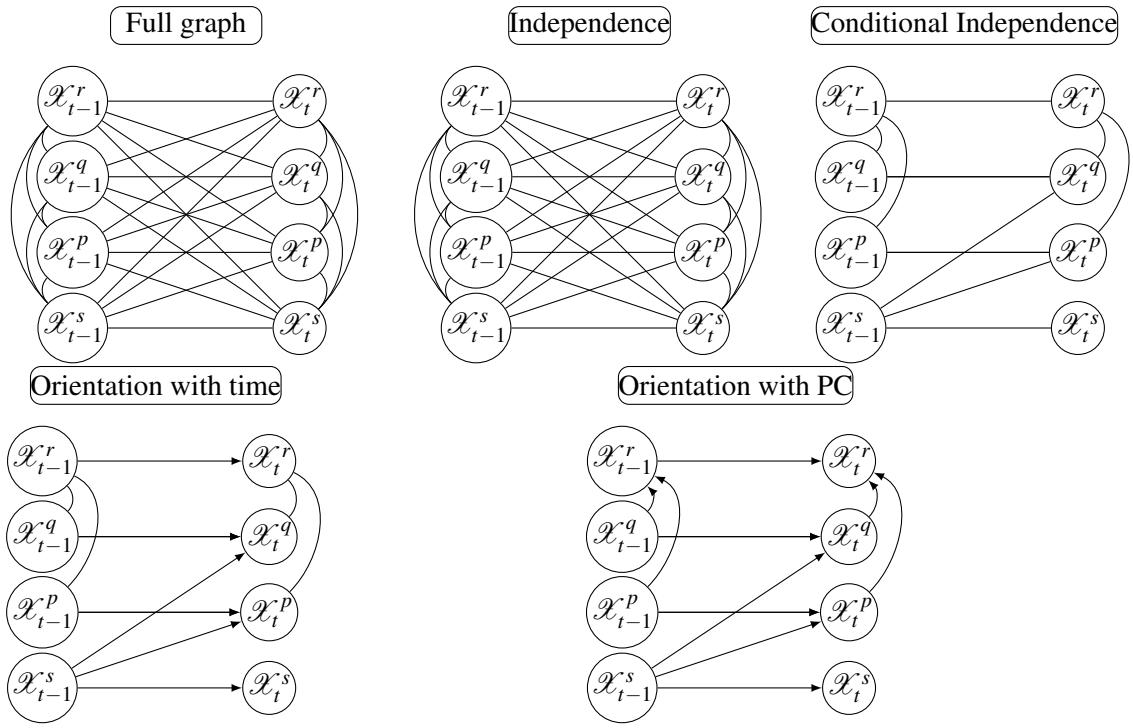


Figure 10: Running example: structure inferred by PCMCI with instantaneous relations.

while testing $\mathscr{X}^p_{t-i} \to \mathscr{X}^q_t$. It is defined as follows: for m a measure of dependence,

$$\text{MCI}(\mathscr{X}^p_{t-i}; \mathscr{X}^q_t) = m\left(\mathscr{X}^p_{t-i}; \mathscr{X}^q_t \,\middle|\, \text{Par}(\mathscr{X}^q_t)\backslash\{\mathscr{X}^p_{t-i}\}, \text{Par}(\mathscr{X}^p_{t-i})\right),$$

and estimates an interpretable notion of causal strength as it quantifies the causal effect on $\mathscr{X}^q_t$ of a hypothetical perturbation in $\mathscr{X}^p_{t-\tau}$. Thus, the value of the MCI statistics allows to rank causal links in large-scale settings. The algorithm is described in Algorithm 4. The method depends on the significant rate α, which can be selected using the Akaike Information Criterion or cross validation. The computational time is polynomial in the number d of time series and the maximum lag τ_{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 (Section 9) on two measures used by Runge et al. (2019), namely the partial correlation and the mutual information.

Instantaneous causal relations, which were not supported in the initial algorithm, have been integrated by 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.

Figure 10 illustrates the different steps of this algorithm on our running example. Note that, here, all edges in Step 2 are kept as all nodes in the window graph are dependent without conditioning.

4.1.3 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 issue would be to use an asymmetric measure such as transfer entropy (Schreiber, 2000), which can be defined as follows:

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

Algorithm 5 oCSE

Require: $\mathscr{X}$ a d-dimensional time series of length T, α a significance threshold
  Form an empty graph $\mathscr{G}$ with d nodes V
  for $\mathscr{X}^q \in V$ **do**
    $z = \infty$
    while $z > 0$ and $\text{card}(\text{Par}(\mathscr{X}^q, \mathscr{G})) < d$ **do**
      for $\mathscr{X}^p \in V \setminus \text{Par}(\mathscr{X}^q, \mathscr{G})$ **do**
        Compute z_p the p-value that corresponds to the test $\text{CE}(\mathscr{X}_t^p \to \mathscr{X}_{t+1}^q \mid \text{Par}(\mathscr{X}^q, \mathscr{G})_t) > 0$
      $p = \text{argmax}_r z_r$
      if $z_p > \alpha$ **then** add edge $\mathscr{X}^p \to \mathscr{X}^q$ to $\mathscr{G}$
    for $\mathscr{X}^p \in \text{Par}(\mathscr{X}^q, \mathscr{G})$ **do**
      Compute z the p-value that corresponds to the test $\text{CE}(\mathscr{X}_t^p \to \mathscr{X}_{t+1}^q \mid \text{Par}(\mathscr{X}^q, \mathscr{G})_t \setminus \{\mathscr{X}_t^p\}) = 0$
      if $z > \alpha$ **then** remove edge $\mathscr{X}^p \to \mathscr{X}^q$ from $\mathscr{G}$
  Return the summary DiGraph $\mathscr{G}$

where $h(. \mid .)$ denotes the conditional entropy. However, this metric is limited to pairwise relations and assumes that nodes are self causal. To overcome this, 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:

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

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, summarized in Algorithm 5, to find, for each node $\mathscr{X}_t^p$, 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:

$$\mathscr{X}_t^q = f_q(a_1 \mathscr{X}_{t-1}^1, a_2 \mathscr{X}_{t-1}^2, \cdots, a_d \mathscr{X}_{t-1}^d, \xi_t^p),$$

where for all $p \in \{1, \ldots, d\}$, a_p is the weight of the link from $\mathscr{X}^p$ to $\mathscr{X}^q$. Note that the parents of $\mathscr{X}_t^q$ can only be attributed to the time $t-1$, known as the Temporally Markov assumption: for all t, $\Pr(\mathscr{X}_t \mid \mathscr{X}_{t-1}, \mathscr{X}_{t-2}, \cdots) = \Pr(\mathscr{X}_t \mid \mathscr{X}_{t-1})$. oSCE 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.

4.2 Without Causal Sufficiency

As explained in Section 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 4, page 771, 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.

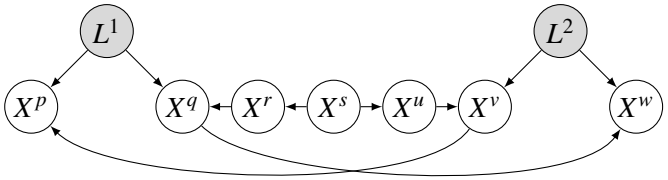


Figure 11: Causal graph with two hidden common causes (Spirtes et al., 2001)

Definition 11 (Discriminating path, Zhang, 2007) *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 by 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 ($\longleftrightarrow$), 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 sets, introduced by Spirtes et al. (2001). 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. (2001) but that simplifies the exposition of the overall procedure.

Definition 12 (Possible-Dsep, Spirtes et al., 2001; Zhang, 2008) *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 11, which displays two hidden comon causes 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.

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.

4.2.1 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 by, *e.g.*, Zhang (2008), are applied recursively[11]. 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 $X^p *\!\!-\!\!\circ X^r \circ\!\!-\!\!* X^q$, if $X^r \notin Sepset(p,q)$, then orient the unshielded triple as a collider: $X^p *\!\!\rightarrow X^r \leftarrow\!\!* X^q$.*

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

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 $X^s \longleftrightarrow X^r \longleftrightarrow X^q$.*

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 $V_1, \cdots, V_l$, that is such that an edge between two consecutive vertices V_{i-1} and V_i has no arrow on V_{i-1}'s side and has either an arrow or a circle on V_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 $X^p - X^r \ast\!\!-\!\!\circ X^q$ (X^p and X^q are not necessarily adjacent), then orient the triple as $X^p - X^r \rightarrow\!\!\ast X^q$.*

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

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, 2008) 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. Really Fast Causal Inference (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.

4.2.2 TEMPORAL EXTENSION THROUGH ADDITIVE NON-LINEAR TIME SERIES MODEL

Inspired by FCI, Chu and Glymour (2008) proposed a method that can deal with hidden confounders when they are linear and instantaneous. Constraint-based methods originally use nonparametric conditional independence tests that are subjects to the curse of dimensionality. To avoid this issue, Chu and Glymour (2008) assumed additive non-linear time series models (ANLTSM) that can be represented as:

$$\mathscr{X}_t^q = \sum_{1\leq p\leq d, p\neq q} a_{q,p}\mathscr{X}_t^p + \sum_{1\leq p\leq d, 1\leq l\leq \tau} f_{q,p,l}(\mathscr{X}_{t-l}^p) + \sum_{r=1}^{h} b_{q,r}\mathscr{U}_t^r + \xi_t, \qquad \text{(ANLTSM)}$$

where $b_{q,r}$s and $a_{q,p}$s are constants, and $f_{q,p,l}$s are smooth univariate functions. $(\mathscr{U}_t^r)_{1\leq r\leq h}$ and ξ_t are unobserved multi-dimensional Gaussian white noise. ξ_t represents latent causes, which can only be direct causes of the observable variables, while $(\mathscr{U}_t^r)_{1\leq r\leq h}$, represents latent common causes. The latter are allowed only for variables at the same time instant, and $\mathscr{X}_t^p$ and $\mathscr{X}_t^q$ have a latent common cause $\mathscr{U}_t^r$ if and only if there exists $1 \leq r \leq h$ such that $b_{q,r}b_{p,r} \neq 0$.

Assuming additive non-linear time series models, Chu and Glymour (2008) test if two nodes $\mathscr{X}_t^p$ and $\mathscr{X}_t^q$ are independent conditionally on the set $\mathbf{S}$ by estimating the conditional expectation of $\mathscr{X}_t^p$ given $\mathscr{X}_t^q \cup \mathbf{S}$ using additive regression models, and check if $\mathscr{X}_t^q$ is a significant predictor for $\mathscr{X}_t^p$ using statistical tests such as the F-test (Bell et al., 1996) or the BIC scores (Huang & Yang, 2004). The insignificancy of the predictor implies that $\mathscr{X}_t^p$ and $\mathscr{X}_t^q$ are conditionally independent.

Using the above test in the FCI algorithm, one first identifies instantaneous causal relations. Lagged causal relations $\mathscr{X}_{t-i}^p \to \mathscr{X}_t^q$ are then identified through one of the following two conditions:

(a) If $\mathscr{X}_{t-i}^p$ and $\mathscr{X}_t^q$ are still dependent given any subset of instantaneous direct causes of $\mathscr{X}_t^q$ and any subset of lagged neighbours and if $\mathscr{X}_{t-i}^p$ is not adjacent to any other node at time t;

(b) Or if $\mathscr{X}_{t-\tau}^p$ and $\mathscr{X}_t^q$ are still dependent given any subset of instantaneous relations and any subset of lagged neighbours.

Finally, the remaining edges are oriented, whenever possible, by two additional rules that detect instantaneous causes in unshielded triples by testing conditional independence with the third variable given the past. This method has been shown to be consistent when the data is generated from an ANLTSM.

4.2.3 TEMPORAL EXTENSION THROUGH WINDOW REPRESENTATIONS AND SVAR

Entner and Hoyer (2010) adapted FCI to time series by transforming the original time series $\mathscr{X}_t = (\mathscr{X}_t^1, \cdots, \mathscr{X}_t^d)_{1\leq t\leq T}$ into a sample of random vectors with a sliding window of size τ. This leads to the consideration of $(T-\tau+1)$ 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. It is described in Algorithm 6.

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.

4.2.4 TEMPORAL EXTENSION USING CONDITIONAL INDEPENDENCE TESTS

Very recently, Gerhardus and Runge (2020) extended PCMCI, introduced in Section 4.1.2, to LPCMCI (for Latent PCMCI) to take into account latent variables, which contrasts with previous methods that rather extend FCI. In LPCMCI, known parents are used as default conditions whereas non-ancestors are not tested in conditioning sets. Furthermore, a new type of edge, with a middle mark, is used to facilitate early orientation of edges. In a preliminary phase, ancestors are detected during the

Algorithm 6 tsFCI

Require: $\mathscr{X}$ a d-dimensional time series of length T, $\tau_{max} \in \mathbb{N}$ the window length
Form a complete undirected graph $\mathscr{G}$ with $d\tau_{max}$ nodes V with all edges of the form $\circ\!\!-\!\!\circ$
n = 0
while there exists $\mathscr{X}_t^q \in V$ such that $\text{card}(\text{Adj}(\mathscr{X}_t^q, \mathscr{G})) \geq n+1$ **do**
  for $\mathscr{X}_t^q \in V$ such that $\text{card}(\text{Adj}(\mathscr{X}_t^q, \mathscr{G})) \geq n+1$ **do**
    for $\mathscr{X}_{t-i}^p \in \text{Adj}(\mathscr{X}_t^q, \mathscr{G})$ such that $\text{card}(\text{Adj}(\mathscr{X}_t^q, \mathscr{G}) \setminus \{X_{t-i}^p\}) \geq n$ **do**
      for $\mathscr{X}_{\mathbf{T}}^{\mathbf{R}} \subset \text{Adj}(\mathscr{X}_t^q, \mathscr{G}) \setminus \{\mathscr{X}_{t-i}^p\}$ such that $\text{card}(\mathscr{X}_{\mathbf{T}}^{\mathbf{R}}) = n$ **do**
        $(z, y) =$ test if $X_{t-i}^p \perp\!\!\!\perp \mathscr{X}_t^q \mid \mathscr{X}_{\mathbf{T}}^{\mathbf{R}}$
        Compute z the p-value that corresponds to the test $X_{t-i}^p \perp\!\!\!\perp \mathscr{X}_t^q \mid \mathscr{X}_{\mathbf{T}}^{\mathbf{R}}$
        if $z > \alpha$ **then**
          Remove edge $\mathscr{X}_{t-i}^p - \mathscr{X}_t^q$ from $\mathscr{G}$
          for $(\mathscr{X}_{j-i}^p, \mathscr{X}_j^q) \in \text{Hom}(\mathscr{X}_{t-i}^p, \mathscr{X}_t^q, \mathscr{G})$ **do** remove edge $\mathscr{X}_{j-i}^p - \mathscr{X}_j^q$ from $\mathscr{G}$
          $\text{Sepset}(\mathscr{X}_{t-i}^p, \mathscr{X}_t^q) = \text{Sepset}(\mathscr{X}_t^q, \mathscr{X}_{t-i}^p) = \mathscr{X}_{\mathbf{T}}^{\mathbf{R}}$
  n=n+1
for each unshielded triple in $\mathscr{G}$ **do** apply F-Rule 0
for $\mathscr{X}_t^q \in V$ **do**
  for $\mathscr{X}_{t-i}^p \in \text{Adj}(\mathscr{X}_t^q, \mathscr{G})$ **do**
    Compute z the p-value that corresponds to the test $\mathscr{X}_{t-i}^p \perp\!\!\!\perp \mathscr{X}_t^q \mid \text{Possible-Dsep}(\mathscr{X}_{t-i}^p, \mathscr{X}_t^q)$
    if $z > \alpha$ **then**
      Remove edge $\mathscr{X}_{t-i}^p - \mathscr{X}_t^q$ from $\mathscr{G}$
      for $(\mathscr{X}_{j-i}^p, \mathscr{X}_j^q) \in \text{Hom}(\mathscr{X}_{t-i}^p, \mathscr{X}_t^q, \mathscr{G})$ **do** remove edge $\mathscr{X}_{j-i}^p - \mathscr{X}_j^q$ from $\mathscr{G}$
      $\text{Sepset}(\mathscr{X}_{t-i}^p, \mathscr{X}_t^q) = \text{Sepset}(\mathscr{X}_t^q, \mathscr{X}_{t-i}^p) = \text{Possible-Dsep}(\mathscr{X}_{t-i}^p, X_t^q)$
Reorient all edges as $\circ\!\!-\!\!\circ$
for each adjacent pair $(\mathscr{X}_{t-i}^p, \mathscr{X}_t^q)$ in $\mathscr{G}$ with $i > 0$ **do** orient the pair as $(\mathscr{X}_{t-i}^p \ast\!\!\rightarrow \mathscr{X}_t^q)$
for each adjacent pair $(\mathscr{X}_t^p, \mathscr{X}_t^q)$ in $\mathscr{G}$ **do** orient the pair as $(\mathscr{X}_t^p \longleftrightarrow \mathscr{X}_t^q)$
for each unshielded triple in $\mathscr{G}$ **do** apply F-Rule 0
while no more edges can be oriented **do** apply FCI-Rules 1 to 10
Return the window PAG $\mathscr{G}$

classical skeleton construction through additional orientation rules, easily adapted from the one introduced in Section 4.2.1. Then, in a final phase, edges are re oriented using the same rules. This algorithm is order independent, sound and complete.

