Chaos
An Interdisciplinary Journal of Nonlinear Science

RESEARCH ARTICLE | JULY 23 2018

Causal network reconstruction from time series: From theoretical assumptions to practical estimation

Special Collection: Causation Inference and Information Flow in Dynamical Systems: Theory and Applications

J. Runge

Chaos 28, 075310 (2018)
https://doi.org/10.1063/1.5025050

Causal network reconstruction from time series: From theoretical assumptions to practical estimation

J. Runge
German Aerospace Center, Institute of Data Science, Jena 07745, Germany

(Received 06 February 2018; accepted 25 May 2018; published online 23 July 2018)

Causal network reconstruction from time series is an emerging topic in many fields of science. Beyond inferring directionality between two time series, the goal of causal network reconstruction or causal discovery is to distinguish direct from indirect dependencies and common drivers among multiple time series. Here, the problem of inferring causal networks including time lags from multivariate time series is recapitulated from the underlying causal assumptions to practical estimation problems. Each aspect is illustrated with simple examples including unobserved variables, sampling issues, determinism, stationarity, nonlinearity, measurement error, and significance testing. The effects of dynamical noise, autocorrelation, and high dimensionality are highlighted in comparison studies of common causal reconstruction methods. Finally, method performance evaluation approaches and criteria are suggested. The article is intended to briefly review and accessibly illustrate the foundations and practical problems of time series-based causal discovery and stimulate further methodological developments.
https://doi.org/10.1063/1.5025050

Reconstructing interaction networks from observed time series is a common problem in fields where active experiments are impossible, unethical, or expensive. Pairwise association networks, for example based on correlations, cannot be interpreted causally. The goal of causal network reconstruction or causal discovery is to distinguish direct from indirect dependencies and common drivers among multiple time series. Here, we briefly recapitulate the theoretical assumptions underlying causal discovery from time series, discuss practical estimation problems, and illustrate each aspect with accessible examples.

I. INTRODUCTION

Reconstructing the causal relations behind the phenomena we observe is a fundamental problem in all fields of science. The traditional approach is to conduct active experiments, but in many fields such as Earth system science or neuroscience, manipulations of the complex system under study are either impossible, unethical, or very expensive. On the other hand, modern science generates an ever-growing amount of data from these systems, in particular time series data. Concurrently, novel computing hardware today allows efficient processing of massive amounts of data. These developments have led to emerging interest in the problem of reconstructing causal networks or causal discovery from observational time series.

In the past few decades, a number of original causality concepts have been developed, such as Granger causality (Granger, 1969) or transfer entropy (Schreiber, 2000). Since the 1990s, computer scientists, statisticians, and philosophers have grounded causal reasoning and inference in a robust mathematical framework (Pearl, 2000; Spirtes *et al.*, 2000). The (quite natural) definition of causality underlying this framework is that $X \to Y$ if and only if an intervention or manipulation in X has an effect on Y (Pearl, 2000; Spirtes *et al.*, 2000). This effect may be in changing Y's mean or any change in its *post-interventional* distribution denoted $P[Y \mid do(X = x)]$ which is different from the conditional distribution $P(Y \mid X = x)$. Unfortunately, all we can measure from observational data are statistical dependencies. These can be visualized in a *graphical model* (Lauritzen, 1996) or *time series graph* (Eichler, 2011) that represents the *conditional independence* relations among the variables and their time lags (Fig. 1). The theory of causal discovery lays out the assumptions under which the underlying causal dependencies can be inferred from observational data.

There are different sets of assumptions that allow us to *identify* a causal graph. Here, we focus on time-lagged causal discovery in the framework of conditional independence testing using the assumptions of time-order, *Causal Sufficiency*, the *Causal Markov Condition*, and *Faithfulness*, among others, which are all discussed thoroughly in this paper. But some of these assumptions can be replaced. Recent work (Peters *et al.*, 2017) shows ways to use assumptions on the noise structure and dependency types in the framework of structural causal models which can complement the approach studied here and we will include references to recent work from this framework throughout the sections.

The paper is organized as follows: In Sec. II, we relate Granger causality and similar concepts to the conditional independence-framework (Spirtes *et al.*, 2000). Section III provides the necessary definitions and notation and in Sec. IV we recapitulate the assumptions underlying time-lagged causal discovery from time series alongside illustrative examples. The practical estimation aspect from introducing

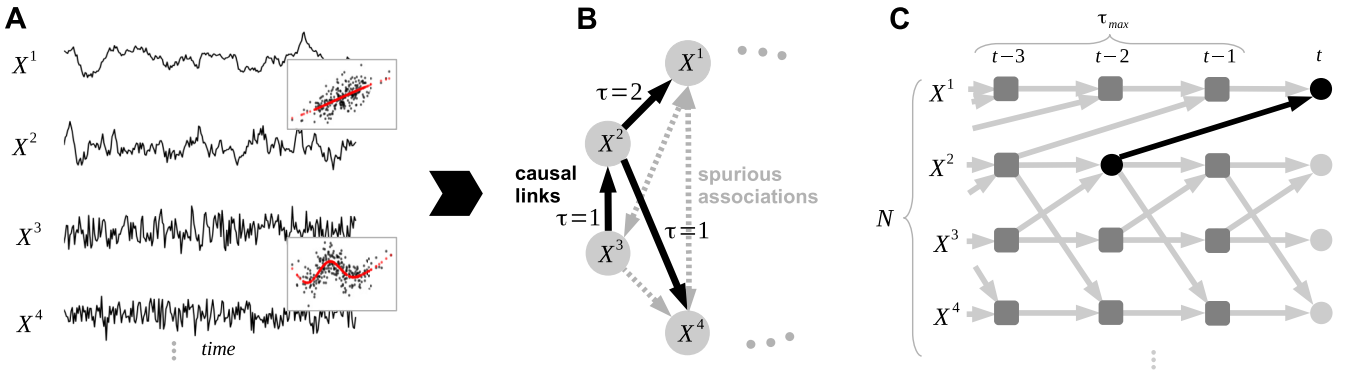

FIG. 1. Causal network reconstruction. Consider a time series dataset (panel **A**) from a complex system of which we try to reconstruct the underlying *causal* dependencies (panel **B**), accounting for linear and nonlinear dependencies and including their time lags (link labels). Causal discovery aims to unveil spurious associations (gray arrows) which necessarily emerge due to common drivers (e.g., $X^1 \leftarrow X^2 \rightarrow X^4$) or transitive indirect paths (e.g., $X^3 \rightarrow X^2 \rightarrow X^1$). Correlation matrices are, therefore, often very dense, while causal networks are typically sparse. (**C**) The time series graph defined in Definition 1 resolves also the time-dependence structure up to some maximum time lag $\tau_{\max}$. A link $X^i_{t-\tau} \rightarrow X^j_t$ (black edge) exists if $X^i_{t-\tau}$ and Y^j_t are *not* independent conditionally on the past of the whole process (gray boxes).

some causal discovery algorithms to significance testing is presented in Sec. V, while Sec. VI discusses suggestions for performance evaluation. In Sec. VII, we present some comparison studies of common causal methods and conclude the paper with a brief discussion (Sec. VIII). The paper is accompanied by a python jupyter notebook on https://github.com/jakobrunge/tigramite to reproduce some of the examples.

II. FROM GRANGER CAUSALITY TO CONDITIONAL INDEPENDENCE

Granger (1969), based on work by Wiener (1956), was the first to propose a practical, operational definition of causality based on prediction improvement. The underlying idea of measuring whether *X Granger causes Y* is that there is some unique information in X relevant for Y that is not contained in Y's past as well as the past of "all the information in the universe" (Granger, 1969). In practice, typically only Y's past is used (bivariate Granger causality). Measuring prediction improvement can be operationalized in different ways. The most common framework are *vector autoregressive models* (*VAR*),

$$\mathbf{X}_t = \sum_{\tau=1}^{\tau_{\max}} \Phi(\tau)\mathbf{X}_{t-\tau} + \eta_t\,, \tag{1}$$

where $\mathbf{X}_t = (X_t^1, \ldots, X_t^N)$, $\Phi(\tau)$ is the $N \times N$ coefficient matrix at lag τ, $\tau_{\max}$ some maximum time lag, and η denotes an independent noise term. Here, X^i *Granger-causes* X^j if *any* of the coefficients $\Phi_{ji}(\tau)$ at lag τ is non-zero. A non-zero $\Phi_{ji}(\tau)$ can then be denoted as a causal link $X^i_{t-\tau} \rightarrow X^j_t$ at lag τ. Another option is to compare the residual variances of the VAR fitted with and without including the variable X^i. The use of VARs restricts this notion of causality to a *causality in mean* (Granger, 1969). A more general definition is that of (*bivariate*) *transfer entropy* (Schreiber, 2000; Barnett *et al.*, 2009)

$$I^{\text{TEbiv}}_{X\rightarrow Y} = I(X_t^-; Y_t \mid Y_t^-), \tag{2}$$

where $I(X;Y \mid Z)$ denotes the *conditional mutual information* (CMI). *Bivariate TE* is a common term, another naming option would be *bivariable TE* since X and Y could also be multivariate variables. Transfer entropy can also be phrased in a multivariate (or multi-variable) lag-specific version (Runge *et al.*, 2012a). Many current methods are advancements of the concept of transfer entropy (Wibral *et al.*, 2013; Staniek and Lehnertz, 2008; Vejmelka and Palus, 2008), in particular in its multivariate version (Sun and Bollt, 2014; Sun *et al.*, 2015; Runge *et al.*, 2012b; 2012a; Runge, 2015).