5. 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) based on two random variables X^p and X^q such that $X^p \to X^q$ with the underlying relation $X^q = 2X^p + \xi^q$, where ξ^q represents some noise. Given enough observations, one can detect a relation 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 a uniform distribution on $\{-1, 0, 1\}$, then one

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 2: 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. The first two columns correspond to observed values of X^p and X^q.

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 2.

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 (potentially with noise) and 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[12].

3. Noises associated to 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 $X^p \perp\!\!\!\perp \xi^q$.

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 given by $X^q = f_q(X^p, \xi^q)$, 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.

12. *Identifiability* in this context refers to the fact that it is possible to infer causal relations from observational data.

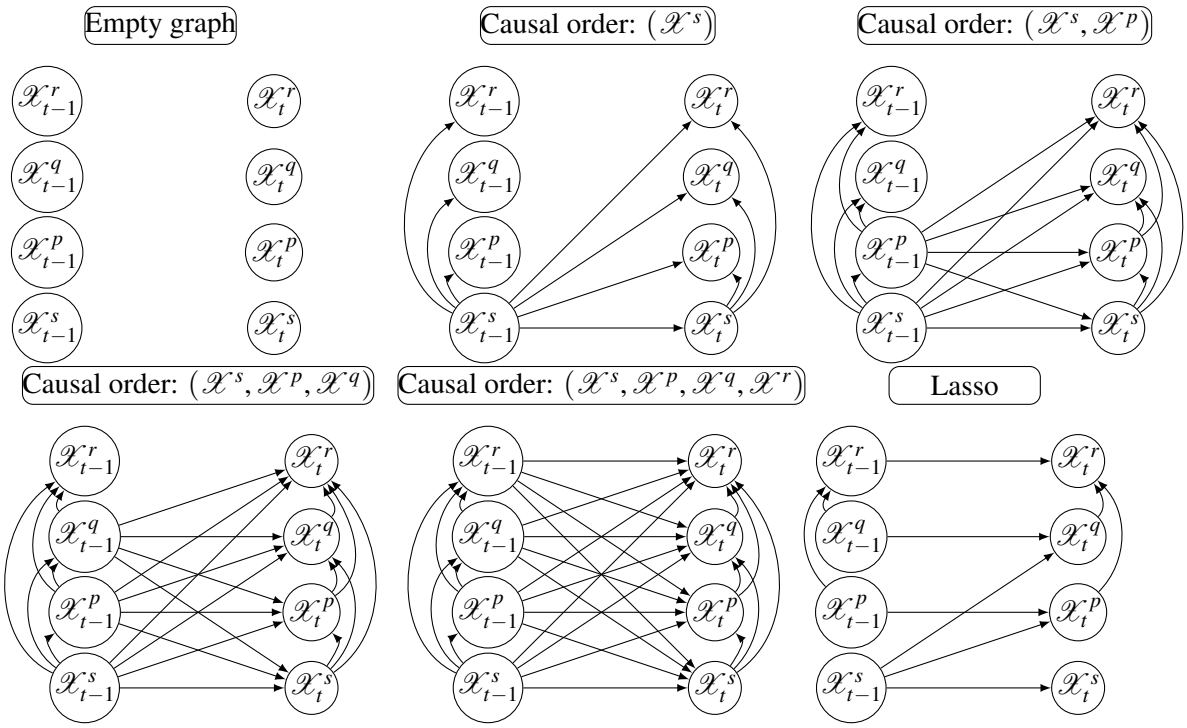


Figure 12: Running example: structured inferred by VarLiNGAM. The first line corresponds to causal ordering, and the second line to graph pruning.

5.1 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 3 (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.

5.1.1 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:

$$\mathscr{X}_t = \sum_{i=0}^{\tau} \mathbf{A}_i \mathscr{X}_{t-i} + \mathbf{e}_t, \tag{SVAR}$$

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

$$\mathscr{X}_t = \sum_{i=1}^{\tau} \mathbf{M}_i \mathscr{X}_{t-i} + \mathbf{e}_t. \tag{VAR}$$

The above model, estimated through a least-square procedure, is used to obtain residuals of the prediction of $\mathscr{X}_t$. A standard LiNGAM analysis is then used on these residuals to obtain an instantaneous causal model $\mathbf{A}_0$. Finally, $(\mathbf{A}_i)_{i>0}$ are deduced by a reparametrization of $(\mathbf{M}_i)_{i>0}$:

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

An intensive illustration of this approach on economic data is provided by Moneta et al. (2013), while Algorithm 7 further details its different steps. Figure 12 further illustrates on our running example the ordering steps and the pruning process yielding a sparse graph.

Huang et al. (2015) extended VarLiNGAM by considering linear and nonlinear time-varying models, both with unobserved confounders. For a multivariate time series $\mathscr{X}_t = (\mathscr{X}_t^1, \ldots, \mathscr{X}_t^d)$, these models take the form:

$$\begin{cases} \mathscr{X}_t^i = \sum_{j=1}^{d} \sum_{p=1}^{P} a_t^{ijp} \mathscr{X}_{t-p}^j + \sum_{k \neq i} b_t^{ik} \mathscr{X}_t^k + g_t^i + \xi_t^i & \text{(VAR-t)}, \\ \mathscr{X}_t^i = f^i(t, \{\mathscr{X}_{t-p}^j\}_j, \{\mathscr{X}_t^k\}_{k \neq i}) + \xi_t^i & \text{(VAR-t-nl)}, \end{cases}$$

where ξ_t^i are *i.i.d.* noise independent of the causes, a_t^{ijp} represent the time-varying lagged causal coefficients, b_t^{ik} give the instantaneous causal coefficients, and g_t^i represent the causal influences from unobserved confounders that are assumed to be smooth functions of time. The (non necessarily linear) functions f^i take into account time-varying causal relations as well as the influence of confounders. Conditions to ensure identifiability for both models are provided, as well as a non-parametric method to estimate the time-dependent causal models based on Gaussian processes. Note that this formulation is more general than the classical vector autoregressive model, but is based on the same steps as VarLiNGAM.

Geiger et al. (2015) has extended the vector autoregressive model for $\mathscr{X}$ to take into account hidden components $\mathscr{Z}$, while considering a lag of size 1:

$$\begin{pmatrix} \mathscr{X}_t \\ \mathscr{Z}_t \end{pmatrix} = \begin{pmatrix} B & C \\ D & E \end{pmatrix} \begin{pmatrix} \mathscr{X}_{t-1} \\ \mathscr{Z}_{t-1} \end{pmatrix} + \mathscr{N}_t,$$

Algorithm 7 VarLiNGAM

Require: $\mathscr{X}$ a d-dimensional time series of length T, $\tau_{\max} \in \mathbb{N}$ the maximum number of lags, α a significance threshold
Form an empty graph $\mathscr{G}$ with $d\tau$ nodes V
Find the optimal lag $\tau \in \{1, \cdots, \tau_{max}\}$
Fit (VAR): $\mathscr{X} \mapsto \mathscr{X}$
Compute $(\mathbf{M}_\tau)_{1 \leq \tau \leq \tau_{\max}}$ the coefficients of (VAR) and $\xi =$ its residuals
$\mathbf{S} = \{1, \cdots, d\}$
while $\text{card}(\mathbf{S}) > 1$ **do**
  for $p \in \mathbf{S}$ **do**
    for $q \in \mathbf{S} \setminus \{p\}$ **do**
      Fit least squares regressions: $\xi^p \mapsto \xi^q$ and compute its residuals $\varepsilon^{p,q}$
    Compute y_p the statistics that corresponds to the test $\varepsilon^{p,\cdot} \perp\!\!\!\perp \xi^p$
  $p^* = \text{argmin}_{q \in \mathbf{S}} y_p$
  $\mathbf{S} = \mathbf{S} \setminus \{p^*\}$
  for $\mathscr{X}^q \in \mathscr{X}^{\mathbf{S}}$ **for** $i \in \{0, \cdots, \tau\}$ **do** add edge $\mathscr{X}^{p^*}_{t-i} \to \mathscr{X}^q_t$ to $\mathscr{G}$
Construct a strictly lower triangular matrix $\mathbf{A}_0$ by following the order in $\mathscr{G}$, and estimate the connection strengths $[\mathbf{A}_0]_{i,j}$ by using some conventional covariance-based.
for $i \in \{1, \cdots, \tau\}$ **do**
  $\mathbf{A}_i = (I - \mathbf{A}_0)\mathbf{M}_i$
Apply Adaptive Lasso on $\mathbf{A}$
for $i \in \{0, \cdots, \tau\}$ **do**
  for $q \in \{1, \cdots, d\}$ **do**
    for $p \in \{1, \cdots, d\}$ **do**
      if $[\mathbf{A}_i]_{p,q} = 0$ **then** remove edge $\mathscr{X}^p_{t-i} \to \mathscr{X}^q_t$ from $\mathscr{G}$
Return the window DAG $\mathscr{G}$

where the noise $(\mathscr{N}_t)_t$ is *i.i.d.*. The authors showed that the model is identifiable when the noise terms are mixtures of Gaussian and propose a variational EM algorithm to estimate the causal model in that case. The model is also identifiable (up to scaling and permutation indeterminacies, because scale and ordering of the components of $\mathscr{Z}$ are arbitrary) when there is no influence from $\mathscr{X}$ on $\mathscr{Z}$.

Lastly, more recently, Lanne et al. (2017) generalized the initial VarLiNGAM model by considering graphs that can contain cycles. They further proved that the proposed model is identifiable and introduced an estimation method based on maximum likelihood. The proposed estimator is furthermore proven to be asymptotically efficient.

5.2 Additive Noise Models

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 4. 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 4 (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*

Algorithm 8 TiMINo

Require: $\mathscr{X}$ a d-dimensional time series of length T, $\tau_{\max} \in \mathbb{N}$ the maximum number of lags, α a significance threshold
Form an empty graph $\mathscr{G}$ with d nodes V
$\mathbf{S} = \{1, \cdots, d\}$
while $\text{card}(\mathbf{S}) > 1$ **do**
  for $\mathscr{X}^q \in \mathscr{X}^{\mathbf{S}}$ **do**
    $\mathscr{X}^{\mathbf{R}} = \mathscr{X}^{\mathbf{S}} \setminus \{\mathscr{X}^q\}$
    Fit: $(\mathscr{X}^q_{t-\tau_{\max}}, \cdots, \mathscr{X}^q_{t-1}, \mathscr{X}^{\mathbf{R}}_{t-\tau_{\max}}, \cdots, \mathscr{X}^{\mathbf{R}}_t) \mapsto \mathscr{X}^q_t$ and compute its residuals ξ^q_t
    Compute z_q the p-value that corresponds to the test $\mathscr{X}^{\mathbf{R}} \perp\!\!\!\perp \xi^q_t$
  $q^* = \text{argmax}_q z_q$
  $q = q^*$-th element of $\mathbf{S}$
  if $z_q > \alpha$ **then**
    $\mathbf{S} = \mathbf{S} \backslash \{q\}$
    for $\mathscr{X}^p \in \mathscr{X}^{\mathbf{S}}$ **do** add edge $\mathscr{X}^p \to \mathscr{X}^q$ to $\mathscr{G}$
  else break and output: "I do not know, bad model fit"
for $\mathscr{X}^q \in V$ **do**
  for $\mathscr{X}^p \in \text{Par}(\mathscr{X}^q, \mathscr{G})$ **do**
    $\mathscr{X}^{\mathbf{R}} = \text{Par}(\mathscr{X}^q, \mathscr{G}) \setminus \{\mathscr{X}^p\}$
    Fit: $(\mathscr{X}^q_{t-\tau_{\max}}, \cdots, \mathscr{X}^q_{t-1}, \mathscr{X}^{\mathbf{R}}_{t-\tau_{\max}}, \cdots, \mathscr{X}^{\mathbf{R}}_t) \mapsto \mathscr{X}^q_t$ and compute its residuals ξ^q_t
    Compute z the p-value that corresponds to the test $\mathscr{X}^{\mathbf{R}} \perp\!\!\!\perp \xi^q_t$
    if $z > \alpha$ **then** remove edge $\mathscr{X}^p \to \mathscr{X}^q$ from $\mathscr{G}$
Return the summary DAG $\mathscr{G}$

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.

5.2.1 TiMINo: Times Series Model with Independent Noise

A class of restricted FCM called Time Series Models with Independent Noise (TiMINo) is studied by Peters et al. (2013). For a multivariate time series $\mathscr{X}$ 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 $\tau > 0$ such that for all $p \in V$ there are sets $\text{Par}(\mathscr{X}^p_0, \mathscr{G}) \subseteq V \backslash \{\mathscr{X}^p\}, \text{Par}(\mathscr{X}^p_k, \mathscr{G}) \subseteq V$ for $1 \leq k \leq \tau$ such that for all t:

$$\mathscr{X}^p_t = f_p(\text{Par}(\mathscr{X}^p_\tau, \mathscr{G})_{t-\tau}, \ldots, \text{Par}(\mathscr{X}^p_1, \mathscr{G})_{t-1}, \text{Par}(\mathscr{X}^p_0, \mathscr{G})_t, \xi^p_t), \tag{TiMINo}$$

where ξ^i_t 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 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 graph, and acyclicity in the summary graph), the summary graph can be recovered from the joint distribution of $\mathscr{X}$. To infer the causal graph in the additive noise 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.

6. Score-Based Approaches

In score-based approaches, a causal graph corresponds to a *probabilistic (or Bayesian) network*; furthermore, a *dynamic probabilistic (or dynamic Bayesian) network* (DPN) is a probabilistic network in which variables are time series. We make use of this terminology in this section. We also want to make clear that there is no guarantee that the Bayesian network inferred by score-based approaches belongs to the equivalence class of the graph underlying the observed (stable) probability distribution. Indeed, score-based methods aim at finding sparse structural equation models that best explain the data, without any guarantee on the corresponding DAG (Kaiser & Sipos, 2021). This contrasts with, *e.g.*, constraint-based approaches.

The problem of learning a probabilistic network from observational data can be formulated as: given a set of instances, find the network that best matches them. In score-based approaches, the notion of best-match is based on a score that typically strikes a balance between the likelihood of the data given the network and a penalty term related to the complexity of the network. Compared to constraint-based approaches, score-based approaches have the advantage of assigning a score to the network inferred, a score that can then be used to assess the validity of the network. However, the solution obtained is in general suboptimal as finding a globally optimal network is known to be NP-hard (Chickering, 1995). In addition, hidden variables have to be "postulated", a fact due to the use of the *Expectation-Maximization* algorithm, and are not "discovered" as in constraint-based methods.

A standard algorithm for inferring probabilistic networks is the *Structural Expectation-Maximization* (SEM) algorithm, introduced by Friedman (1997). This algorithm combines parametric and structural modifications, the former aiming at finding better parameters and the latter at finding better structures. This algorithm has been extended by Friedman (1998) to deal with scoring functions based on true Bayesian scores, and by Friedman et al. (1998) to deal with dynamic probabilistic networks.

A dynamic probabilistic network can be decomposed in a prior network, which provides dependencies between variables in a given time stamp, and a transition network, which provides dependencies over time. The transition networks considered by Friedman et al. (1998) are Markovian, in the sense that $\Pr(\mathscr{X}_{t+1}|\mathscr{X}_0,\cdots,\mathscr{X}_t) = \Pr(\mathscr{X}_{t+1}|\mathscr{X}_t)$ where $\mathscr{X}$ is a multivariate time series, and stationary, meaning that $\Pr(\mathscr{X}_{t+1}|\mathscr{X}_t)$ is independent of t. These assumptions are limitations, as high-order temporal dependencies are not considered, imposed by the complexity of the inference process when they are removed. The Bayesian scores taken into account are the Bayesian information criterion (BIC, Schwarz, 1978) and the Bayesian Dirichlet equivalence score (BDe, Heckerman et al., 1995), based on Dirichlet priors on the structural parameters. At each iteration, the overall process consists in first improving the parameters of the prior and transition networks, and then searching over DPN structures using expected counts to select the best scoring structures. The search over DPN structures is based on heuristics that typically consider neighboring structures of a given structure, obtained by arc additions, removals and reversals. Experimental results obtained on simulated data for traffic patterns, in which mixtures of different driving tendencies are considered, as well as on molecular biology data, with the aim to infer the structure of regulatory pathways, validate the SEM algorithm for DPNs in the context of noisy and missing data, provided of course that the assumptions made are realistic, which is the case in the two applications retained.

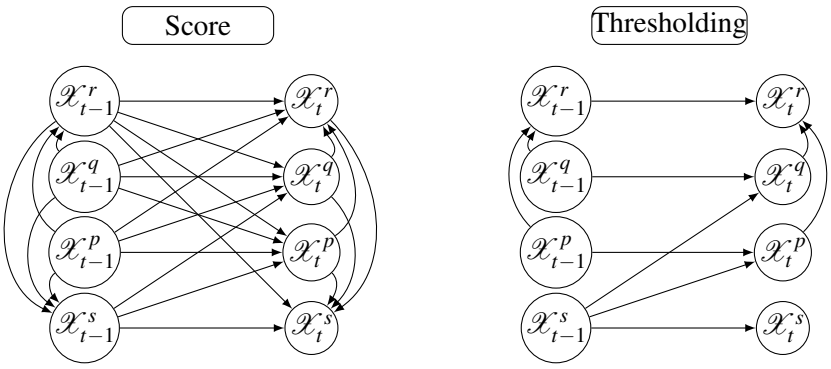


Figure 13: Running example: a diamond structure inferred by DYNOTEARS.

Within the same framework and assumptions, Peña et al. (2005) proposed to use cross-validation (CV) as the Bayesian score underlying the learning process. Indeed, one is typically interested in finding DPNs $(\mathscr{G}, \hat{\theta})$ that generalize well, where $\mathscr{G}$ is the underlying graph and $\hat{\theta}$ the maximum likelihood (or maximum a posteriori) estimates of the parameters which define the transition probabilities. One way to do so is to look for the DPN, trained on S instances, that maximizes the expectation of the log-likelihood on any unseen instance D_{S+1}: $\mathbb{E}[\log \Pr(D_{S+1}|\mathscr{G}, \hat{\theta})]$. As this expectation cannot be computed directly, the authors propose to use a cross-validation approach and compute for each fold D^k of the training data D the quantity $\log \Pr(D^k|\mathscr{G}, \hat{\theta})$. The above expectation can then be estimated by:

$$\frac{1}{S}\sum_{k=1}^{K} \log \Pr(D^k|\mathscr{G}, \hat{\theta})$$

where S is the number of instances in D and K the number of folds. In practice, when there are more than a few tens of nodes, one cannot do an exhaustive search over all graphs. The authors thus relied on a greedy hill-climbing search that gradually improves a graph through a highest scoring single edge addition or removal. This can be seen as a special case, with no reversal, of the greedy equivalence search algorithm introduced by Meek (1997) and studied by Chickering (2002) who proves the conjecture on which it is based. Furthermore, as CV may overfit (Ng, 1997), the authors modified the hill-climbing search by adding an edge only if the improvement in CV is significant, according to a statistical test. Experiments, conducted on data generated from random DPNs and the Yeast dataset, show that the CV-based scoring method leads to models that generalize better than those based on BIC and BDe for a wide range of sample sizes, in particular the range of sizes one usually encounters in practice. One can also note another extension of SEM for fMRI data, referred to as extended unified SEM, presented by Gates and Molenaar (2012) and based on a bilinear system to describe brain regions of interest.

More recently, Pamfil et al. (2020) have proposed a method, named DYNOTEARS, which simultaneously estimates instantaneous and time-lagged relationships between time series through $d \times d$ adjacency matrices $W, A_1, \cdots A_{\tau_{max}}$ that respectively represent the importance of the relation between two time series with a time lag 0 (instantaneous relations, W), 1 (A_1), ..., τ_{max} ($A_{\tau_{max}}$). These matrices are learned by minimizing the following penalized loss based on the Frobenius norm of the residuals of a linear model:

$$f(W,A) = \frac{1}{2d(T+1-\tau_{max})} \| \mathscr{X}_t - \mathscr{X}_t^T W - \mathscr{X}_{t+1:t+\tau_{max}}^T A \|_F^2 + \lambda_W \|W\|_1 + \lambda_A \|A\|_1, \qquad \text{(Score)}$$

where ^{T} denotes the transpose, $\|.\|_1$ stands for the element-wise ℓ_1 norm and λ_W and λ_A represent regularization constants. The causal graph is then obtained by successively considering all relations at different time lags, as described in Algorithm 9. To avoid cycles, an acyclicity constraint on the instantaneous adjacency matrix W is used, which is solved using an equivalent formulation based on the trace exponential function (Zheng et al., 2020). This method is illustrated in Figure 13 on

our running example. First, the best sparse window DAG is selected using a score. A thresholding step is then used to prune some spurious correlations.

Algorithm 9 DYNOTEARS

Require: $\mathscr{X}$ a d-dimensional time series of length T, $\tau_{\max} \in \mathbb{N}$ the maximum number of lags, λ_A, λ_W, α
  $W, A = \min_{W,A} f(W,A)$ from (Score)
  for $w_{pq} \in W$ **do**
    if $w_{pq} \geq \alpha$ **then**
      Add $X_t^p \to X_t^q$ to $\mathscr{G}$
      for $(\mathscr{X}_j^p, \mathscr{X}_j^q) \in \text{Hom}(\mathscr{X}_t^p, \mathscr{X}_t^q, \mathscr{G})$ **do** add edge $\mathscr{X}_j^p \to \mathscr{X}_j^q$ to $\mathscr{G}$
  for $i \in \{1, \cdots, \tau_{\max}\}$ **do**
    for $a_{pq} \in A_i$ **do**
      if $a_{pq} \geq \alpha$ **then**
        Add $X_{t-i}^p \to X_t^q$ to $\mathscr{G}$
        for $(\mathscr{X}_{j-i}^p, \mathscr{X}_j^q) \in \text{Hom}(\mathscr{X}_{t-i}^p, \mathscr{X}_t^q, \mathscr{G})$ **do** add edge $\mathscr{X}_{j-i}^p \to \mathscr{X}_j^q$ to $\mathscr{G}$
  Return window DAG $\mathscr{G}$

To reduce the computational complexity of inferring DPNs, Dojer (2006) showed that there exist, under three assumptions, polynomial algorithms to learn a globally optimal structure. Both the minimum description length and the DBe scores are studied in this work. Vinh et al. (2011) extend this approach to a score based on the mutual information test, leading to a model known as *GlobalMIT*. The first two assumptions considered by Dojer (2006) are related to the fact that the score can be decomposed across variables and rewritten as the sum of a term penalizing the complexity of the network and of a term explaining the data from the network. These are relatively mild assumptions which hold for different scores. The third assumption is stronger and states that the complexity term only depends on the cardinality of the set of parents for any variable. Unfortunately, this assumption is not always met in practice. Another important limitation of this approach lies in the fact that the degree of the polynomial controlling the complexity of the algorithm increases with the number of examples in the training set.

Several authors have proposed hybrid approaches that aim at combining constraint-based approaches and score-based approaches, the former providing relatively efficient algorithms while the latter providing scores on the inferred models and the possibility to directly orient pairs of nodes (Dash and Druzdzel (1999), Claassen and Heskes (2012), Jabbari et al. (2017)[13]). Malinsky and Spirtes (2018) also proposed a hybrid algorithm, called SVAR-GFCI, based on SVAR-FCI and on the score-based GES method (Chickering, 2002).

In this latter line and dealing with time series, Sanchez-Romero et al. (2019) made use of a variant of the PC-stable algorithm (Colombo & Maathuis, 2014), known as the Fast Adjacency Search stable (FAS-stable), to build a skeleton on which pairwise rules are used to orient edges. The overall process is referred to as FASK, for *Fast Adjacency Skewness*. The FAS-stable algorithm is an order independent adjacency search that avoids spurious connections between parents of variables. FAS-stable builds an undirected graph by iteratively testing conditional independencies; the BIC criterion is used by Sanchez-Romero et al. (2019) for this testing. The orientation of two adjacent nodes $\mathscr{X}^p$ and $\mathscr{X}^q$ in the graph obtained is then based on a score comparing the conditional correlation of $\mathscr{X}$ and $\mathscr{X}^q$ given $\mathscr{X}^p > 0$ with the one given $\mathscr{X}^q > 0$: if $\text{corr}(\mathscr{X}^p, \mathscr{X}^q | \mathscr{X}^p > 0) > \text{corr}(\mathscr{X}^p, \mathscr{X}^q | \mathscr{X}^q > 0)$, then $\mathscr{X}^p \to \mathscr{X}^q$; otherwise, $\mathscr{X}^q \to \mathscr{X}^p$. One important feature of FASK is its ability to identify cycles, especially 2-cycles in between two variables which are obtained when $\text{corr}(\mathscr{X}^p, \mathscr{X}^q | \mathscr{X}^p > 0) > \text{corr}(\mathscr{X}^p, \mathscr{X}^q)$ and $\text{corr}(\mathscr{X}^p, \mathscr{X}^q | \mathscr{X}^q > 0) > \text{corr}(\mathscr{X}^p, \mathscr{X}^q)$. Note that Sanchez-Romero et al. (2019) also introduced a hybrid algorithm, referred to as the Two-

13. Jabbari et al. (2017) provides a brief survey of these studies.

Section	Method	Causal graph	Faithfulness / Minimality	Causal Markov Condition	Instantaneous rel.	Lag > 1	Inference of self causes	Confounders	Inst. Hidden Conf	Lagged Hidden Conf.	Model based	Linear model	< 5 Hyper-parameters
3. Granger	PWGC*	S			✗	✓	✗	✗	✗	✗	✓	✓	✓
	MVGC *	S			✗	✓	✗	✓	✗	✗	✓	✓	✓
	TCDF*	W			✓	✓	✓	✓	✓	✗	✓	✗	✗
4. Constraint -based	PCMCI*	W	F	✓	✗	✓	✓	✓	✗	✗	✗	✗	✓
	oCSE*	S	F	✓	✗	✗	✓	✓	✗	✗	✗	✗	✓
	ANLTSM	W	F	✓	✓	✓	✓	✓	✓	✗	✓	✗	✓
	tsFCI*	W	F	✓	✗	✓	✓	✓	✗	✓	✗	✗	✓
	SVAR-FCI	W	F	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
5. Noise-based	VarLiNGAM*	W	M	✓	✓	✓	✓	✓	✗	✗	✓	✓	✓
	TiMINo*	S	M	✓	✓	✓	✗	✓	✗	✗	✓	✗	✓
6. Score-based	DYNOTEARS*	W			✓	✓	✓	✓	✗	✗	✓	✓	✓

Table 3: Summary of the main characteristics of representative algorithms in all the families discussed in this survey. Methods with * are also illustrated in the experimental section (Section 9). For causal graphs, S means that the method provides a summary causal graph whereas W means that the method provides a window causal graph; F corresponds to faithfulness and M to minimality.

Step algorithm, combining again the FAS-stable algorithm but this time with an identification of the causal relations based on independent component analysis.

7. Summary of the Main Methods Reviewed So Far

Table 3 displays the main characteristics of representative algorithms in all the families considered so far. The characteristics retained concern:

- The assumptions made on the graph: does a method infer a window causal graph? If not, the method directly infers a summary causal graph. Does the method rely on faithfulness, only minimality or none of the two? Lastly, does the method rely on the causal Markov condition?
- The type of relations inferred: can a method infer instantaneous relations? Can a method infer relations with a lag strictly greater than 1? Can a method infer self causes? Note that the methods which do not infer self causes assume that self causes always exist.
- The treatment of confounders[14]: can a method detect confounders? Can it detect instantaneous hidden confounders? Can it detect lagged hidden confounders? Instantaneous hidden confounders correspond to the situation where the two effects are instantaneous; the hidden confounder can either be instantaneous or lagged.
- The type of underlying models: we use here two main categories corresponding to whether a method relies on a model or not and whether the model is linear or not. In addition, we distinguish between methods that rely on few (less than 5) hyper-parameters and those that rely on more than 5 hyper-parameters. This latter characteristic is an indication of the complexity of fine-tuning a given method.