Tests for causality are then based on testing whether a particular CMI is greater than zero. Looking at the definition of CMI,

$$I(X;Y \mid Z) = \iiint p(x,y,z) \log \frac{p(x,y|z)}{p(x|z)\cdot p(y|z)}\, dx\, dy\, dz. \tag{3}$$

TEbiv and its advancements essentially test for *conditional independence* of X and Y given Z, denoted $X \perp\!\!\!\perp Y|Z$ since

$$X \perp\!\!\!\perp Y|Z \Longleftrightarrow p(x,y|z) = p(x|z)p(y|z) \quad \forall\, x,y,z \tag{4}$$

$$\Longleftrightarrow I(X;Y \mid Z) = 0. \tag{5}$$

Z then represents Y's past and other included variables. The lag-specific generalization of the VAR model (1) then is the *full conditional independence* (FullCI) approach

$$I^{\text{FullCI}}_{i\rightarrow j}(\tau) = I(X^i_{t-\tau}; X^j_t \mid \mathbf{X}_t^{(t-1,\ldots,t-\tau_{\max})} \backslash \{X^i_{t-\tau}\}), \tag{6}$$

where $\mathbf{X}_t^{(t-1,\ldots,t-\tau_{\max})} = (\mathbf{X}_{t-1}, \ldots, \mathbf{X}_{t-\tau_{\max}})$. I can be CMI or any other conditional dependence measure. In the case of partial correlation, a non-zero entry $\Phi_{ji}(\tau)$ corresponds to a non-zero $I^{\text{FullCI}}_{i\rightarrow j}(\tau)$. A general concept to represent *conditional independence* relations among multiple variables and their time lags is that of *time series graphical models* (Eichler, 2011).

III. DEFINITIONS AND NOTATION

A. Definition of time series graphs

Consider a multivariate process $\mathbf{X}$ of dimension N. We define the *time series graph* $\mathcal{G} = (V \times \mathbb{Z}, E)$ of $\mathbf{X}$ as follows

[Fig. 1(c)]: The set of nodes in that graph consists of the set of components V at each time $t \in \mathbb{Z}$. That is, the graph is actually infinite, but in practice defined up to some maximum time lag $\tau_{\max}$. Compared to the general concept of graphical models (Lauritzen, 1996) for data without time-ordering, for time series graphs the time-dependence is explicitly used to define directional links in the set of edges E (Eichler, 2011). For convenience, we treat $\mathbf{X}$, $\mathbf{X}_t$, and $\mathbf{X}_t^-$ as sets of random variables here and use the difference symbol "\" for sets.

Definition 1. (Definition of links).

Variables $X^i_{t-\tau}$ *and* X^j_t *are connected by a lag-specific* directed link "$X^i_{t-\tau} \to X^j_t$" $\in \mathcal{G}$ [Fig. 1(c)] *for* $\tau > 0$ *if and only if*

$$X^i_{t-\tau} \not\perp\!\!\!\perp X^j_t \mid \mathbf{X}^-_t \setminus \{X^i_{t-\tau}\}, \tag{7}$$

i.e., if they are not *independent conditionally on the past of the whole process denoted by* $\mathbf{X}^-_t = (\mathbf{X}_{t-1}, \mathbf{X}_{t-2}, \ldots)$, *which implies a lag-specific conditional dependence with respect to* $\mathbf{X}$.

For $i = j$, we call links *autodependencies*. For $\tau = 0$, the same definition can also be used to define undirected contemporaneous links (see Eichler, 2011; Runge, 2015), which would lead to the time series graph being a *mixed graph* instead of a *directed acyclic graph*. The arrow of time is a convenient way to disambiguate independence relationships: If we do not have access to the time ordering of the variables (or there is none) and observe as the only conditional independence relation $X \perp\!\!\!\perp Y \mid Z$ while all other relations are dependent, then this relation can be generated by any of the three causal motifs $X \to Z \to Y$, $Y \to Z \to X$, or $X \leftarrow Z \to Y$ which form a *Markov equivalence class*.

Here, we define links with lags $t - \tau$ relative to some time point t, but throughout this paper we assume stationarity (discussed in Sec. IV E). Then a link is repeated for every $t' < t$ if a link exists at time t. Alternatively, the links can be estimated from different realizations at time t. In practice, however, the links will mostly be estimated from single time series realizations requiring stationarity.

The parents of a node X^j_t are defined as

$$\mathcal{P}(X^j_t) = \{X^k_{t-\tau} : X^k \in \mathbf{X}, \tau > 0, X^k_{t-\tau} \to X^j_t\}. \tag{8}$$

In the following, A, B, S denote nodes or sets of nodes in the graph and $X_{t-\tau}, Y_t, Z_{t-\tau}, U_{t-\tau}, \mathbf{Z}$ random variables of the process $\mathbf{X}$, sometimes dropping the subscript. We will denote a general conditional dependence measure as $I(X; Y|Z)$ which can be CMI or also some other measure such as partial correlation, depending on the context.

B. Separation

When considering the dependency between two variables X, Y given a set of conditions $\mathbf{Z}$ as in $I(X; Y|Z)$, the idea of open and blocked paths or separation between the corresponding nodes in the time series graph $\mathcal{G}$ [Fig. 1(c)] is important. A *directed path* is a sequence of linked nodes containing only motifs $\to \bullet \to$. But there are also other paths on which information is shared even though no causal interventions could "travel" along these. In general (Eichler, 2011), in the above defined time series graph, a path between two single nodes A and B is called *open* if it contains only the motifs $\to \bullet \to$ or $\leftarrow \bullet \to$. For notational convenience, we will sometimes use left-pointing arrows, while still in the time series graph all directed links are forward in time. For example, $\bullet \leftarrow \bullet \to \bullet \to \bullet$ is an open path. On the other hand, if *any motif* on a path is $\to \bullet \leftarrow$, the path is blocked. Nodes in such motifs are also called *colliders*. If we now consider a *separating or conditioning set* S, openness and blockedness of motifs reverse, i.e., denoting a conditioned node by $\blacksquare$, the motifs $\to \blacksquare \to$ and $\leftarrow \blacksquare \to$, are blocked and the motif $\to \blacksquare \leftarrow$ becomes open. For example, the path $\bullet \leftarrow \blacksquare \to \bullet \to \bullet$ is blocked, while $\bullet \leftarrow \bullet \to \blacksquare \leftarrow \bullet$ is open. Note that paths can also traverse links repeatedly, e.g., forward and backward.

Definition 2. (Separation).

Two nodes A *and* B *are* separated given a conditioning set S *with* $A, B \notin S$ (S *may also be empty*) *if* all paths *between* A *and* B *are blocked, denoted*

$$A \bowtie B \mid S. \tag{9}$$

Conversely, two nodes are connected given a set S *if* at least one path *between the two is open*.

Intuitively, if two nodes are separated, no information is shared between the two. For example, in Fig. 1(c) X^1_t and X^4_{t-1} are separated by $S = \{X^2_{t-2}, X^4_{t-2}\}$ or also by $S = \{X^1_{t-1}, X^2_{t-2}\}$. Then $I(X^4_{t-1}; X^1_t|\mathcal{S}) = 0$. Conversely, X^1_t and X^3_{t-3} are still connected given $S = \{X^2_{t-2}\}$ since the "back door"-path $X^3_{t-3} \leftarrow X^3_{t-4} \to X^2_{t-3} \to X^1_{t-1} \to X^1_t$ is still open. These definitions are important for the relations between the graph and the underlying process.

IV. ASSUMPTIONS OF CAUSAL DISCOVERY FROM OBSERVATIONAL TIME SERIES

Causal information cannot be obtained from associations of measured variables without some assumptions. A variety of different assumptions have been shown to be sufficient to estimate the true causal graph (Spirtes *et al.*, 2000; Peters *et al.*, 2017). Here, we focus on three main assumptions under which the time series graph represents causal relations: *Causal Sufficiency*, the *Causal Markov Condition*, and *Faithfulness*. For time-lagged causal discovery from observational time series, we also need the assumptions of *no instantaneous effects* and *stationarity*. Further added to these are *dependence type assumptions* (e.g., linear or nonlinear) and *no measurement error*, and we will also assume that the joint distribution of the process has a positive density. All of these are discussed in the following.

For illustrating some of the assumptions in this paper, we will estimate the time series graphs by directly testing Definition 1 via Eq. (6) [Fig. 1(c)], mostly in the partial correlation version.

A. Causal sufficiency

As Granger (1969) already notes, "[t]he one completely unreal aspect of the above definitions is the use of the series $\mathbf{U}^-_t$ representing all available information [in the universe]." This definition makes sure that the measured variables include all of the common causes of X and Y. However, we do not always need the whole universe. If we have available only

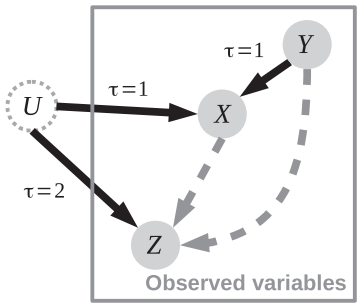

FIG. 2. The problem of latent (unobserved) variables. Here, U is a latent confounder and leads to spurious links between (X, Y, Z).

a limited number of measured variables, we only need to assume that there exist no other unobserved (or latent) variables that directly or indirectly influence any other pair of our set of variables which is the assumption of *Causal Sufficiency* (Spirtes *et al.*, 2000).

Definition 3. (Causal Sufficiency).

A set $W \subset V \times \mathbb{Z}$ of variables is causally sufficient *for a process* $\mathbf{X}$ *if and only if in the process every common cause of any two or more variables in W is in W or has the same value for all units in the population.*

Example 1. (Unobserved variables)

What happens if such an unobserved (latent) confounder exists? Consider the example depicted in Fig. 2 where we assume no autodependencies. Here, U is an unobserved variable and drives both X and Z. With the time lags considered, this common driver leads to an association between X and Z. The *estimated* time series graph [via Eq. (6)] among the observed variables (X, Y, Z) will then contain a link $X_{t-1} \to Z_t$ even though it is a spurious association and any manipulation of X would not have any effect on Z. Here, U acts as a direct common driver leading to an induced association. But the estimated time series graph additionally contains a link $Y_{t-2} \to Z_t$ even though there is no direct confounder between Y and Z. The reason is that $Y_{t-2} \not\perp\!\!\!\perp Z_t \mid \mathbf{X}_t^-$ because the path $Y_{t-2} \to X_{t-1} \leftarrow U_{t-2} \to Z_t$ is open through the condition on $X_{t-1} \in \mathbf{X}_t^-$ (see Definition 2).

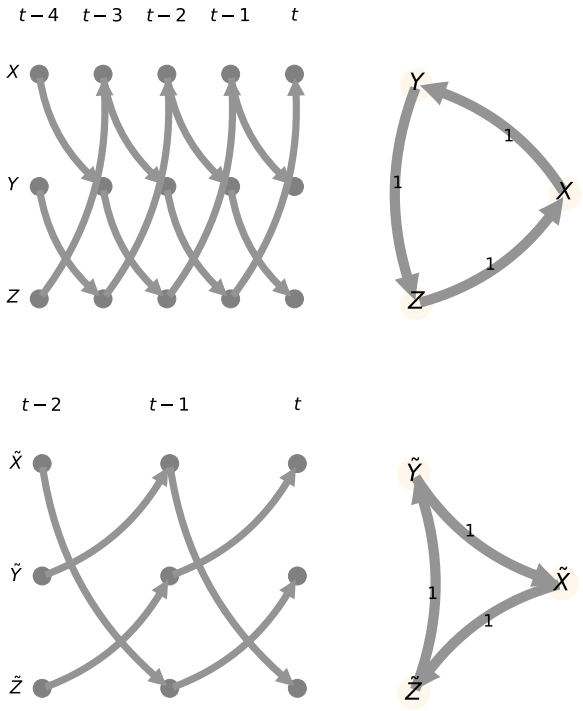

FIG. 3. Sub-sampled time series. If the time series of the true underlying process shown in the top panels (time series graph on the left, aggregated process graph on the right) is sampled at $\Delta t = 2$, the time series graph of the sub-sampled process (bottom left panel, note that t here refers to the sub-sampled time) here even has a reversed causal loop.

The *Fast Causal Discovery* (FCI) algorithm (Spirtes *et al.*, 2000; Zhang, 2008) does not assume causal sufficiency and allows us to partially identify which links are spurious due to unobserved confounders and also for which links confoundedness cannot be determined. The underlying idea is that if conditional independence holds between two variables for *any subset* (including the empty set) of W, then these variables are not linked. In the example above, this idea can be used to remove the link $Y_{t-2} \to Z_t$ since $Y_{t-2} \perp\!\!\!\perp Z_t$, i.e., Y_{t-2} and Z_t are unconditionally independent. Latent causal discovery is further addressed, for example, in Entner and Hoyer (2010), Eichler (2013), Ramb *et al.* (2013), Smirnov (2013), Hyttinen *et al.* (2014), and Geiger *et al.* (2015).

Example 2. (Sub-sampled time series)

Causal sufficiency can also be violated if all variables are observed, but they are sampled at too coarse time intervals relative to the causal links. Consider a process with time series graph depicted in the top panel of Fig. 3 featuring a causal loop $X \to Y \to Z \to X$ with all causal links at lag $\tau = 1$. If we sub-sample the time series with an original resolution of $\Delta t = 1$ at $\Delta t = 2$, we would estimate the causal graph from $(\tilde{X}, \tilde{Y}, \tilde{Z})$ as shown in the bottom panel of Fig. 3 that has a completely reversed causal loop. Looking at the top panel time series graph again, this spurious reversal can be understood: For example, in the path $Z_{t-2} \to X_{t-1} \to Y_t$ the node X_{t-1} is not sampled and, thus, unobserved, leading to a spurious link $\tilde{Z}_{t-1} \to \tilde{Y}_t$ in the sub-sampled time series graph in the bottom panel (note that t is measured with twice the sampling rate then). Sub-sampled time series are an active area of research (Smirnov, 2013; Barnett and Seth, 2015; Spirtes and Zhang, 2016), to some extent sub-sampled time series graphs can be identified as addressed in Gong *et al.* (2015) and Hyttinen *et al.* (2016).

B. Causal Markov condition

All independence-based causal discovery methods necessitate the Causal Markov Condition (Spirtes *et al.*, 2000) which constitutes a close relationship between the process $\mathbf{X}$ and its graph $\mathcal{G}$.

Definition 4. (Causal Markov Condition).

The joint distribution of a time series process $\mathbf{X}$ *with graph $\mathcal{G}$ fulfills the Causal Markov Condition if and only if for all $Y_t \in \mathbf{X}_t$ with parents $\mathcal{P}_{Y_t}$ in the graph*

$$\mathbf{X}_t^- \setminus \mathcal{P}_{Y_t} \bowtie Y_t \mid \mathcal{P}_{Y_t} \Rightarrow \mathbf{X}_t^- \setminus \mathcal{P}_{Y_t} \perp\!\!\!\perp Y_t \mid \mathcal{P}_{Y_t}, \qquad (10)$$

that is, from separation in the graph (since the parents $\mathcal{P}_{Y_t}$ separate Y_t from $\mathbf{X}_t^- \setminus \mathcal{P}_{Y_t}$ in the graph) follows independence.

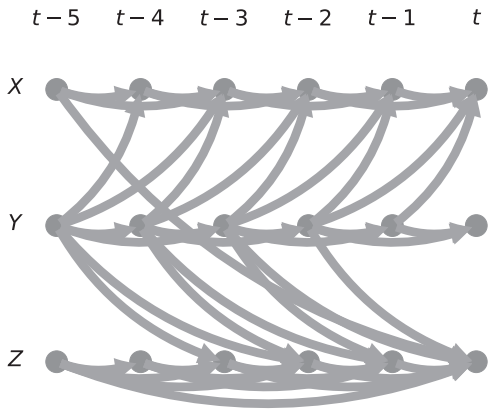

FIG. 4. Example of time series graph for non-markovian process given by Model (12). Since the present is not independent of the past given its parents, there are many spurious links if the graph is estimated with FullCI [Eq. (6)].