14. As mentioned in Section 2, confounders play a major role in the identification of causal relations.

As one can note, most methods infer a temporal causal graph. It is of course possible to deduce both window and summary causal graphs from a full time causal graph if the graph is consistent throughout time (Section 2). This said, methods that directly aim at inferring a summary causal graph may have advantage over methods that first infer a full time causal graph when considering the summary graph only. Indeed, the former methods can be faster and directly aim at solving a simpler problem. The distinction on the type of graphs inferred is thus not a way to rank causal discovery methods; it just reflects the fact that the objectives differ from one method to the other.

The detection of instantaneous relations is important from a practical point of view as the difference in time between two events associated to two time series may not be observed if the sampling frequencies of the time series are small. Roughly only half of the methods address this particular problem[15]. Being able to detect relations with a gap greater than 1 is also important in practical situations and only oCSE is restricted to a gap of 1. Methods that are not able to infer self causes usually assume that self causes always exist, which seems reasonable in real-life examples.

Most methods (with the exception of the traditional Granger method PWGC) can detect confounders. However, only three of them (TCDF, ANLTSM and SVAR-FCI) can detect instantaneous hidden confounders and only two (tsFCI and SVAR-FCI) can detect lagged hidden confounders. More generally, very few methods can deal with hidden variables, which violates the causal sufficiency assumption.

Regarding the type of underlying models, almost all methods rely on a particular model (except PCMCI and oCSE). Among the methods relying on a model, roughly half of them rely on a linear model. Concerning ANLTSM, if the underlying model considered is non-linear for observed variables, it is linear for hidden ones. Relying on a specific model can be an advantage when the data considered arises from a similar model. It can be of course a disadvantage when this is not the case. We illustrate this point in Section 9. Lastly, as one can note, most models use few (less than 5) hyper-parameters, with the exception of TCDF which is based on deep neural networks.

8. Other Approaches

We present in this section three other families of methods which differ from the previous ones by the type of data they use (discrete data for time series) or the type of models they rely on (dynamical systems for topology-based and difference-based methods).

8.1 Logic-Based Approaches

Another approach that has been explored is the one based on logical formulas, enabling inference of complex relationships. The most prominent framework is the one based on probabilistic computation tree logic (PCTL, Hansson & Jonsson, 1994) and its extension to numerical constraints, referred to as PCTLc, that expresses temporal properties over continuous and discrete variables. This line of research is exemplified in the work by S. Kleinberg (Kleinberg & Mishra, 2009; Kleinberg, 2011; Huang & Kleinberg, 2015).

To allow readers to easily relate our description to the original papers it is based on, we rely in this section on their notation. When continuous causes and effects are considered, one can consider that c plays the role of $\mathscr{X}^p > \theta_p$, where θ_p is a threshold on the values taken by $\mathscr{X}^p$, and e the role of $\mathscr{X}^q$.

Two types of (boolean) formulas are considered in PCTLc: state formulas that describe properties of individual states, and path formulas that describe properties along sequences of states. A particular relation in this formalism is the "leads-to" relation defined as:

$$c \rightarrow_{\geq p}^{\geq r, \leq s} e$$

where $[r,s]$ characterizes a window of time between c and e such as $1 \leq r \leq s \leq \infty$ and $r \neq \infty$. This relation states that if c is true, then e will become true in between r and s time units with at least

15. Note that the most recent version of PCMCI includes this possibility. We are discussing here the standard version.

probability p, where p is obtained by summing the probabilities of all paths from states where c is true to states where e is true in the $[r,s]$ time window. This relation can be extended to continuous effects by considering the expected value $\mathbb{E}[e]$ of e. For example,

$$c \rightarrow_{\geq p}^{\geq r,\leq s} [e > \mathbb{E}[e]]$$

states that if c is true, then e will be increased in between r and s time units with at least probability p. A similar relation for a possible decrease of e can of course be stated. As an illustration of this relation, one can consider the use of a drug (c is *true* when the drug is used, and *false* otherwise) and its effect resulting in a decrease of weight ($[e < \mathbb{E}[e]]$).

Potential causes (similar to the *prima facie* causes of Suppes, 1970) are then defined in a way reminiscent of the probability raising principle (Reichenbach, 1956; Suppes, 1970; Eells, 1991).

Definition 13 (Potential cause, Kleinberg, 2011) *When both c and e are formulas, c is a potential cause of e if the probability of c eventually occurring at some time is greater than zero, the probability of e is less than p and: $c \rightarrow_{\geq p}^{\geq r,\leq s} e$. When c is a formula and e is a continuous values variable taking values in $\mathbb{R}$, c is a potential cause of e if, with c being earlier than c: $\mathbb{E}(e \mid c) \neq \mathbb{E}(e)$, where the expectations are defined relative to time windows in which e occurs.*

Kleinberg (2011) further measures the significance of potential causes so as to retain only those causes deemed sufficiently significant for the effect. Let C denotes the set of potential causes of a continuous variable e. The causal significance of a potential cause c of e is measured by the difference of the conditional expectation of e when c is true and when c is false.

Definition 14 (Causal significance, Kleinberg, 2011) *A potential cause c of a continuous effect e is an ε-insignificant cause of e if $|\varepsilon_{avg}(c,e)| < \varepsilon$, where $\varepsilon_{avg}(c,e)$ is defined by:*

$$\varepsilon_{avg}(c,e) = \sum_{x \in C \setminus c} \frac{\mathbb{E}[e \mid c \wedge x] - \mathbb{E}[e \mid \neg c \wedge x]}{|C \setminus c|}.$$

A similar definition based on conditional probabilities is stated for effects that correspond to formulas. The overall approach to identify causes of an effect e is finally based on the identification of all potential causes of e using Definition 13, followed by the filtering of the potential causes deemed insignificant using Definition 14. Note that the complexity for computing ε_{avg} with d variables and T timepoints is $O(d^3T)$.

As an illustration, consider the toy example in Figure 14 from Kleinberg (2011) where $p(e)$ is uniform over the possible values of e. To determine if c is a potential cause of e in exactly one time unit, we first compute $\mathbb{E}[e \mid c] = (0+5+3.5+0)/4 = 2.125$ and $\mathbb{E}[e] = 1.9$. c increased the expected value of e which implies that c is a potential cause of e. To get the causal significance, assuming c is the only potential cause of e, we compute $\mathbb{E}[e \mid \neg c] = (2+2)/2 = 2$ therefore $\varepsilon_{avg}(c,e) = 2.125 - 2 = 0.125$. Assuming this value is greater than ε, one can state that c is a significant cause of e. An illustration of how this approach can be used on real data is given by Kleinberg (2011) with the data collected from Wharton Research Data Services (WRDS) that represents daily returns and the set of stocks in 2007.

The above framework was later extended by (Huang & Kleinberg, 2015) to obtain a faster procedure and overcome some of the problems associated with Definition 14. Indeed, there may be causes of e that occur only with $c \wedge x$ or $\neg c \wedge x$ so that the difference between $\mathbb{E}[e \mid c \wedge x]$ and $\mathbb{E}[e \mid \neg c \wedge x]$ may in practice not be large enough for c to be considered as a significant cause of e. One way to solve this problem is to compute the contribution to e that comes solely from c, leading to a new version of the causal significance measure of c for e, denoted $\alpha(c,e)$:

$$\alpha(c,e) = \frac{|T(e|c)|}{N(e|c)}(E[e|c \bigwedge_{x \in C \setminus c} \neg x] - E[e| \bigwedge_{x \in C} \neg x]).$$

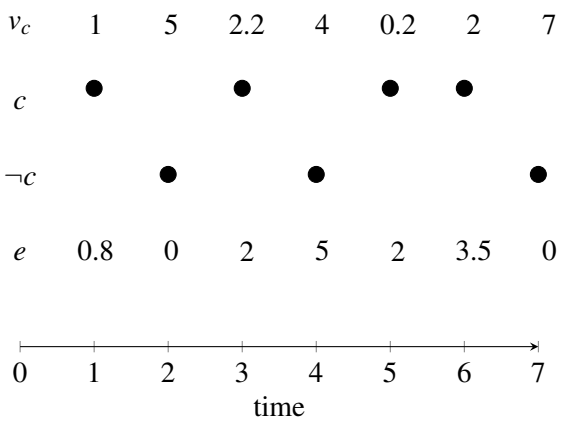


Figure 14: Toy example: v_c and e are two continuous variables. c and $\neg c$ are a discretization of v_c such as $c = v_c \leq 3$ and $\neg c = v_c > 3$

C is as before the set of potential causes of e, $|T(e|c)|$ is the number of unique timepoints where e is measured in window $[r,s]$ after each instance of c, and $N(e|c)$ is the total number of such timepoints. Let x_i and x_j be two elements of X, $|T(e)|$ be the number of time points where e occurs and let $f(e|x_i,x_j)$ be defined by:

$$f(e|x_i,x_j) = \frac{|T(e)| \times |T(e|c)|}{N(e|c) \times (|T(e)| - |T(e|c))}.$$

Huang and Kleinberg (2015) showed that when (a) the causal relationships are linear and additive (that is, the value of a variable at any time is given by the sum of the impact of its causes that are present plus a constant), (b) the causal relationships are deterministic and constant (that is, c's impact on e is the same every time c occurs), (c) the value of a variable when no cause is present is constant, (d) all genuine causes are measured, and (e) the matrix A defined by $A_{ij} = f(e|x_i,x_j)$ is of full rank, then $\alpha(c,e)$ is exactly the impact of c on e. Note that the complexity to compute $\alpha(c,e)$ is now $O(d^2T)$.

More recently, Bruto da Costa and Dasgupta (2021) formulated the problem of causal discovery as learning a causal sequence that explains a target effect E which is considered to be Boolean (or can be converted to Boolean). The sequences are learned using decision trees in which each path from the root to a leaf represents a causal sequence and each node a predicate, which is chosen on the basis of its utility in separating the cases where E is true from the cases where E is false.

8.2 Topology-Based Approaches

When considering a deterministic dynamic system (even a noisy one), Takens' theorem (Takens, 1981) states that the phase space can be reconstructed through time-delayed observations from the system, which implies that the effect should help in predicting the cause, given that it must in some way encode information about the cause. As the dynamic system is supposed to be deterministic, there is an underlying manifold that governs its dynamics.

Inspired by this idea, Sugihara et al. (2012) suggested a new method, called Convergent Cross Mapping (CCM) which tests for causality between $\mathscr{X}^p$ and $\mathscr{X}^q$ in the following sense:

Definition 15 (CCM causality, McCracken & Weigel, 2014) *Given two time series $\mathscr{X}^p$ and $\mathscr{X}^q$, we say that $\mathscr{X}^p$ CCM-causes $\mathscr{X}^q$ if $C(\mathscr{X}^q, \mathscr{X}^p) > C(\mathscr{X}^p, \mathscr{X}^q)$, where $C(\mathscr{X}^p, \mathscr{X}^q)$ is the squared Pearson correlation coefficient between the original time series $\mathscr{X}^p$ and an estimate of $\mathscr{X}^p$ made using its convergent cross-mapping with $\mathscr{X}^q$.*

As one can note, this method is grounded on dynamic system theory through the use of a convergent cross-mapping, and can be interpreted as: two variables are CCM causally linked if they share a common attractor manifold. However, it has been showed that CCM causality differs from true causality (McCracken & Weigel, 2014). Moreover, it is possible to arrive at the conclusion that both $\mathscr{X}^p$ and $\mathscr{X}^q$ are CCM causes of one another even though the true causal relation holds in only one direction (Ye, Deyle, Gilarranz, & Sugihara, 2015).

To overcome these problems, a variant of CCM, called Pairwise Asymmetric Inference (PAI), has been proposed by McCracken and Weigel (2014):

Definition 16 (PAI causality, McCracken & Weigel, 2014) *Given two time series $\mathscr{X}^p$ and $\mathscr{X}^q$ from the same attractor, we say that $\mathscr{X}^p$ PAI-causes $\mathscr{X}^q$ if $\tilde{C}(\mathscr{X}^p, \mathscr{X}^p\mathscr{X}^q) > \tilde{C}(\mathscr{X}^q, \mathscr{X}^q\mathscr{X}^p)$, where $\tilde{C}(\mathscr{X}^p, \mathscr{X}^p\mathscr{X}^q)$ is the squared Pearson correlation coefficient between the original time series $\mathscr{X}^p$ and an estimate of $\mathscr{X}^p$ made using its convergent cross-mapping with X^q and the past of X^p.*

The reader may have noticed the similarity with Granger's causality as *past values of $\mathscr{X}^p$ provide unique, statistically significant information about future values of $\mathscr{X}^q$*. it is nevertheless still grounded on dynamic systems through the use of the convergent cross-mapping.

Both CCM and PAI have been originally developed for bivariate analysis. In recent works, more variables have been taken into account: Feng et al. (2019) proposed a Bayesian version of CCM using deep Gaussian processes (DGPs), which are naturally connected with deep neural networks, whereas Leng et al. (2020) proposed an extension of CCM through Partial Cross Mapping, PCM, that is looking for conditional (in)dependencies.

8.3 Difference-Based Approaches

Difference-Based Causal Models (DBCMs) is a class of discrete-time dynamic models which represent all causal relations across time by means of difference equations driving changes in the system. This means that all causal relations across time are due to a derivative causing a change in its integral (cross-temporal restriction). Difference equations are supposed not to vary across time. DBCMs are defined as follows.

Definition 17 (Difference-Based Causal Model, Voortman et al., 2010) *A DBCM is a structural equation model in which the set of variables is given by time series (evaluated on the first two time points due to consistency thorough time of difference equations), and the set of equations is such that there exists a cross-temporal parent of some variable if and only if the corresponding equation is the integral equation for this variable.*

A method, named Difference-Based Causality Learner (DBCL), has been developed in Voortman et al. (2010) to learn DBCMs which relies on faithfulness (this implies that the model does not reach equilibrium). DBCL first finds relevant latent derivatives, computed by differences of variables, and then learns the contemporaneous structure using any correct causal discovery algorithm under causally sufficient data.

8.4 Drawbacks and Conclusion

The logic-based approach presented in Section 8.1 is interesting for inferring causes of effects that can be either continuous or discrete (note that causes are always discrete). However, there is no guarantee that the graph obtained with the cause-effect relations is in the equivalence class of the causal graph underlying the observations. There is also no simple way to deal with latent variables in this approach. In addition, the discretization of continuous variables for identifying causes is a limiting factor as this process needs to rely on background knowledge provided by experts (Malinsky & Danks, 2018). Topology-based methods presented in Section 8.2 are interesting when considering deterministic dynamic systems. They however aim at discovering specific correlation

and the concepts of CMM and PAI causality they rely on render them not truly causal. For the sake of completeness, we also discussed the difference-based approach (Section 8.3), even if this approach has not been the subject of many studies in the causality literature. Furthermore it is not, to our knowledge, widely used in practice.

For all these reasons, we do not include these methods in our experimental comparison.

9. Experimental Illustration

We present in this section an experimental comparison of the major causal discovery methods we have reviewed. To do so, we first describe the selected evaluation measures and discuss the retained methods as well as the artificial datasets corresponding to basic causal structures and the standard real world benchmark we have considered. We then present the results of all experiments.

9.1 Evaluation Measures

Among all existing evaluation metrics to assess the quality of causal inference, as Structural Hamming Distance (Peters & Bühlmann, 2015) or Frobenius norm (Shimizu et al., 2011), we use the standard *F1-score*, referred to as F1 and defined by: $\mathrm{F1} = 2\mathrm{TP}/\left(2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}\right)$, where TP, FP and FN respectively correspond to true positives, false positives and false negatives. This score can be used to assess both the quality of the skeleton of the causal graph obtained and the quality of the causal graph itself. In his latter case, we refer to this score as $\overrightarrow{\mathrm{F1}}$ to emphasize the fact that the orientation of the edges is taken into account when comparing to a gold standard. Furthermore, the F1 score is based on both precision and recall which measure different characteristics of a system. We provide results in terms of precision and recall in the Supplementary Material.

9.2 Evaluated Methods

From the Granger family (Section 3), we retain the pairwise and multivariate methods (referred to as PWGC and MVGC). The full model is compared to the restricted model using an F-test. We rely on our implementation of PWGC and use for MVGC the code available at `http://www.sussex.ac.uk/sackler/mvgc/`. In addition, we include in our comparison TCDF and rely for this method on the implementation available at `https://github.com/M-Nauta/TCDF`. For the hyper-parameters in this latter method, we used the values suggested by the authors: a kernel of size 4, a dilation coefficient equal to 4, 1 hidden layer, a learning rate of 0.01, and 5000 epochs.

From the constraint-based family (Section 4), we retain PCMCI using both partial correlation and mutual information to measure independence. Both scores are available in the implementation provided at `https://github.com/jakobrunge/tigramite`. We also include oCSE, which we implemented. In all those methods, mutual information is estimated using the k-Nearest Neighbour method (Runge, 2018) and a permutation test is used to assess whether the mutual information scores are significantly different from 0 or not. Finally, we also consider tsFCI, with the implementation provided at `https://sites.google.com/site/dorisentner/publications/tsfci`, in which independence and conditional independence are tested respectively with tests of zero correlation and zero partial correlation.

Among the noise-based approaches (Section 5), we retain 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 directly use the value of the statistics. TiMINo uses a test based on cross-correlation that can be derived from Brockwell and Davis (1986, Thm 11.2.3.).

We have retained the most recent score-based method, namely DYNOTEARS (Pamfil et al., 2020) available at `https://github.com/quantumblacklabs/causalnex`, the hyperparameters of which are set to their recommended values ($\lambda_W = \lambda_A = 0.05$ and $\alpha_W = \alpha_A = 0.01$).

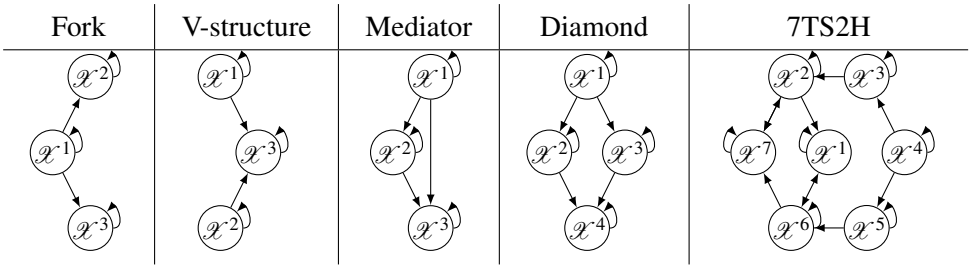


Table 4: Structures corresponding to the artificial datasets. $A \to B$ means that A causes B and $A \longleftrightarrow$ B represents the existence of a hidden common cause between A and B.

For all the methods, when doing a statistical test, we use a significance level of 0.05. Our implementations of PWGC and oCSE are available at `https://github.com/ckassaad/causal_discovery_for_time_series`; furthermore, all methods can be used through a Python routine available at `https://github.com/ckassaad/causal_discovery_for_time_series`.

9.3 Datasets

The artificial datasets, available at `https://dataverse.harvard.edu/dataverse/basic_causal_structures_additive_noise`, correspond to five basic causal structures presented in Table 4: fork, v-structure, mediator, diamond, as well as to a nine nodes structure introduced by Spirtes et al. (2001) and referred to as 7ts2h. In 7ts2h, seven nodes correspond to observational time series and two to hidden common causes, represented by double arrows. The underlying generating process is based on nonlinear functions between time series and linear relations for self causation, as defined below:

$$\forall q,\ \mathscr{X}_0^q = 0;\ \forall t > 0,\ \mathscr{X}_t^q = a_{t-1}^{qq}\mathscr{X}_{t-1}^q + \sum_{\substack{(p,\gamma) \\ \mathscr{X}_{t-\gamma}^p \in Par(\mathscr{X}_t^q)}} a_{t-\gamma}^{pq} f(\mathscr{X}_{t-\gamma}^p) + 0.1\xi_t^q, \tag{3}$$

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

To evaluate the performance of the inference with respect to the length of the time series, we make the length vary from 125 to 4000 time points. For each structure and each length, we generate 10 different datasets over which the performance of each method is averaged.

The real-world benchmark we have retained here is FMRI (Functional Magnetic Resonance Imaging) which contains BOLD (Blood-oxygen-level dependent) datasets for 28 different underlying brain networks (Smith et al., 2011)[16]. Each dataset contains the neural activity, based on the change of blood flow, of at most 50 different regions. Each region corresponds to a time series which contains between 50 and 5000 time points. Since not all the methods retained can handle more than a few times series, we excluded the larger dataset and make use here of the 27 datasets that contain at most 15 time series. Note that these datasets are considered causally sufficient.

9.4 Numerical Results

Here we assess how the retained methods behave on the artificial datasets corresponding to basic causal structures and on the FMRI benchmark.

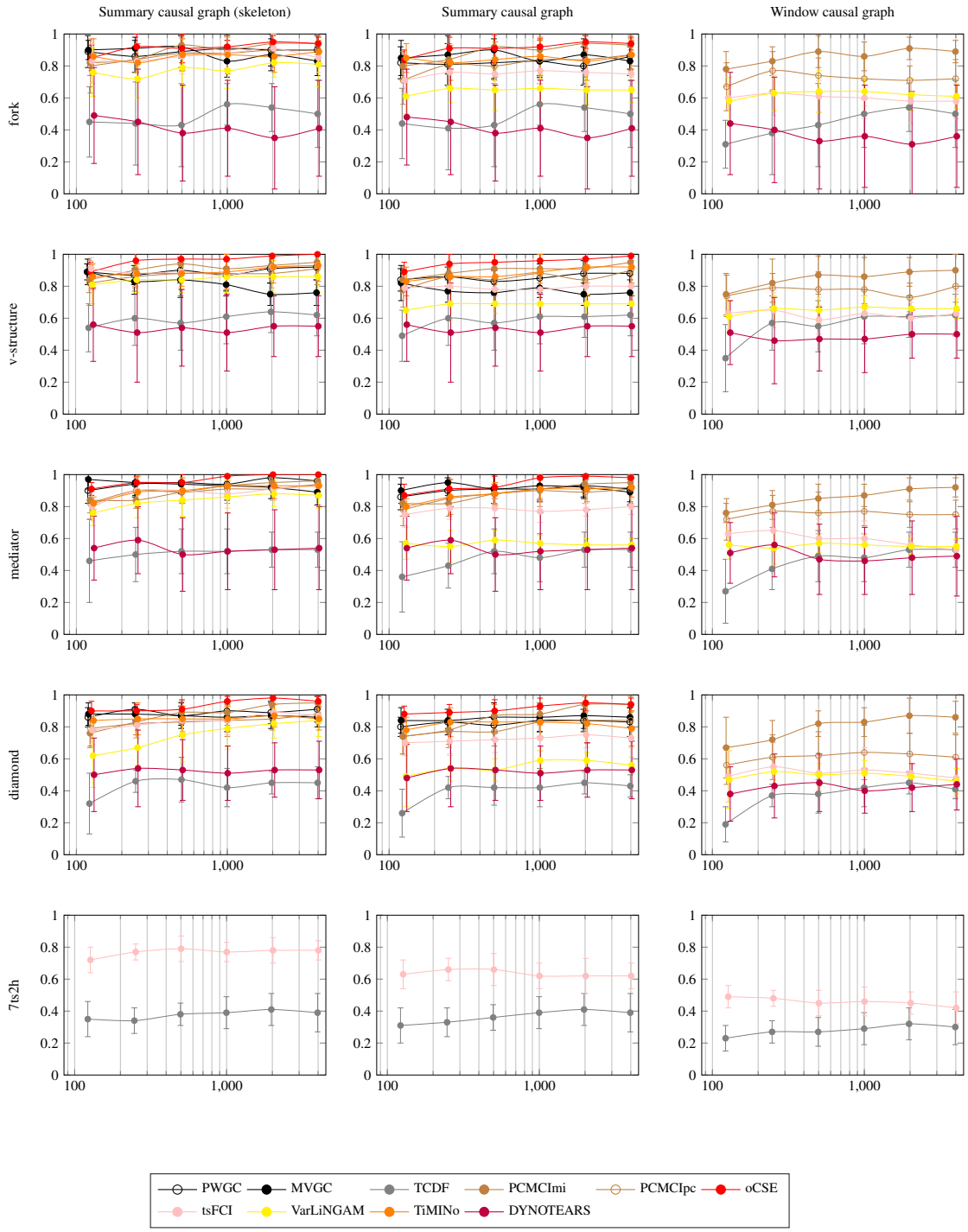


Figure 15: Performance of all methods on the 5 artificial datasets. The results are computed over 10 runs for which we report the mean ($\pm$ the standard deviation) of the F1 score without taking into account the orientation of edges for the skeleton of the summary causal graph (left column), and while taking into account the orientation of the edges for the summary (middle column) and window (right column) causal graphs. The results are computed for various lengths of the time series: 125, 250, 500, 1000, 2000 and 4000 time points (a log-scale is used for the x-axis).

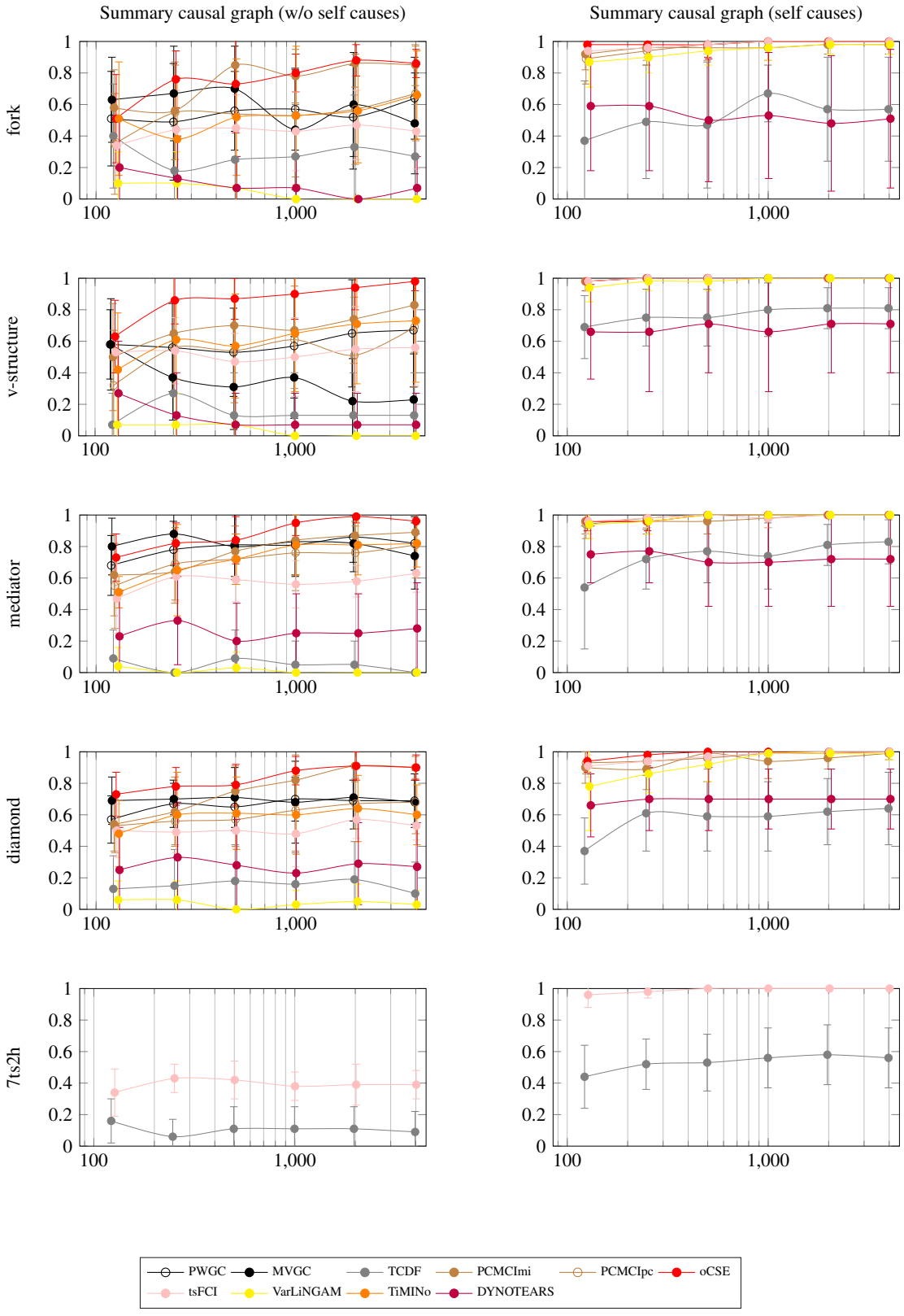


Figure 16: Performance of all methods on the 5 artificial datasets for the inference of the summary causal graph with two settings: excluding self causes (left column) and focusing only on self causes (right column). The results are computed over 10 runs for which we report the mean ($\pm$ the standard deviation) of the F1 score taking into account the orientation of the edges. The results are computed for various lengths of the time series: 125, 250, 500, 1000, 2000 and 4000 time points (a log-scale is used for the x-axis).

9.4.1 ARTIFICIAL DATASETS

We are interested here in assessing the capacity of each method to identify correct summary and window causal graphs. The overall results obtained by the different methods are given in Figure 15, for the skeleton of the summary causal graph (right column), the summary causal graph (middle column) and the window causal graph (right column). Note that the orientation in the window causal graph relies (except for instantaneous relations) on temporal information so that the accuracy on the skeleton of the window causal graph and on the window causal graph itself are almost identical. For this reason, we solely present here the results on the window causal graph. The first four datasets (fork, v-structure, mediator and diamond) are causally sufficient whereas the last one, 7ts2h, is not as it contains two hidden variables. Only two methods, tsFCI and TCDF, are able to deal with this latter dataset. They can infer both a summary and a window causal graph. Among all methods, only five, namely TCDF, PCMCI (mi and pc), tsFCI and VarLiNGAM, aim at building both a summary and a window causal graph. The remaining 5 methods, PWGC, MVGC, oCSE, TiMINo and DYNOTEARS, directly infer a summary causal graph.

As one can note from Figure 15, for all methods and all datasets, the performance on the (skeleton of the) summary causal graph slightly increases with the number of time points considered and reaches a plateau after roughly 250 time points have been considered. We attribute this to the fact that the summary causal graph is a relatively simple structure in this case that can be inferred without much information. The increase with the number of time points considered is more important for the window causal graph which requires a certain amount of information for the inference to be correct.