This includes its contraposition

$$\mathbf{X}_t^- \setminus \mathcal{P}_{Y_t} \not\perp\!\!\!\perp Y_t \mid \mathcal{P}_{Y_t} \Rightarrow \mathbf{X}_t^- \setminus \mathcal{P}_{Y_t} \not\bowtie Y_t \mid \mathcal{P}_{Y_t}, \qquad (11)$$

from dependence follows connectedness.

Note that if Causal Sufficiency is not fulfilled, then also generally the Markov condition will not hold (Spirtes *et al.*, 2000). Intuitively, the Causal Markov Condition implies that once we know the values of Y_t's parents, all other variables in the past ($t-\tau$ for $\tau > 0$) become irrelevant for predicting Y_t. Of course, Y's descendants at future time points can also "predict" Y_t.

Example 3. (Non-markovian processes)

A typical example of a non-markovian process is an autoregressive process driven by $1/f$ noise where the power spectrum is inversely proportional to the frequency of the signal. Consider a process generated by

$$\begin{aligned} X_t &= 0.4Y_{t-1} + \eta_t^X \\ Y_t &= \eta_t^Y \\ Z_t &= 0.3Y_{t-2} + \eta_t^Z, \end{aligned} \qquad (12)$$

where η_t^i for $i = X, Y, Z$ is $1/f$ noise. Such noise terms are not independent in time anymore, even though the noise terms between each individual variable are still independent, i.e., $\eta_t^X \perp\!\!\!\perp \eta_t^Y \perp\!\!\!\perp \eta_t^Z$. Here, $\mathcal{P}_{Z_t} = \{Y_{t-2}\}$, but still we observe many more links in the time series graph (estimated with FullCI in Fig. 4), both within one variable and between variables. This means that $Z_t \not\perp\!\!\!\perp \mathbf{X}_t^- \setminus \mathcal{P}_{Z_t} \mid \mathcal{P}_{Z_t}$—separation in the graph does not imply independence in the process, and the Markov condition is violated.

Example 4. (Time aggregation)

Another example where noise terms become dependent is time-aggregation. Consider the causal chain of processes $X_{t-2} \to Y_{t-1} \to Z_t$ shown in the top panel of Fig. 5. Time aggregation of a time series realization with $\Delta t = 2$ is here done by constructing the new time series $\tilde{X}_t = \frac{1}{2}(X_t + X_{t-1}) \forall t$ and correspondingly for Y, Z. Now, we observe additional contemporaneous links next to the directed links $\tilde{X}_{t-1} \to \tilde{Y}_t$ and $\tilde{Y}_{t-1} \to \tilde{Z}_t$ in the time series graph of the aggregated process due to the too coarse time resolution. But, furthermore, we also get spurious directed links, for example between $\tilde{X}$ and $\tilde{Z}$. In general, the causal structure of an aggregated process may be very different from the original process.

Time aggregation is an important issue in many applied fields. For example, in climate research time series are frequently measured daily and then aggregated to a monthly scale to investigate dependencies (Runge *et al.*, 2014). Ideally, the time resolution is at least as short as the shortest causal time lag (assuming no instantaneous effects, see Sec. IV D). See Breitung and Swanson (2002) and Barnett and Seth (2015) for a discussion on temporal aggregation in time series models. Ignoring time order, in some cases the recent methods discussed in Peters *et al.* (2017) can help.

C. Faithfulness

The Causal Markov Condition guarantees that separation in the graph implies independence in the process. But what can be concluded from an estimated conditional independence relation, that is, the reverse direction? *Faithfulness* guarantees that the graph entails *all* conditional independence relations that are implied by the Markov condition.

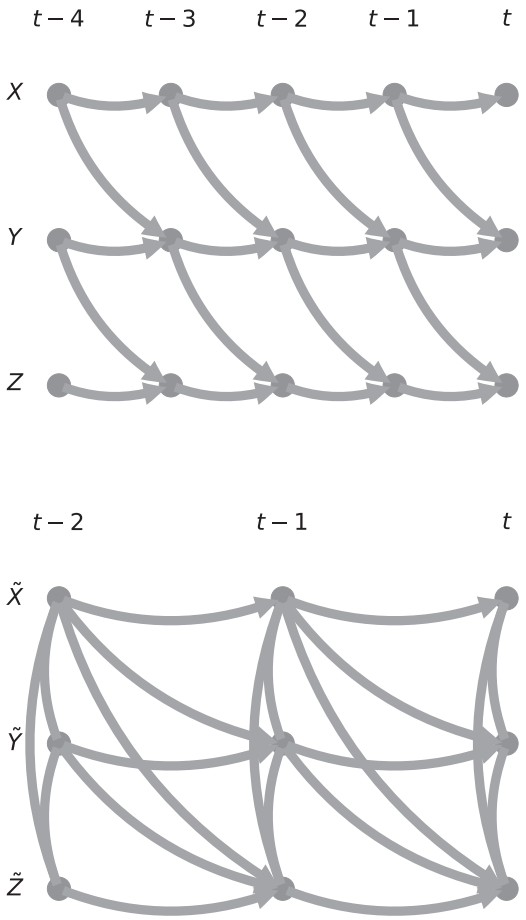

FIG. 5. Time aggregation. If the time series of the true underlying process shown in the top panel is aggregated at $\Delta t = 2$, the time series graph of the aggregated process (bottom panel, note that t here refers to the aggregated time) has contemporaneous dependencies due to the too coarse time resolution compared to the lags of the causal links, but also many spurious directed links.

Definition 5. (Faithfulness).

The joint distribution of a time series process $\mathbf{X}$ *with graph* $\mathcal{G}$ *fulfills the Faithfulness condition if and only if for all disjoint subsets of nodes (or single nodes)* $A, B, S \subset \mathcal{G}$ *it holds that*

$$X_A \perp\!\!\!\perp X_B \mid X_S \Rightarrow A \bowtie B \mid S, \tag{13}$$

that is, from independence follows separation, which includes its logical contraposition

$$A \not\bowtie B \mid S \Rightarrow X_A \not\perp\!\!\!\perp X_B \mid X_S, \tag{14}$$

from connectedness follows dependence.

The combination of Faithfulness and the Markov property implies that $A \bowtie B \mid S \Leftrightarrow X_A \perp\!\!\!\perp X_B \mid X_S$ and its logical contraposition $A \not\bowtie B \mid S \Leftrightarrow X_A \not\perp\!\!\!\perp X_B \mid X_S$. Both conditions are an important assumption for causal discovery algorithms as discussed in Spirtes *et al.* (2000). Intuitively, Faithfulness together with the Causal Markov Condition allow us to conclude that (in the limit of infinite sample size) a measured statistical dependency is actually due to some (not necessarily direct) causal mechanism and, conversely, a measured independence (given any set of conditions) implies that no *direct* causal mechanism exists (see also Remark 1 in the Discussion).

Example 5. (Counteracting mechanisms)

In a linear model [e.g., Eq. (1)], the coefficient values form a real space and the set of points in this space that create vanishing partial correlations not implied by the Causal Markov Condition have Lebesgue measure zero (Spirtes *et al.*, 2000). One can, thus, argue that non-faithful distributions arise from an unrealistic fine-tuning of dependence parameters. However, *approximately* vanishing partial correlations despite connectedness in the graph can also occur for a distribution that is faithful, but *almost* unfaithful, if we have only a limited sample size available as discussed in Uhler *et al.* (2013). An example of an unfaithfully fine-tuned process is

$$\begin{aligned} X_t &= \eta_t^X \\ Y_t &= 0.6X_{t-1} + \eta_t^Y \\ Z_t &= 0.6Y_{t-1} - 0.36X_{t-2} + \eta_t^Z. \end{aligned} \tag{15}$$

As shown in Fig. 6, here X influences Z directly as well as indirectly through Y. Now simple algebra shows that

$$Z_t = 0.6Y_{t-1} - 0.36X_{t-2} + \eta_t^Z \tag{16}$$

$$= 0.6(0.6X_{t-2} + \eta_{t-1}^Y) - 0.36X_{t-2} + \eta_t^Z \tag{17}$$

$$= \underbrace{0.36X_{t-2} - 0.36X_{t-2}}_{=0} + 0.6\eta_{t-1}^Y + \eta_t^Z \tag{18}$$

implying that Z and X are unconditionally independent since both mechanisms counteract each other even though there is a link in the graph. In Runge (2015) such counteracting interdependencies are analyzed information-theoretically.

Example 6. (Determinism)

One may argue that we live in a deterministic world and the assumption of "an independent noise term" that is pertinent to statistics is unrealistic. On the other hand, for a given observed variable Y, the complexity of the underlying processes will almost always imply that Y does not deterministically depend on its parents, but some unresolved processes constitute "intrinsic" or "dynamical" noise that is also driving Y.