For all the causally sufficient structures, the performance obtained on the summary causal graph (for both its skeleton and graph itself) is high (in between 0.8 and 1 in terms of F1 score) and comparable for all methods, except for VarLiNGAM, TCDF and DYNOTEARS which do not perform as well as the other methods. This is not really surprising for VarLiNGAM and DYNOTEARS as both methods are based on a linear assumption whereas the generation process of the datasets considered is based on non-linear relations between different time series. Surprisingly, both the Granger pairwise method and its multivariate extension have good performance, whereas they do not aim at inferring true causality by definition. The best performing method overall on the summary graph is oCSE, but the difference with other methods is not significant (with respect to a two sided t-test with level 0.01). The results obtained on the window causal graph are lower than the ones obtained on the summary causal graph and the difference between the methods are more marked. This is not really surprising as the former graph is more complex than the latter one. The best performing method here is PCMCI, and in particular the version based on mutual information (mi). Note however that this version is more computationally demanding (Runge et al., 2019) than the one based on partial correlation (pc). Lastly, for 7ts2h, a dataset with hidden confounders, the applicable methods have difficulties in identifying both the summary and the window causal graphs. The problem is definitely more complex and no satisfying solution has been proposed yet, even though TiMINo reaches 0.6 and 0.5 in terms of F1 score on the summary causal graph and the window causal graph respectively.

Self causes Among the methods we have reviewed, some (PWGC, MVGC, TiMINo) assume that a time series always causes itself, which seems a reasonable assumption for time series, whereas others (DYNOTEARS, oCSE, PCMCI, TCDF, tsFCI and VarLiNGAM) do not make such an assumption and try to infer self causes as any other causes, which is more difficult *a priori*. As all the artificial datasets we have considered contain self-causes (which represent roughly 50% of the causal relations on each dataset), methods of the first type have an advantage over the ones of the second type. To further compare all methods, we have examined the performance of each method in two cases: one in which self causes are excluded and one in which only self causes are considered.

16. Original data: `https://www.fmrib.ox.ac.uk/datasets/netsim/index.html`
Preprocessed version: `https://github.com/M-Nauta/TCDF/tree/master/data/fMRI`

The results obtained are displayed in Figure 16. As one can note, the performance of all methods when excluding self causes are lower than when considering all causes and the differences between the methods are more important. oCSE outperforms, with a larger margin than before, all other methods on the causally sufficient structures. In addition, VarLiNGAM obtains poor results which makes sense as the relations between different time series in the datasets retained are not linear. On self causes only, all methods but DYNOTEARS and TCDF perform very well and make no mistake when the number of time points is sufficient. As mentioned before, an important difference between self causes and causes between different time series is that the former are linear whereas the latter are not (and are thus more difficult to identify). This explains the difference in performance between the left and right columns of Figure 16 as well as the fact that VarLiNGAM, which relies on a linear model, behaves well on self causes.

Limit cases We conclude this comparison by considering two limit cases, one in which the Markov equivalence class contains more than one graph and one in which the faithfulness assumption is no longer valid. For the first case, we generate a new fork structure in which all relations are instantaneous, so that one cannot differentiate between common and intermediate causes (by definition, the fork structure does not contain any collider). For the second case, we generate new mediator and diamond structures in which all relations are linear with coefficients set in such a way that different causal paths eliminate each other. This is obtained by setting, in Eq. (3), $a^{13} = -a^{12}a^{23}$ for mediator, following Zhalama et al. (2016), and $a^{34} = -a^{12}a^{23}/a^{13}$ for diamond. In each case, we exclude self causes and simulate 10 datasets each with 1000 time points.

	Fork (Markov equi.)		Mediator (unfaith.)		Diamond (unfaith.)	
	F1	$\overrightarrow{\text{F1}}$	F1	$\overrightarrow{\text{F1}}$	F1	$\overrightarrow{\text{F1}}$
PWGC	0.12 ± 0.24	0.05 ± 0.15	0.28 ± 0.37	0.12 ± 0.27	0.32 ± 0.28	0.14 ± 0.23
MVGC	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
PCMCImi	0.28 ± 0.29	0.07 ± 0.019	0.27 ± 0.29	0.05 ± 0.15	0.41 ± 0.25	0.20 ± 0.22
PCMCIpc	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
tsFCI	$\mathbf{0.71} \pm 0.29$	0.44 ± 0.17	0.88 ± 0.09	0.55 ± 0.04	0.86 ± 0.03	0.55 ± 0.03
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	$\mathbf{0.45} \pm 0.15$	$\mathbf{0.95} \pm 0.15$	$\mathbf{0.64} \pm 0.08$	0.78 ± 0.06	$\mathbf{0.49} \pm 0.03$
DYNOTEARS	0.00 ± 0.00	0.00 ± 0.00	0.41 ± 0.29	0.37 ± 0.26	0.29 ± 0.33	0.29 ± 0.33

Table 5: Results obtained on the two limit cases corresponding to a rich Markov equivalence class and to unfaithful data. The first limit case is evaluated on the fork structure whereas the second one is evaluated on both the mediator and diamond structures. We report the mean and the standard deviation (over 10 runs) of the oriented and non-oriented F1 scores. Best results are in bold and methods are grouped according to their family (Granger, constraint-based, noise-based, score-based).

The results (averaged over 10 runs) obtained by the different methods on these two limit cases are displayed in Table 5. As one can see, the results differ from the ones obtained before and methods that rely strongly on faithfulness and causal sufficiency (as PCMCI and oCSE) now perform poorly even though they were among the best methods before. Furthermore, all methods have difficulties in orienting edges as the $\overrightarrow{\text{F1}}$ score is systematically lower than the F1 score, the drop being particularly important for most methods on the unfaithful datasets. The best method overall for the $\overrightarrow{\text{F1}}$ score is TiMINo. For the F1 score, the best method on fork is tsFCI, followed by TiMINo, and VarLiNGAM on the two unfaithful datasets, followed by either TiMINo or tsFCI. The good behaviour of TiMiNo, and to a certain extent of TCDF which obtains consistent, good results on all datasets for both measures, can be explained by the fact that these methods are not restricted to the Markov equivalence class and do not rely on the faithfulness assumption. All in all, this experiment

confirms that different methods are adpated to different datasets according to the assumptions they rely on.

9.4.2 REAL DATASET: FMRI

The results obtained for all methods on the real dataset FMRI are displayed in Table 6. In order to compare all methods, we focus here on the summary graph, evaluating as before the capacity of each method to retrieve a causal relation between two time series (*i.e.* to obtain a correct skeleton) and to orient such relations. We also evaluate the method through their capacity to detect causal relations between different time series (w/o self causes) and within a time series (self causes only).

All methods but TCDF are able to retrieve the skeleton of the summary causal graph, tsFCI and VarLiNGAM being the two best methods here, reaching a F1 score above 0.8. For the summary causal graph, VarLiNGAM is by far the best method which may suggest that the relations between time series in FMRI may be well approximated by linear relations. Apart from that, the different families of approaches perform similarly (even though some methods are worse than others in the different families, as TCDF here). The same conclusion can be drawn when only causal relations between different times series are considered. When considering self causes only (which amount to roughly 50% of all causal relations) one can see that all methods perform very well (the methods with a *1.00* in italics always assume self causes) except TCDF. Lastly, one can see that the standard deviations vary from one method to the other. They are more important without self causes than when considering self causes only and similar for the skeleton and the summary causal graph.

All in all, all families of approaches obtain more or less the same results on the summary causal graph, with an advantage to the noise-based method VarLiNGAM on this dataset. As mentioned before, this method relies on certain assumptions that seem to be appropriate for the FMRI dataset considered here.

	Sum. graph (skel.)	Sum. graph	Sum. graph (details)	
			W/o self causes	Self causes only
PWGC	0.74 ± 0.08	0.63 ± 0.08	0.31 ± 0.17	*1.00* $\pm$ *0.00*
MVGC	0.76 ± 0.09	0.59 ± 0.11	0.24 ± 0.18	*1.00* $\pm$ *0.00*
TCDF	0.33 ± 0.25	0.30 ± 0.22	0.07 ± 0.13	0.47 ± 0.35
PCMCImi	0.67 ± 0.14	0.59 ± 0.13	0.22 ± 0.19	0.90 ± 0.19
PCMCIpc	0.72 ± 0.11	0.64 ± 0.12	0.29 ± 0.20	0.96 ± 0.13
oCSE	0.68 ± 0.084	0.63 ± 0.07	0.16 ± 0.20	0.91 ± 0.14
tsFCI	0.80 ± 0.09	0.60 ± 0.10	0.44 ± 0.10	**0.97** ± 0.09
VarLiNGAM	**0.84** ± 0.16	**0.71** ± 0.17	**0.49** ± 0.28	0.92 ± 0.22
TiMINo	0.75 ± 0.13	0.56 ± 0.12	0.32 ± 0.11	*1.00* $\pm$ *0.00*
DYNOTEARS	0.77 ± 0.12	0.58 ± 0.12	0.38 ± 0.15	0.97 ± 0.12

Table 6: Results for FMRI in terms of the F1 score (mean $\pm$ standard deviation) averaged over the 27 networks of this dataset. For the skeleton of the summary causal graph (Sum. caus. graph (skel.)), the orientation of the edges is not taken into account when computing the F1 score. The third column (Sum. caus. graph (details)) illustrates the capacity of the methods to detect causal relations between different time series (W/o self causes) and within a time series (Self causes only). In this latter case, a *1.00* in italics indicates that the method assumes that self causes always exist. Best results are in bold and methods are grouped according to their family (Granger, constraint-based, noise-based, score-based).

10. Conclusion

We have presented in this survey different methods, pertaining to different families of approaches, for causal discovery in time series. We furthermore have illustrated their behaviour through ex-

periments conducted on both artificial and real datasets for inferring either a window causal graph or a summary causal graph. The families we have retained correspond to approaches *à la* Granger, constraint-based approaches, noise-based approaches, score-based approaches logic-based approaches, topology-based approaches, and difference-based approaches.

The main conclusions one can draw from this survey is that causal discovery in times series is an active research field in which new methods (in every family of approaches) are regularly proposed, and that no family or method stands out in all situations. Indeed, they all rely on assumptions that may or may not be appropriate for a particular dataset. Constraint-based and noise-based methods often come with theoretical guarantees on their optimality. If this is clearly an interesting feature, these guarantees also rely on assumptions which are not always met in practice.

Several extensions to the methods we have presented have been, and still are, investigated. For example, Gong et al. (2015) considers the problem of subsampling which amounts to recover relations between time instants that were not observed as their difference is smaller than the sampling rate of the time series. Hyttinen et al. (2017) further studies the subsampling in the context of time series with hidden variables. Gong et al. (2017) studies methods to infer causal relations on time series which correspond to aggregate (local averages or sums of observations) of other time series. Both Zhang et al. (2017) and Huang et al. (2019) address the problem of causal discovery and forecasting on non-stationary time series. Lastly, the problem of time series with different sampling rates has partially been explored by Mogensen et al. (2018) through the consideration of continuous time series. This is an important problem in practice that remains largely unexplored.

If causal discovery is an important aspect of the research conducted on causality in time series, causal reasoning on causal graphs certainly opens the door to practical applications beyond the reach of current tools. The reliance of causal reasoning on causal graphs explains the importance of causal discovery and our focus on this aspect in this survey. As we have seen, causal discovery in time series is a difficult problem and, facing the performance of the methods so far developed, we believe that a promising approach is to have experts interact with causal discovery tools to infer causal graphs that can then be used for reasoning and problem solving.

Appendix A. Additional Experimental Results

We provide in this section results in terms of precision and recall for the experiments described in the main paper.

As one can note from Figure 17 and 18, for all methods and all datasets, the performance on the skeleton of the summary causal graph slightly increases with the number of time points considered and reaches a plateau after roughly 250 time points have been considered. Focusing on precision, we observe a decrease in performance for VarLiNGAM and tsFCI when considering the oriented graphs. This is not true for recall, which increases with the number of time points. For all the causally sufficient structures, the performance obtained on the summary causal graph (for both its non oriented and oriented versions) is particularly high (in between 0.8 and 1) in terms of precision and comparable for all methods, TCDF and DYNOTEARS being slightly lower. The performance obtained for the summary causal graph (skeleton and oriented) and the window causal graph in terms of recall is a bit lower, but still on a good range for most of the methods (except for TCDF and VarLiNGAM and DYNOTEARS). On can in particular notice that tsFCI is one of the best methods in terms of recall, whereas it was one of the worst in terms of precision.

The results obtained for all methods on the real dataset FMRI are displayed in Table 7. In order to compare all methods, we focus here on the summary graph, evaluating as before the capacity of each method to retrieve a causal relation between two time series (*i.e.*, to obtain a correct skeleton) and to orient such relations (providing the summary causal graph). We also evaluate the method through their capacity to detect causal relations between different time series (w/o self causes) and within a time series (self causes only). As one can see, VarLiNGAM behaves quite well in terms of precision, the difference with some constraint-based methods as PCMCImi being however small.

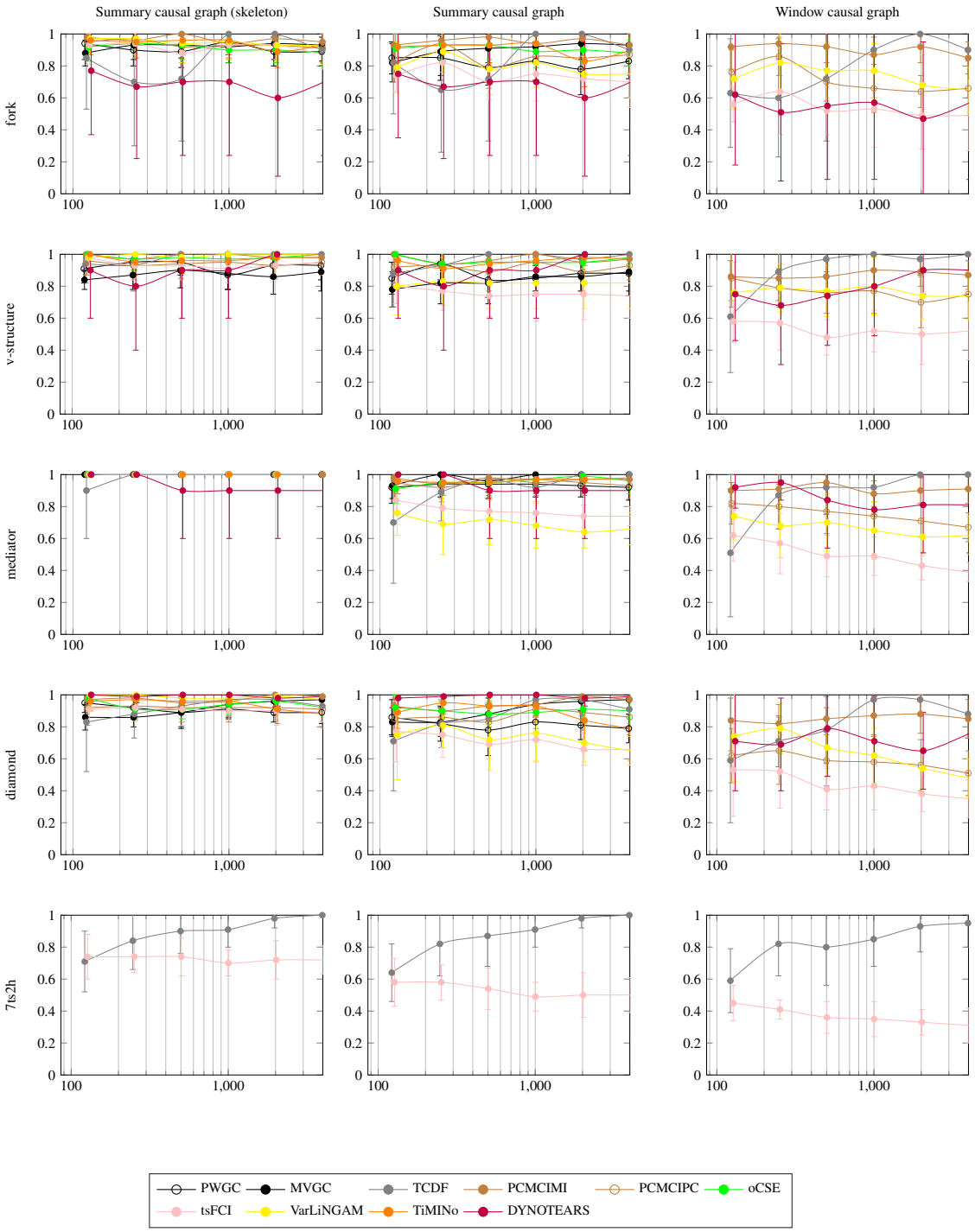


Figure 17: Performance of all methods on the 5 artificial datasets. The results are computed over 10 runs for which we report the mean ($\pm$ the standard deviation) of the precision without taking into account the orientation of edges for the skeleton of the summary causal graph (left column), and while taking into account the orientation of the edges for the summary (middle column) and window (right column) causal graphs. The results are computed for various lengths of the time series: 125, 250, 500, 1000, 2000 and 4000 time points (a log-scale is used for the x-axis).

In termes of recall, tsFCI and DYNOTEARS obtain very good results (above 0.9), closely followed by TiMINo. The three families of methods, constraint-based, noise-based and score-based, are thus well represented in this case.

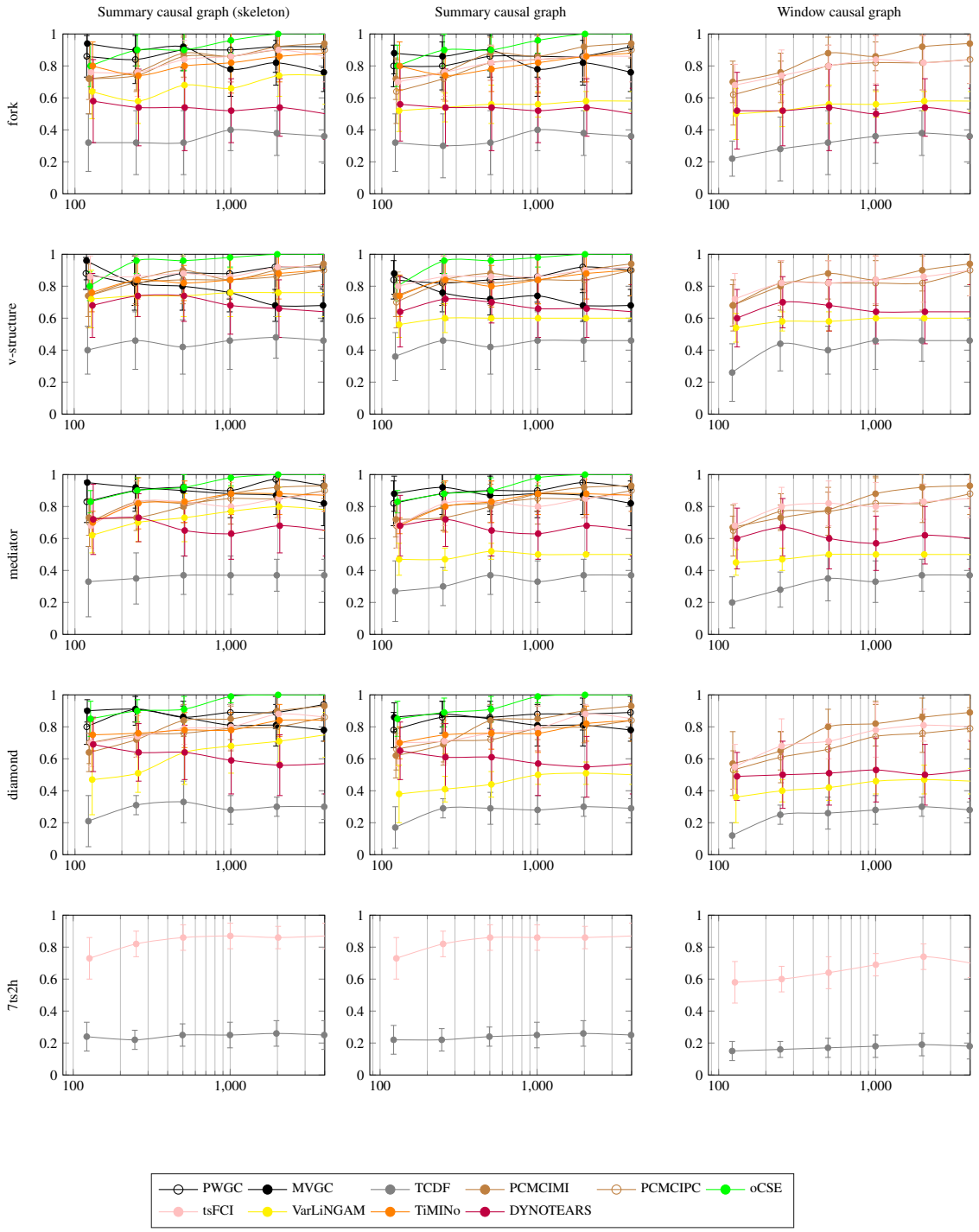


Figure 18: Performance of all methods on the 5 artificial datasets. The results are computed over 10 runs for which we report the mean ($\pm$ the standard deviation) of the recall without taking into account the orientation of edges for the skeleton of the summary causal graph (left column), and while taking into account the orientation of the edges for the summary (middle column) and window (right column) causal graphs. The results are computed for various lengths of the time series: 125, 250, 500, 1000, 2000 and 4000 time points (a log-scale is used for the x-axis).

Finally, it is interesting to point out that some methods are more precision-oriented, as they detect few but relevant relations (TCDF, VarLiNGAM), whereas other methods are more recall-oriented and focus on the detection of all relevant relations (tsFCI, DYNOTEARS).

Precision	Sum. graph (skel.)	Sum. graph	Sum. graph (details)	
			W/o self causes	Self causes only
PWGC	0.75 ± 0.10	0.62 ± 0.16	0.31 ± 0.24	*1.00* $\pm$ *0.00*
MVGC	0.73 ± 0.17	0.57 ± 0.18	0.22 ± 0.17	*1.00* $\pm$ *0.00*
TCDF	0.73 ± 0.41	0.68 ± 0.41	0.14 ± 0.28	0.70 ± 0.46
PCMCImi	0.84 ± 0.14	0.73 ± 0.16	0.32 ± 0.29	**1.0** ± 0.0
PCMCIpc	0.80 ± 0.13	0.61 ± 0.14	0.37 ± 0.28	**1.0** ± 0.0
oCSE	0.82 ± 0.12	**0.75** ± 0.16	0.17 ± 0.21	**1.0** ± 0.0
tsFCI	0.72 ± 0.14	0.46 ± 0.13	0.30 ± 0.09	**1.0** ± 0.0
VarLiNGAM	**0.90** ± 0.11	0.73 ± 0.17	**0.48** ± 0.27	0.96 ± 0.18
TiMINo	0.71 ± 0.21	0.49 ± 0.22	0.28 ± 0.18	*1.00* $\pm$ *0.00*
DYNOTEARS	0.63 ± 0.17	0.40 ± 0.15	0.21 ± 0.09	**1.0** ± 0.0
Recall	Sum. graph (skel.)	Sum. graph	Sum. graph (details)	
			W/o self causes	Self causes only
PWGC	0.77 ± 0.17	0.71 ± 0.16	0.44 ± 0.32	*1.00* $\pm$ *0.00*
MVGC	0.86 ± 0.14	0.65 ± 0.12	0.32 ± 0.23	*1.00* $\pm$ *0.00*
TCDF	0.24 ± 0.21	0.22 ± 0.18	0.06 ± 0.12	0.38 ± 0.33
PCMCImi	0.58 ± 0.17	0.52 ± 0.15	0.19 ± 0.18	0.86 ± 0.25
PCMCIpc	0.68 ± 0.15	0.61 ± 0.14	0.29 ± 0.23	0.95 ± 0.17
oCSE	0.62 ± 0.16	0.58 ± 0.15	0.19 ± 0.29	0.86 ± 0.2
tsFCI	0.95 ± 0.10	0.94 ± 0.10	**0.92** ± 0.17	0.96 ± 0.13
VarLiNGAM	0.84 ± 0.21	0.72 ± 0.22	0.56 ± 0.35	0.90 ± 0.24
TiMINo	0.89 ± 0.16	0.84 ± 0.19	0.70 ± 0.36	*1.00* $\pm$ *0.00*
DYNOTEARS	**0.97** ± 0.11	**0.95** ± 0.13	0.91 ± 0.26	**0.99** ± 0.03

Table 7: Results for FMRI in terms of precision and recall (mean $\pm$ standard deviation) averaged over the 27 networks of this dataset. For the skeleton of the summary causal graph (Sum. caus. graph (skel.)), the orientation of the edges is not taken into account when computing the measures. The third column (Sum. caus. graph (details)) illustrates the capacity of the methods to detect causal relations between different time series (W/o self causes) and within a time series (Self causes only). In this latter case, a 1.00 in italics indicates that the method assumes that self causes always exist. Best results are in bold and methods are grouped according to their family (Granger, constraint-based, noise-based, score-based).

References

Ali, A. R., Richardson, T. S., Spirtes, P., & Zhang, J. (2005). 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, pp. 10–17, Arlington, Virginia, USA. AUAI Press.

Ancona, N., Marinazzo, D., & Stramaglia, S. (2004). Radial basis function approach to nonlinear granger causality of time series. *Physical Review E*, *70*, 056221.

Andersson, S. A., Madigan, D., & Perlman, M. D. (1997). A characterization of markov equivalence classes for acyclic digraphs. *Annals of Statistics*, *25*(2), 505–541.

Barrett, A. B., Barnett, L. C., & Seth, A. K. (2010). Multivariate granger causality and generalized variance. *Physical review E*, *81*, 041907.

Bell, D., Kay, J., & Malley, J. (1996). A non-parametric approach to non-linear causality testing. *Economics Letters*, *51*(1), 7 – 18.

Blom, T., Bongers, S., & Mooij, J. M. (2019). Beyond structural causal models: Causal constraints models. In Globerson, A., & Silva, R. (Eds.), *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2019, Tel Aviv, Israel, July 22-25, 2019*.

Brockwell, P. J., & Davis, R. A. (1986). *Time Series: Theory and Methods*. Springer-Verlag, Berlin, Heidelberg.

Brovelli, A., Ding, M., Ledberg, A., Chen, Y., Nakamura, R., & Bressler, S. L. (2004). 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.

Bruto da Costa, A. A., & Dasgupta, P. (2021). Learning temporal causal sequence relationships from real-time time-series. *J. Artif. Int. Res.*, *70*, 205–243.

Casile, A., Faghih, R. T., & Brown, E. N. (2021). Robust point-process granger causality analysis in presence of exogenous temporal modulations and trial-by-trial variability in spike trains. *PLOS Computational Biology*, *17*(1), 1–22.

Chen, Y., Rangarajan, G., Feng, J., & Ding, M. (2004). Analyzing multiple nonlinear time series with extended granger causality. *Physics Letters A*, *324*, 26–35.

Chickering, D. M. (1995). Learning bayesian networks is np-complete. In Fisher, D., & Lenz, H. (Eds.), *Learning from Data - Fifth International Workshop on Artificial Intelligence and Statistics, AISTATS 1995, Key West, Florida, USA, January, 1995. Proceedings*, pp. 121–130. Springer.

Chickering, D. M. (2002). Learning equivalence classes of bayesian-network structures. *Journal of Machine Learning Research*, *2*, 445–498.

Chu, T., & Glymour, C. (2008). Search for additive nonlinear time series causal models. *Journal of Machine Learning Research*, *9*, 967–991.

Claassen, T., & Heskes, T. (2012). A bayesian approach to constraint based causal inference. In de Freitas, N., & Murphy, K. P. (Eds.), *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence, Catalina Island, CA, USA, August 14-18, 2012*, pp. 207–216. AUAI Press.

Climenhaga, N., DesAutels, L., & Ramsey, G. (2019). Causal inference from noise. *Noûs*.

Colombo, D., & Maathuis, M. H. (2014). Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, *15*(116), 3921–3962.

Colombo, D., Maathuis, M. H., Kalisch, M., & Richardson, T. S. (2012). Learning high-dimensional directed acyclic graphs with latent and selection variables. *Annals of Statistics*, *40*(1), 294–321.

Dash, D., & Druzdzel, M. J. (1999). A hybrid anytime algorithm for the construction of causal models from sparse data. In Laskey, K. B., & Prade, H. (Eds.), *UAI '99: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, Stockholm, Sweden, July 30 - August 1, 1999*, pp. 142–149. Morgan Kaufmann.

Ding, M., Chen, Y., & Bressler, S. (2006). Granger causality: Basic theory and application to neuroscience. *Handbook of Time Series Analysis*.

Dojer, N. (2006). Learning bayesian networks does not have to be np-hard. In Kralovic, R., & Urzyczyn, P. (Eds.), *Mathematical Foundations of Computer Science 2006, 31st International Symposium, MFCS 2006, Stará Lesná, Slovakia, August 28-September 1, 2006, Proceedings*, Vol. 4162 of *Lecture Notes in Computer Science*, pp. 305–314. Springer.

Eells, E. (1991). *Probabilistic Causality*. Cambridge Studies in Probability, Induction and Decision Theory. Cambridge University Press.

Eichler, M. (2008). 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*.

Entner, D., & Hoyer, P. O. (2010). On causal discovery from time series data using fci. In *Proceedings of the Fifth European Workshop on Probabilistic Graphical Models (PGM-2010)*, pp. 121–128, Helsinki, Finland. HIIT Publications.

Faes, L., Nollo, G., & Chon, K. H. (2008). Assessment of granger causality by nonlinear model identification: application to short-term cardiovascular variability. *Annals of biomedical engineering*, *36*(3), 381–395.

Feng, G., Quirk, J. G., & Djurić, P. M. (2019). Detecting causality using deep gaussian processes. In *2019 53rd Asilomar Conference on Signals, Systems, and Computers*, pp. 472–476.

Friedman, N. (1997). Learning belief networks in the presence of missing values and hidden variables. In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML 1997), Nashville, Tennessee, USA, July 8-12, 1997*, pp. 125–133.