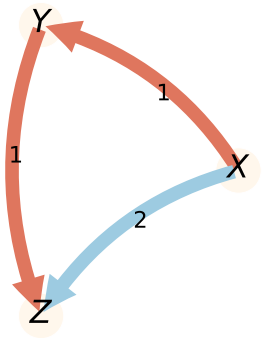

FIG. 6. Example of unfaithfully fine-tuned process where X and Z are independent even though they are connected in the graph. Here, the indirect mechanism $X \to Y \to Z$ with a positive effect counteracts the direct link $X \to Z$ with a negative effect.

Determinism violates Faithfulness as follows. Consider the model

$$\begin{aligned} Z &= \eta^Z \\ X &= f(Z) \\ Y &= g(Z) + cX + \eta^Y, \end{aligned} \tag{19}$$

with $c > 0$ and some functions f, g. Here, we have $I(X; Y \mid Z) = I[f(Z); Y \mid Z] = 0$ {since $H[f(Z) \mid Z] = 0$ (Cover and Thomas, 2006)} implying $X \perp\!\!\!\perp Y \mid Z$ even though Y depends on X in the model. One can argue that X should not be considered as an autonomous causal variable in this example and instead consider $Z \to Y$ as the causal graph for this model writing the model above as

$$\begin{aligned} Z &= \eta^Z \\ Y &= g(Z) + cf(Z) + \eta^Y. \end{aligned} \tag{20}$$

Determinism in causal inference can to some extent be addressed in the conditional independence framework (Spirtes *et al.*, 2000) or using structural causal models in Janzing *et al.* (2012) and Daniusis *et al.* (2012).

The former example only illustrated a static case of determinism. The field of nonlinear dynamics studies the properties of nonlinear and chaotic dynamical processes which has led to a plethora of nonlinear time series analysis methods (Kantz and Schreiber, 2003), often from an information-theoretic perspective (Hlaváčková-Schindler *et al.*, 2007) including transfer entropy. Many of these methods built on the assumption that no system is perfectly deterministic, for example, due to the coarse-graining of the system's phase-space in the measurement process. In Sec. VII A, we study the effect of dynamical noise on several common time-series based causal discovery approaches for chaotic systems.

Example 7. (Non-pairwise dependencies)

Next to the fine-tuned example on counteracting mechanisms, Faithfulness can also be violated for a dependency

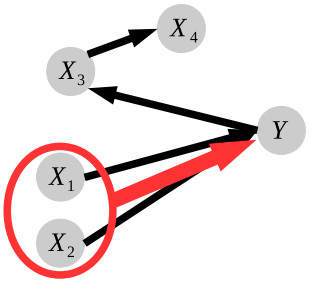

FIG. 7. Schematic view of dependency network with pairwise dependencies and a synergistic dependency of Y on (X_1, X_2) with $X_1 \perp\!\!\!\perp Y$ and $X_2 \perp\!\!\!\perp Y$, but $(X_1, X_2) \not\perp\!\!\!\perp Y$.

realized in an XOR-gate. Suppose, as shown in Fig. 7 among several "faithful" pairwise dependencies, we have that $Y = X_1 \oplus X_2 + \eta^Y$ with X_1, X_2 being binary random processes with equal probability for all values of X_1 and X_2. Then $I(X_1; Y) = I(X_2; Y) = 0$ even though both are connected to Y—a violation of Faithfulness. Here, the full information is *synergistically* contained in $I[(X_1, X_2); Y]$. Note that from the chain rule it follows that $0 < I[(X_1, X_2); Y] = I(X_1; Y) + I(X_2; Y \mid X_1) = I(X_2; Y \mid X_1)$. Thus, the MI is zero, but the CMI is not and there is, hence, a link in the time series graph. Note that already if $P(X_1) \neq P(X_2)$, Faithfulness is not violated anymore as analyzed in Sun *et al.* (2015), showing that Faithfulness violations are rather pathological.

Another form of synergy without violation of Faithfulness is the case that $Y = X_1 X_2 + \eta^Y$, where we have a rather weak MI $I(X_1; Y)$, but again a much larger $I[(X_1, X_2); Y]$. In Runge *et al.* (2015a), synergy is analyzed in the context of optimal prediction schemes.

As pointed out in James *et al.* (2016) for synergistic dependencies, the problem is that the concept of a pairwise dependency graphical model does not apply, but hyper-graphs are needed to represent such dependencies. Causal discovery of such graphs, however, carries the problem of combinatorial explosion if links between sets of nodes are considered.

D. Instantaneous effects

Granger causality and the definition of time series graphs are examples for lagged definitions of causality. To guarantee that the lagged parents defined in Eq. (8) are sufficient for the Causal Markov Condition to hold, we need to assume that there are *no instantaneous (contemporaneous) causal effects*, i.e., $X_t^i \to X_t^j$. One may argue that causality between dynamical systems cannot have instantaneous effects because the speed of light is finite and, if the process is sampled with sufficient resolution (otherwise, see the Examples of sub-sampling and aggregation), we only need to consider lagged causal effects. However, we often do not have a sufficiently sampled time series. Here, recent developments in causal inference theory (Zhang and Hyvärinen, 2009; Peters *et al.*, 2013; Lopez-Paz *et al.*, 2015; Spirtes and Zhang, 2016; Peters *et al.*, 2017) address instantaneous causality within the framework of structural causal models which can be applied to determine causal directionality for contemporaneous links. These models work under assumptions on the noise in the model such as non-Gaussianity. Also, the logical causal orientation rules of the causal discovery approaches in Spirtes *et al.* (2000) can be used to partially orient contemporaneous links.

E. Stationarity

To estimate the time series graph defined in Definition 1 from time series data, we assume stationarity. Another option would be to utilize independent ensembles of realizations of lagged processes. Here, we define stationarity with respect to a time index set $\mathcal{T}$. For example, $\mathcal{T}$ can contain all time indices belonging to a certain regime of a dynamical process, e.g., only winter months in the climate system.

Definition 6. (Causal stationarity).

The time series process $\mathbf{X}$ *with graph defined in Definition 1 is called* causally stationary over a time index set $\mathcal{T}$ *if and only if for all links* $X_{t-\tau}^i \to X_t^j$ *in the graph*

$$X_{t-\tau}^i \not\perp\!\!\!\perp X_t^j \mid \mathbf{X}_t^- \setminus \{X_{t-\tau}^i\} \text{ holds for all } t \in \mathcal{T}. \tag{21}$$

This constitutes actually a weaker form of stationarity than the common definition of stationarity in mean, variance, spectral properties, or of the value of individual coefficients in a linear model. For example, one could require that all CMIs are stationary,

$$I(X_{t-\tau}^i; X_t^j \mid \mathbf{X}_t^- \setminus \{X_{t-\tau}^i\}) \text{ has the same value for all } t \in \mathcal{T}, \tag{22}$$

which is a much stronger statement. The strength of causal mechanisms may fluctuate over time and the causal stationarity assumption only requires conditional independence to be stationary.

Example 8. (Non-stationarity due to confounding)

Consider the data shown in Fig. 8 and suppose we first only have access to the variables (X, Y, Z). Clearly, the time series of this subprocess are nonstationary in a classical sense, varying over time not only in their mean but also in their spectral properties. Estimating the time series graph on these three variables results in the graph shown in the bottom left panel of Fig. 8, where the common nonstationary trend leads to an almost fully connected graph.

A typical example of a common nonstationarity, albeit not the same as in our example, is found in climate time series which are usually all driven by solar forcing leading to a common seasonal signal. In climate research the time series are typically *anomalized*, that is, the seasonal signal is estimated and subtracted from the data (Storch and Zwiers, 1999) which is equivalent to regressing out its influence. But this is not always possible, in our example the common signal is not purely periodic and cannot easily be estimated from the data. Another option for the case of piecewise stationary processes is to include background knowledge on the stationary regimes and estimate the graphs separately for the stationary subsets of $\mathcal{T}$. For example, the climatic seasons El Niño and La Niña lead to different causal directions of surface temperature anomalies in the tropical Pacific (Philander, 1985). Prior knowledge of when the seasons start and end allow us to restrict the estimation of time series graphs to samples within a particular season.