Friedman, N. (1998). The bayesian structural em algorithm. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, UAI'98, pp. 129–138, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Friedman, N., Murphy, K., & Russell, S. (1998). Learning the structure of dynamic probabilistic networks. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, UAI'98, pp. 139–147, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Gates, K. M., & Molenaar, P. C. (2012). Group search algorithm recovers effective connectivity maps for individuals in homogeneous and heterogeneous samples. *NeuroImage*, *63*(1), 310 – 319.

Geiger, P., Zhang, K., Schölkopf, B., Gong, M., & Janzing, D. (2015). Causal inference by identification of vector autoregressive processes with hidden components. In Bach, F., & Blei, D. (Eds.), *Proceedings of the 32nd International Conference on Machine Learning*, Vol. 37 of *Proceedings of Machine Learning Research*, pp. 1917–1925, Lille, France. PMLR.

Gerhardus, A., & Runge, J. (2020). High-recall causal discovery for autocorrelated time series with latent confounders..

Geweke, J. (1982). Measurement of linear dependence and feedback between multiple time series. *Journal of the American Statistical Association*, *77*(378), 304–313.

Glymour, C., Zhang, K., & Spirtes, P. (2019). Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, *10*, 524.

Gong, M., Zhang, K., Schölkopf, B., Glymour, C., & Tao, D. (2017). Causal discovery from temporally aggregated time series. In *Proceedings Conference on Uncertainty in Artificial Intelligence (UAI) 2017*, p. ID 269. Association for Uncertainty in Artificial Intelligence (AUAI).

Gong, M., Zhang, K., Schölkopf, B., Tao, D., & Geiger, P. (2015). Discovering temporal causal relations from subsampled data. In Bach, F., & Blei, D. (Eds.), *Proceedings of the 32nd International Conference on Machine Learning*, Vol. 37 of *Proceedings of Machine Learning Research*, pp. 1898–1906, Lille, France. PMLR.

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

Granger, C. (1980). Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and Control*, *2*, 329–352.

Granger, C. (1988). Some recent development in a concept of causality. *Journal of Econometrics*, *39*(1-2), 199–211.

Guo, R., Cheng, L., Li, J., Hahn, P. R., & Liu, H. (2020). A survey of learning causality with data: Problems and methods. *ACM Comput. Surv.*, *53*(4).

Hansson, H., & Jonsson, B. (1994). A logic for reasoning about time and reliability. *Formal Aspects of Computing*, *6*, 102–111.

Heckerman, D., Geiger, D., & Chickering, D. M. (1995). Learning bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, *20*(3), 197–243.

Hiemstra, C., & Jones, J. D. (1994). Testing for linear and nonlinear granger causality in the stock price-volume relation. *The Journal of Finance*, *49*(5), 1639–1664.

Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., & Schölkopf, B. (2009). Nonlinear causal discovery with additive noise models. In Koller, D., Schuurmans, D., Bengio, Y., & Bottou, L. (Eds.), *Advances in Neural Information Processing Systems 21*, pp. 689–696. Curran Associates, Inc.

Hu, M., & Liang, H. (2014). A copula approach to assessing granger causality. *NeuroImage*, *100*, 125 – 134.

Huang, B., Zhang, K., Gong, M., & Glymour, C. (2019). Causal discovery and forecasting in nonstationary environments with state-space models. In Chaudhuri, K., & Salakhutdinov, R. (Eds.), *Proceedings of Machine Learning Research*, Vol. 97, pp. 2901–2910, Long Beach, California, USA. PMLR.

Huang, B., Zhang, K., & Schölkopf, B. (2015). Identification of time-dependent causal model: A gaussian process treatment. In *24th International Joint Conference on Artificial Intelligence, Machine Learning Track*, pp. 3561–3568, Palo Alto, California USA. AAAI Press.

Huang, J. Z., & Yang, L. (2004). Identification of non-linear additive autoregressive models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *66*(2), 463–477.

Huang, Y., & Kleinberg, S. (2015). Fast and accurate causal inference from time series data. In *FLAIRS Conference*, pp. 49–54. AAAI Press.

Hume, D. (1738). *A Treatise of Human Nature*. Oxford University Press.

Hyttinen, A., Plis, S. M., Järvisalo, M., Eberhardt, F., & Danks, D. (2017). A constraint optimization approach to causal discovery from subsampled time series data. *International Journal of Approximate Reasoning*, *90*, 208–225.

Hyvärinen, A., Zhang, K., Shimizu, S., & Hoyer, P. O. (2010). Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, *11*, 1709–1731.

Jabbari, F., Ramsey, J., Spirtes, P., & Cooper, G. (2017). Discovery of causal models that contain latent variables through bayesian scoring of independence constraints. In *Proceedings of the European Conference on Machine learning and Knowledge Discovery in Databases, ECML-PKDD*, pp. 142–157.

Jiao, J., Permuter, H. H., Zhao, L., Kim, Y.-H., & Weissman, T. (2013). Universal estimation of directed information. *IEEE Transactions on Information Theory*, *59*(10), 6220–6242.

Kaiser, M., & Sipos, M. (2021). Unsuitability of notears for causal graph discovery..

Kalisch, M., & Bühlmann, P. (2007). Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, *8*, 613–636.

Kim, J.-M., Lee, N., & Hwang, S. Y. (2019). A copula nonlinear granger causality. *Economic Modelling*.

Kim, S., Putrino, D., Ghosh, S., & Brown, E. N. (2011). A granger causality measure for point process models of ensemble neural spiking activity. *PLOS Computational Biology*, *7*(3), 1–13.

Kleinberg, S. (2011). A logic for causal inference in time series with discrete and continuous variables. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two*, IJCAI'11, pp. 943–950. AAAI Press.

Kleinberg, S., & Mishra, B. (2009). The Temporal Logic of Causal Structures. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*, Montreal, Quebec.

Lanne, M., Meitz, M., & Saikkonen, P. (2017). Identification and estimation of non-gaussian structural vector autoregressions. *Journal of Econometrics*, *196*(2), 288 – 304.

Leng, S., Ma, H., & Kurths, J. e. a. (2020). Partial cross mapping eliminates indirect causal influences. *Nat Commun*, *11*(2632).

Luo, L., Liu, W., Koprinska, I., & Chen, F. (2015). Discovering causal structures from time series data via enhanced granger causality. In Pfahringer, B., & Renz, J. (Eds.), *AI 2015: Advances in Artificial Intelligence*, pp. 365–378, Cham. Springer International Publishing.

Malinsky, D., & Danks, D. (2018). Causal discovery algorithms: A practical guide. *Philosophy Compass*, *13*(1).

Malinsky, D., & Spirtes, P. (2018). Causal structure learning from multivariate time series in settings with unmeasured confounding. In *Proceedings of 2018 ACM SIGKDD Workshop on Causal Disocvery*, Vol. 92 of *Proceedings of Machine Learning Research*, pp. 23–47, London, UK. PMLR.

Marinazzo, D., Pellicoro, M., & Stramaglia, S. (2008). Kernel-granger causality and the analysis of dynamical networks. *Physical Review E*, *77*, 056215.

McCracken, J. M., & Weigel, R. S. (2014). Convergent cross-mapping and pairwise asymmetric inference. *Phys. Rev. E*, *90*, 062903.

Meek, C. (1997). *Graphical Models: Selecting causal and statistical models*. PhD thesis, Carnegie Mellon University.

Meek, C. (1995). Causal inference and causal explanation with background knowledge. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI-95, pp. 403–410, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Mogensen, S. W., Malinsky, D., & Hansen, N. R. (2018). Causal learning for partially observed stochastic dynamical systems. In *Proceedings of the 34th conference on Uncertainty in Artificial Intelligence*, pp. 350–360.

Moneta, A., Entner, D., Hoyer, P. O., & Coad, A. (2013). Causal inference by independent component analysis: Theory and applications. *Oxford Bulletin of Economics and Statistics*, *75*(5), 705–730.

Mooij, J., Janzing, D., Peters, J., & Schölkopf, B. (2009). 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, pp. 745–752, New York, NY, USA. ACM.

Nauta, M., Bucur, D., & Seifert, C. (2019). Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction*, *1*(1), 312–340.

Ng, A. Y. (1997). Preventing "overfitting" of cross-validation data. In Fisher, D. H. (Ed.), *Proceedings of the Fourteenth International Conference on Machine Learning (ICML 1997), Nashville, Tennessee, USA, July 8-12, 1997*, pp. 245–253. Morgan Kaufmann.

Nicolaou, N., & Constandinou, T. G. (2016). A nonlinear causality estimator based on non-parametric multiplicative regression. *Frontiers in neuroinformatics*, *10*, 19–19.

Nogueira, A., Gama, J., & Ferreira, C. (2021). Causal discovery in machine learning: Theories and applications. *Journal of Dynamics & Games*, *8*(3), 203–231.

Pamfil, R., Sriwattanaworachai, N., Desai, S., Pilgerstorfer, P., Georgatzis, K., Beaumont, P., & Aragam, B. (2020). Dynotears: Structure learning from time-series data. In Chiappa, S., & Calandra, R. (Eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, Vol. 108 of *Proceedings of Machine Learning Research*, pp. 1595–1605. PMLR.

Papagiannopoulou, C., Miralles, D. G., Decubber, S., Demuzere, M., Verhoest, N. E. C., Dorigo, W. A., & Waegeman, W. (2017). A non-linear granger-causality framework to investigate climate-vegetation dynamics. *Geoscientific Model Development*, *10*(5), 1945–1960.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, NY, USA.

Peña, J. M., Björkegren, J., & Tegnér, J. (2005). Learning dynamic bayesian network models via cross-validation. *Pattern Recognition Letters*, *26*, 2295–2308.

Peters, J., & Bühlmann, P. (2015). Structural intervention distance (sid) for evaluating causal graphs. *Neural Computation*, *27*(3), 771–799.

Peters, J., Janzing, D., & Schölkopf, B. (2013). Causal inference on time series using restricted structural equation models. In *Advances in Neural Information Processing Systems 26*, pp. 154–162.

Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, USA.

Rankin, M., & McCormack, T. (2013). The temporal priority principle: at what age does this develop?. *Frontiers in Psychology*, *4*, 178.

Reichenbach, H. (1956). *The Direction of Time*. Dover Publications.

Richardson, T. (1996). A discovery algorithm for directed cyclic graphs. In *Proceedings of the Twelfth International Conference on Uncertainty in Artificial Intelligence*, UAI'96, pp. 454–461, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Richardson, T., & Spirtes, P. (2002). Ancestral graph markov models. *Annals of Statistics*, *30*(4), 962–1030.

Robins, J. M., Scheines, R., Spirtes, P., & Wasserman, L. (2003). Uniform consistency in causal inference. *Biometrika*, *90*(3), 491–515.

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

Runge, J. (2020). Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In Peters, J., & Sontag, D. (Eds.), *Proceedings of Machine Learning Research*, Vol. 124, pp. 1388–1397. PMLR.

Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., & Sejdinovic, D. (2019). Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, *5*(11).

Sanchez-Romero, R., Ramsey, J. D., Zhang, K., Glymour, M. R. K., Huang, B., & Glymour, C. (2019). Estimating feedforward and feedback effective connections from fmri time series: Assessments of statistical methods. *Network Neuroscience*, *3*(2), 274–306.

Schreiber, T. (2000). Measuring information transfer. *Physical review letters*, *85*, 461–4.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464.

Shah, R. D., & Peters, J. (2020). The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, *48*(3), 1514 – 1538.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., & Kerminen, A. (2006). A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, *7*, 2003–2030.

Shimizu, S., Inazumi, T., Sogawa, Y., Hyvärinen, A., Kawahara, Y., Washio, T., Hoyer, P. O., & Bollen, K. (2011). Directlingam: A direct method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research*, *12*, 1225–1248.

Smith, S. M., Miller, K. L., Khorshidi, G. S., Webster, M. A., Beckmann, C. F., Nichols, T. E., Ramsey, J., & Woolrich, M. W. (2011). Network modelling methods for fmri. *NeuroImage*, *54*, 875–891.

Spirtes, P., Glymour, C., & Scheines, R. (1990). *Causation, Prediction, and Search* (1st edition). MIT press.

Spirtes, P., Glymour, C., & Scheines, R. (2001). *Causation, Prediction, and Search* (2nd edition). MIT press.

Sugihara, G., May, R., Ye, H., hao Hsieh, C., Deyle, E., Fogarty, M., & Munch, S. (2012). Detecting causality in complex ecosystems. *Science*, *338*, 496 – 500.

Sun, J., Taylor, D., & Bollt, E. (2015). Causal network inference by optimal causation entropy. *SIAM Journal on Applied Dynamical Systems*, *14*(1), 73–106.

Sun, X. (2008). Assessing nonlinear granger causality from multivariate time series. In Daelemans, W., Goethals, B., & Morik, K. (Eds.), *Machine Learning and Knowledge Discovery in Databases*, pp. 440–455, Berlin, Heidelberg. Springer Berlin Heidelberg.

Suppes, P. (1970). *A Probabilistic Theory of Causality*. Amsterdam: North-Holland Pub. Co.

Takens, F. (1981). Detecting strange attractors in turbulence. In Rand, D., & Young, L.-S. (Eds.), *Dynamical Systems and Turbulence, Warwick 1980*, pp. 366–381, Berlin, Heidelberg. Springer Berlin Heidelberg.

Verma, T., & Pearl, J. (1991). Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, UAI '90, pp. 255–270, New York, NY, USA. Elsevier Science Inc.

Vinh, N. X., Chetty, M., Coppel, R., & Wangikar, P. P. (2011). GlobalMIT: learning globally optimal dynamic bayesian network with the mutual information test criterion. *Bioinformatics*, *27*(19), 2765–2766.

Voortman, M., Dash, D., & Druzdzel, M. (2010). Learning why things change: The difference-based causality learner. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence, UAI 2010*, pp. 641 – 650.

Vowels, M. J., Camgöz, N. C., & Bowden, R. (2021). D'ya like dags? A survey on structure learning and causal discovery. *CoRR*, *abs/2103.02582*.

Wright, S. (1921). Correlation and causation. *Journal of agricultural research*, *20*(7), 557–585.

Ye, H., Deyle, E. R., Gilarranz, L. J., & Sugihara, G. (2015). Distinguishing time-delayed causal interactions using convergent cross mapping. *Scientific Reports*, *5*.

Zhalama, Zhang, J., & Mayer, W. (2016). Weakening faithfulness: some heuristic causal discovery algorithms. *International Journal of Data Science and Analytics*, *3*, 93–104.

Zhang, D. D., Lee, H. F., Wang, C., Li, B., Pei, Q., Zhang, J., & An, Y. (2011). The causality analysis of climate change and large-scale human crisis. *Proceedings of the National Academy of Sciences*, *108*(42), 17296–17301.

Zhang, J. (2007). 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, pp. 450–457, Arlington, Virginia, USA. AUAI Press.

Zhang, J. (2008). On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, *172*(16), 1873 – 1896.

Zhang, J., & Spirtes, P. (2002). Strong faithfulness and uniform consistency in causal inference. In *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*, UAI'03, pp. 632–639, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Zhang, K., Huang, B., Zhang, J., Glymour, C., & Schölkopf, B. (2017). Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pp. 1347–1353.

Zheng, X., Dan, C., Aragam, B., Ravikumar, P., & Xing, E. (2020). Learning sparse nonparametric dags. In Chiappa, S., & Calandra, R. (Eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, Vol. 108 of *Proceedings of Machine Learning Research*, pp. 3414–3425. PMLR.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429.