Now suppose we actually have access to the common signal U and include it in our analysis (but without estimating the

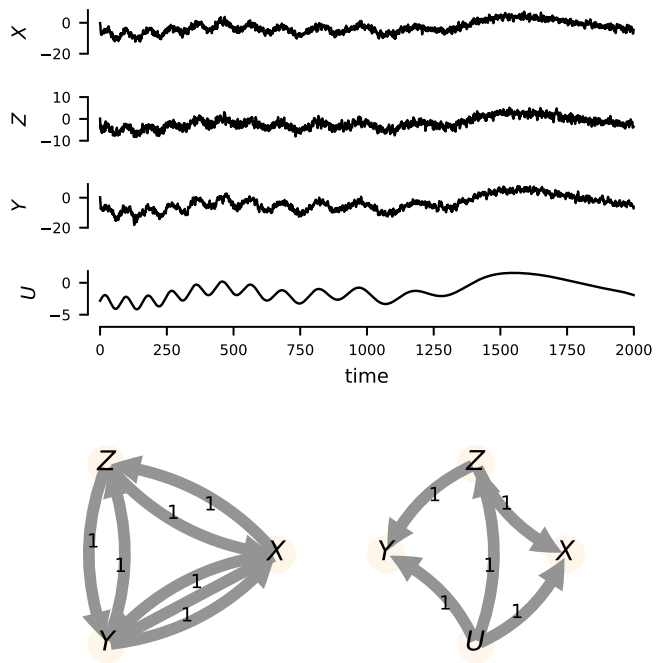

FIG. 8. Nonstationarity due to confounding. **(Top)** Example time series (X, Y, Z) that are nonstationary in mean and spectral properties due to a common signal U. **(Bottom left)** Access to only (X, Y, Z) results in a fully connected causal graph. **(Bottom right)** Including U allows us to identify the correct causal graph.

parents of U, i.e., treating U as exogenous). Then, as shown in the bottom right panel of Fig. 8, we can recover the true causal structure where U is a confounder of the three variables while they are connected through the motif $X \leftarrow Z \rightarrow Y$. Thus, here nonstationarity is a result of confounding and can be removed if we have access to the underlying trend.

The key point is that the causal structure, that is, the time series graph, of the whole process (X, Y, Z, U) *is* invariant in time. One may argue that causal laws are generally invariant and non-stationarity is simply a problem of violation of causal sufficiency. The idea of finding invariant predictors for causal inference is explored in Peters *et al.* (2016).

F. Dependency type assumptions

To test conditional independence hypotheses $X \perp\!\!\!\perp Y \mid Z$, different test statistics can be utilized. These are typically based on making certain assumptions about the type of the underlying dependency structure. While classical statistical methods are often based on the assumption of linearity (which allows us to derive rigorous results), modern statistics, the physics community, and the recent field of machine learning have developed non-parametric or model-free methods that allow us to better capture the nonlinear reality of many dynamical complex systems—at the cost of weaker theoretical results. Conditional independence testing can be classified into regression-based and model-free approaches. Here, we only discuss tests for continuously valued variables, for discrete variables one can, for example, use methods based on contingency tables (Spirtes *et al.*, 2000) or discrete CMI estimation (Cover and Thomas, 2006).

Regression-based conditional independence tests of $X \perp\!\!\!\perp Y|Z$ are based on first regressing out the influence of Z from X and Y and then testing the dependence between the residuals. We first fit a model assuming

$$\begin{aligned} X &= f_X(Z) + \epsilon_X, \\ Y &= f_Y(Z) + \epsilon_Y, \end{aligned} \tag{23}$$

for centered variables X, Y and independent and identically normally distributed $\epsilon_{X,Y}$. Now further restrictions can be laid upon the functional form of $f_{X,Y}$. For example, the partial correlation test assumes linearity, while a non-parametric regression can be based on Gaussian Process regression (Rasmussen and Williams, 2006).

Secondly, from the estimated functions $\hat{f}$, the residuals are formed as

$$\begin{aligned} r_X &= X - \hat{f}_X(Z) \\ r_Y &= Y - \hat{f}_Y(Z). \end{aligned} \tag{24}$$

Finally, the dependence between the residuals can be tested with different pairwise association tests. For partial correlation this is a t-test, while the dependence between the residuals of a non-parametric regression can be tested with non-parametric tests (Gretton *et al.*, 2008; Székely *et al.*, 2007) such as the *distance correlation coefficient* $\mathcal{R}(r_X, r_Y)$ (Székely *et al.*, 2007) (see also Sec. V C). Note that these models all make parametric assumptions and, thus, do not estimate conditional independence in its most general form.

The other extreme to partial correlation are model-free methods that directly test conditional independence. The most prominent test statistic is CMI as defined in Eq. (3), for which non-parametric estimators based on nearest-neighbor statistics exist (Kraskov *et al.*, 2004; Frenzel and Pompe, 2007; Vejmelka and Palus, 2008; Póczos and Schneider, 2012) [see also Gao *et al.* (2015) and Lord *et al.* (2018) for recent progress on nearest-neighbor entropy estimators]. Other possible conditional independence tests are Kernel Conditional Independence Tests (Zhang *et al.*, 2011; Strobl *et al.*, 2017) which essentially test for zero Hilbert-Schmidt norm of the partial cross-covariance operator or *conditional distance correlation* (Wang *et al.*, 2015). Some new recent tests are based on neural networks (Sen *et al.*, 2017) or decision tree regression (Chalupka *et al.*, 2018). In Runge (2018), a conditional independence test based on CMI is introduced.

Example 9. (Nonlinearity)

Figure 9 gives an overview over different types of linear and nonlinear relationships of the form

$$\begin{aligned} Z_t &= \eta_t^Z \sim \mathcal{N}(0,1), \\ X_t &= f_X(Z_{t-1}, \eta_t^X), \\ Y_t &= f_Y(Z_{t-2}, \eta_t^Y). \end{aligned} \tag{25}$$

In all cases, we have $X \perp\!\!\!\perp Y \mid Z$.

For the linear case (first row in Fig. 9), we consider $f_\cdot = cZ_{t-1,2} + \eta_t^\cdot$ and the regression-based techniques correctly fit the dependencies of the pairs (X, Z) and (Y, Z) (red fit lines in gray scatterplots), and, thus, correctly identify the

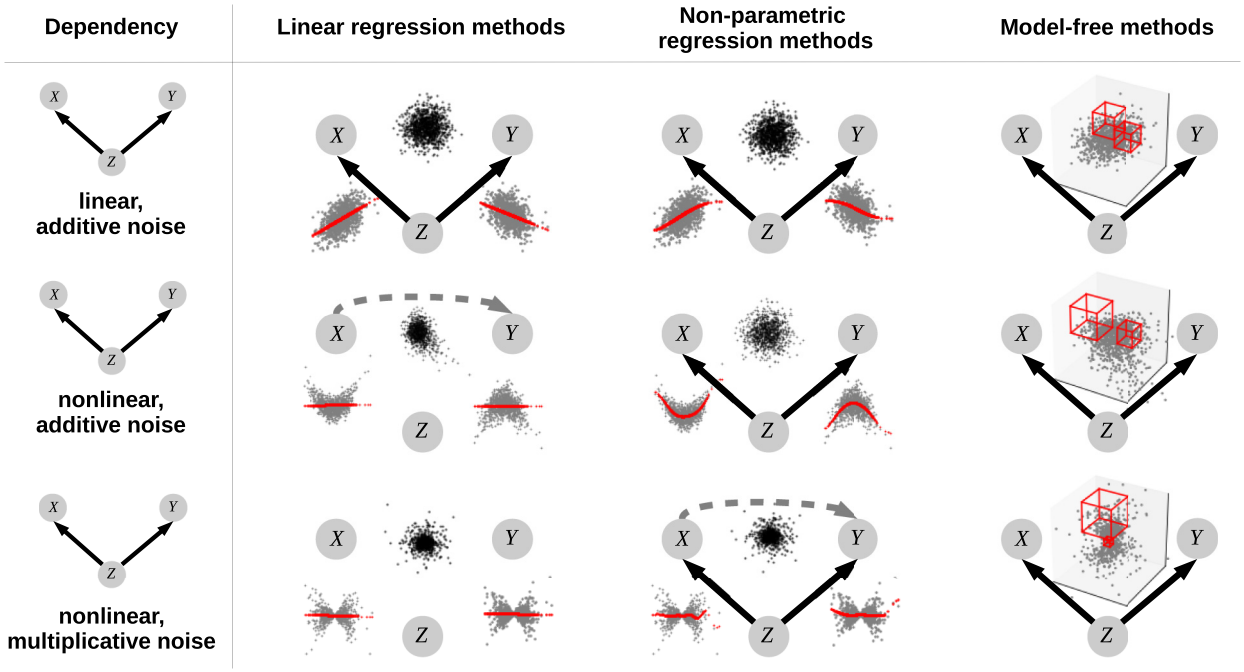

FIG. 9. Illustration of applicability of different conditional independence methods (linear and non-parametric regression-based and model-free) on different types of linear and nonlinear common driver models. Black arrows denote correctly identified causal links and dashed gray arrows indicate spurious links. The gray scatterplots with red fit line illustrate regressions of X and Y on Z and the black scatterplot the dependency between the residuals r_X, r_Y. The three-dimensional scatterplot with red cubes in the right column depicts the CMIknn test (Runge, 2018) which is based on data-adaptive nearest-neighbor estimation (the cubes are smaller for denser regions).

independence of the residuals (black scatterplot). A model-free test also correctly identifies the common driver motif here.

For the nonlinear additive noise case with a quadratic dependency $f_\cdot = c_\cdot Z_{t-1,2}^2 + \eta_t^\cdot$, partial correlation cannot fit the dependencies of the pairs (X, Z) and (Y, Z). As a result, the residuals are still correlated (spurious gray dashed link) and the causal graph is completely wrong: We overlook the links $X \to Z$ and $Y \to Z$ and get a false positive $X \to Y$. Since here the dependencies are still additive functions, non-parametric regressions and model-free tests yield a correct causal graph.

Finally, if the dependencies are multiplicative (bottom row) as in $f_\cdot = c_\cdot Z_{t-1,2} \cdot \eta_t^\cdot$, both regression methods fail. Then the residuals are nonlinearly related which is not detected with a partial correlation test (here two errors somewhat cancel each other out). A non-parametric test on the residuals, on the other hand, then wrongly estimates the spurious link $X \to Y$ (gray dashed in center bottom row).

Model-free methods in principle can deal with all these cases, which might lead to the conclusion that they are superior. But the "no-free-lunch-theorem" tells us that such generality has a price and model-free methods are very data-hungry and computationally expensive. If expert knowledge pointing to a linear or otherwise parametric dependency is available, then regression-based methods will typically greatly outperform model-free methods.

G. Measurement error

Measurement error, unlike dynamical noise, contaminates the variables between which we seek to reconstruct dependencies and constitutes a difficult problem in causal network reconstruction (Scheines and Ramsey, 2016).

Example 10. (Observational noise)

Here, we only discuss measurement error in its simple form as observational noise, which can be modeled as $\tilde{Z} = Z + \epsilon^Z$. Such observational noise presents at least two sorts of problems for causal discovery.

Firstly, observational noise attenuates true associations and, therefore, lowers detection power. This is because in general $I(\tilde{X}; \tilde{Y}) = I(X + \epsilon^X; Y + \epsilon^Y) \leq I(X; Y)$ which is a consequence of the data processing inequality (Cover and Thomas, 2006): Manipulating a variable can only reduce its information content. In Fig. 10, we added normal observational noise with $\sigma = 20$ to Z. Then the links $Z \to X$ and $Z \to Y$ cannot be reconstructed anymore. Secondly, too much noise on conditioning variables makes it impossible to preserve conditional independence. In Fig. 10, we have $I(X; Y \mid \tilde{Z}) = I(X; Y \mid Z + \epsilon^Z) > 0$ even though $I(X; Y \mid Z) = 0$. The

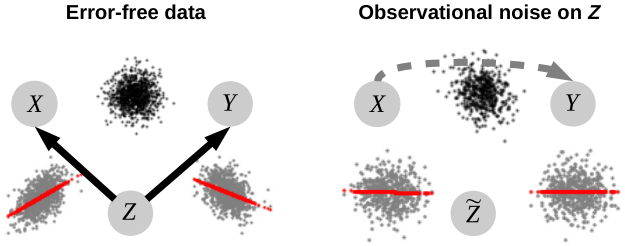

FIG. 10. Effect of observational noise on causal network reconstruction. Shown left is the linear example from Fig. 9. Very strong observational noise on Z (right panel) here leads to a vanishing correlation between X and $\tilde{Z}$ as well as Y and $\tilde{Z}$. Since then the effect of $\tilde{Z}$ cannot be regressed out anymore, we also get a spurious link $X \to Y$.

effect of observational noise on causal discovery is discussed further in Smirnov (2013); Scheines and Ramsey (2016) and Runge *et al.* (2018) give some numerical experiments for high-dimensional causal discovery.

V. PRACTICAL ESTIMATION

The previous sections concerned fundamental assumptions of time-lagged causal discovery based on conditional independence relations. We now turn to the topic of practical estimation where we introduce several common causal discovery methods and discuss their consistency, significance testing, and computational complexity. We restrict the analysis to the class of conditional independence approaches, which flexibly allows us to use different independence tests. But graphical models, in general, can also be estimated with score-based Bayesian methods, e.g., the max-min hill-climbing algorithm (Tsamardinos *et al.*, 2006).

A. Causal discovery algorithms

1. Granger causality/transfer entropy/FullCI

Transfer entropy, as introduced by Schreiber (2000), is a direct information-theoretic implementation of Granger causality (Barnett *et al.*, 2009). In a lag-specific implementation, as given by FullCI [Eq. (6)], it tests for conditional independence between each $X^i_{t-\tau}$ and X^j_t conditioned on the entire past $\mathbf{X}^{(t-1,\ldots,t-\tau_{\max})}_t$ (excluding $X^i_{t-\tau}$). As the experiments in Sec. VII will demonstrate, this approach strongly suffers from the curse of dimensionality.

2. Optimal causation entropy

Sun and Bollt (2014) and Sun *et al.* (2015) developed a discovery algorithm based on the information-theoretic *optimal causation entropy* principle (algorithm abbreviated as OCE) which reconstructs the lagged parents of a variable X^j_t by an iterative procedure alleviating the curse of dimensionality: Starting with an empty parent set $\widehat{\mathcal{P}}^{\rm OCE}(X^j_t) = \emptyset$, first the MIs $I(X^i_{t-\tau}; X^j_t)$ for all $X^i_{t-\tau} \in \mathbf{X}^-_t$ are estimated. As the first parent $X^{(1)}$, the one with the largest MI with X^j_t is selected. The next parent $X^{(2)}$, however, is chosen according to the largest CMI $I(X^i_{t-\tau}; X^j_t | X^{(1)})$ among all remaining variables, the third parent is the one with largest CMI conditional on the two previously selected parents, etc. The process is continued until the CMI of a selected parent is non-significant. This forward-selection stage is followed by a backward elimination whereby the significance of each of the parents $X^i_{t-\tau} \in \widehat{\mathcal{P}}^{\rm OCE}(X^j_t)$ is tested conditional on the remaining parents:

$$\begin{aligned} \text{OCE:} \quad & X^i_{t-\tau} \perp\!\!\!\perp X^j_t \mid \hat{\mathcal{P}}^{\rm OCE}(X^j_t) \setminus \{X^i_{t-\tau}\} \\ & \forall\, X^i_{t-\tau} \in \hat{\mathcal{P}}^{\rm OCE}(X^j_t). \end{aligned} \tag{26}$$

The significance of CMIs can be tested with a nearest-neighbor CMI estimator (Kraskov *et al.*, 2004; Frenzel and Pompe, 2007; Vejmelka and Palus, 2008) in combination with a permutation test where $X^i_{t-\tau}$ is randomly shuffled. Of course, the conditional independencies in Eq. (26) can also be tested with other test statistics.

3. PC algorithm

An alternative to this forward-backward scheme is the PC algorithm (named after its inventors Peter and Clark) (Spirtes and Glymour, 1991). The original PC algorithm was formulated for general random variables without assuming a time order. It consists of several phases where first, in the skeleton-discovery phase, an undirected graphical model (Lauritzen, 1996) is estimated whose links are then oriented using a set of logical rules (Spirtes and Glymour, 1991; Spirtes *et al.*, 2000). A later improvement led to the more robust modification called PC-stable (Colombo and Maathuis, 2014).

For the case of time series, we can use the information of time order which naturally provides an orientation rule for links. The algorithm then is as follows: For every variable $X^j_t \in \mathbf{X}_t$ it starts by initializing the preliminary parents $\hat{\mathcal{P}}(X^j_t) = (\mathbf{X}_{t-1}, \mathbf{X}_{t-2}, \ldots, \mathbf{X}_{t-\tau_{\max}})$. In the first iteration ($p = 0$), we remove a variable $X^i_{t-\tau}$ from $\widehat{\mathcal{P}}(X^j_t)$ if the null hypothesis

$$\text{PC}(p = 0): X^i_{t-\tau} \perp\!\!\!\perp X^j_t, \tag{27}$$

cannot be rejected at a significance threshold α. Then, in each next iteration, we increase $p \to p + 1$ and remove a variable $X^i_{t-\tau}$ from $\hat{\mathcal{P}}(X^j_t)$ if *any* of the null hypotheses

$$\text{PC}(p > 0): X^i_{t-\tau} \perp\!\!\!\perp X^j_t \mid \mathcal{S} \quad \text{for any } \mathcal{S} \text{ with } |\mathcal{S}| = p, \tag{28}$$

cannot be rejected, where $\mathcal{S}$ iterates (in an inner loop) through all combinations of subsets $\mathcal{S} \subseteq \hat{\mathcal{P}}(X^j_t) \setminus \{X^i_{t-\tau}\}$ with cardinality p. The algorithm converges for a link $X^i_{t-\tau} \to X^j_t$ once $\mathcal{S} = \hat{\mathcal{P}}(X^j_t) \setminus \{X^i_{t-\tau}\}$ and the null hypothesis $X^i_{t-\tau} \perp\!\!\!\perp X^j_t \mid \hat{\mathcal{P}}(X^j_t) \setminus \{X^i_{t-\tau}\}$ is rejected (if the null hypothesis cannot be rejected, the link is removed). Runge *et al.* (2018) provide pseudo-code for this algorithm.

The forward-backward scheme of OCE conducts conditional independence tests only using the conditions with highest CMI in the preceding stage and quickly increases the number of conditions. This can lead to wrong parents being kept in $\hat{\mathcal{P}}^{\rm OCE}(X^j_t)$ (see Sec. V B) which are only removed in the backward stage where the dimensionality of the set $\hat{\mathcal{P}}^{\rm OCE}(X^j_t)$ can be already quite high. High dimensionality, in principle, leads to lower detection power (Example VII C). The PC algorithm conducts conditional independence tests not only using the condition with highest association, but it goes through (in theory all) combinations of conditions $\mathcal{S}$ which can help to alleviate the curse of dimensionality regarding the estimation dimension of $I(X^i_{t-\tau}; X^j_t \mid \mathcal{S})$ compared to OCE. On the other hand, the PC algorithm conducts many more tests which increases other problems (see Sec. V C).

4. PCMCI

A more recent approach that addresses some of the shortcomings of the PC algorithm above is PCMCI (Runge *et al.*, 2018). PCMCI is a two-step approach which uses a version of the PC-algorithm only as a condition-selection step (PC_1 algorithm) to obtain $\hat{\mathcal{P}}(X^j_t)$ for all $X^j_t \in \mathbf{X}_t$, followed by the *momentary conditional independence* (MCI) test for

$X^i_{t-\tau} \to X^j_t$ defined as

$$\text{MCI:} \quad X^i_{t-\tau} \perp\!\!\!\perp X^j_t \mid \hat{\mathcal{P}}(X^j_t) \setminus \{X^i_{t-\tau}\}, \hat{\mathcal{P}}(X^i_{t-\tau}) \quad \forall\, X^i_{t-\tau} \in \mathbf{X}^-_t, \tag{29}$$

with $\mathbf{X}^-_t = (\mathbf{X}_{t-1}, \mathbf{X}_{t-2}, \ldots, \mathbf{X}_{t-\tau_{\max}})$. The main difference between PC and PC_1 is that PC_1 tests only the condition subset $\mathcal{S}$ with largest association instead of going through all possible combinations. To this end $\widehat{\mathcal{P}}(X^j_t)$ is sorted after every iteration according to the test statistic value and $\mathcal{S}$ is determined by the first p variables in $\widehat{\mathcal{P}}(X^j_t)$ (excluding $X^i_{t-\tau}$). This leads to less tests, but still provably removes incorrect links.

The MCI test is the most important difference to the PC algorithm and the approach by Sun and Bollt (2014). The additional conditioning on the parents $\widehat{\mathcal{P}}(X^i_{t-\tau})$ in MCI accounts for autocorrelation leading to well-controlled false positive rates at the expected level (Runge *et al.*, 2018). A variant (PCMCI_0) where the condition on the parents $\widehat{\mathcal{P}}(X^i_{t-\tau})$ is dropped leads to a very similar approach to OCE. PCMCI, like FullCI, OCE, and PC, can be implemented with different conditional independence tests. For further details on PCMCI, see Runge *et al.* (2018). In Sec. VII we compare FullCI, OCE, and PCMCI in a number of numerical comparison studies.

B. Consistency

Consistency is an important property of causal methods that tells us whether the method provably converges to the true causal graph in the limit of infinite sample size. Consistency concerns the conditional independence tests on the one hand, but also the causal algorithm in the case of iterative approaches such as those discussed in Sec. V A.

For example, for the consistency of the non-parametric regression independence tests in Eq. (23), we need to assume that the function estimator converges to the true function, that the noise in the model is additive and independent, and finally that we have a consistent unconditional independence test for the residuals. With a consistent test, the time series graph can be directly estimated based on Definition 1. For iterative causal algorithms, we can define *universal* consistency as follows.

Definition 7. (Universal causal consistency).

Denote by $\widehat{\mathcal{G}}_n$ the estimated graph of some causal estimator from a sample of a distribution P with sample size n and by $\mathcal{G}$ the true causal graph. Then a causal estimator is said to be universally consistent if $\widehat{\mathcal{G}}_n$ converges in probability to $\mathcal{G}$ for every distribution P,

$$\lim_{n\to\infty} Pr(\hat{\mathcal{G}}_n \neq \mathcal{G}) = 0. \tag{30}$$

That is, the probability of estimating the wrong graph becomes arbitrarily small if enough data is available, for any distribution P (hence "universal"). Consistency has been proven for classical causal discovery algorithms such as the PC-algorithm (Spirtes *et al.*, 2000), the optimal causation approach Sun and Bollt (2014), Sun *et al.* (2015) and PCMCI Runge *et al.* (2018), as an approach based on PC.

However, universal consistency is a weaker statement than, for example, *uniform consistency* which bounds the error probability as a function of the sample size n giving a *rate of convergence*. Thus, for a non-uniform, but only universally consistent method, the sample size at which a given error can be guaranteed and can be different for every distribution P. Robins *et al.* (2003) showed that no uniformly consistent causal discovery algorithm from the class of independence-based approaches (Spirtes *et al.*, 2000) exists since the convergence can always be made arbitrarily slow by a distribution that is *almost unfaithful* with some dependencies made arbitrarily small. Uniform consistency for conditional-independence based algorithms can only be achieved under further assumptions such as having strong enough dependencies (Kalisch, 2008).

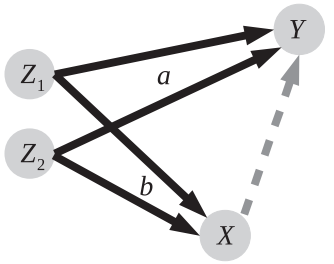

FIG. 11. Example where for certain a, b the MI $I(X;Y)$ can be larger than any of the MIs $I(Z_1;Y)$ or $I(Z_2;Y)$. Thus, the most strongly associated variable with Y is actually not a causal driver.

Example 11. (An inconsistent causal algorithm)

Consider again the forward-selection stage of the OCE algorithm (Sun and Bollt, 2014; Sun *et al.*, 2015) introduced in Sec. V as a standalone method to reconstruct parents of a variable $Y_t \in \mathbf{X}_t$. Even though the scheme sounds appealing and efficient, the scheme alone is not a consistent estimator of causal graphs. It yields a *superset* of the parents (Sun *et al.*, 2015) which may also contain false positives: Consider the example graph shown in Fig. 11. Here, the causal parents of Y are Z_1, Z_2 (dropping time subscripts t here). If forward-selection alone was a causal approach, then in each step the variable with strongest association would also need to be a causal parent. But in this example the MI between X and Y can be larger than the MIs of Z_1 and Z_2 with Y. For example, for $a = 0.5, b = 2$ we have $I(X;Y) \approx 0.13$ *nats* while $I(Z_1;Y) = I(Z_2;Y) \approx 0.06$ *nats*. See Appendix A for an information-theoretic analysis. Hence, the wrong parent X is selected. This scheme, thus, requires the second step of the OCE approach, given by Eq. (26).

C. Significance testing

How can we assess the statistical significance of conditional independence tests on which the causal algorithms in Sec. V A are based, such as the tests discussed in Sec. IV F? Using a test statistic $\widehat{I}_n(\mathbf{x};\mathbf{y} \mid \mathbf{z})$ (I here stands not only for CMI, but any conditional independence test statistic) for the observed samples $(\mathbf{x},\mathbf{y},\mathbf{z}) = \{x_i, y_i, z_i\}_{i=1}^n$ we wish to test the hypothesis

$$H_0 : X \perp\!\!\!\perp Y \mid Z, \tag{31}$$

versus the general alternative

$$H_1 : X \not\perp\!\!\!\perp Y \mid Z. \tag{32}$$

To assess the significance of an outcome of such a test using p-values, we need to know the distribution $Pr(I_n \mid H_0)$ of the test statistic under the null hypothesis. For partial correlation tests, exact analytical expressions of the null distribution exist under certain assumptions, but for nonlinear tests (such as CMI or also non-parametric regression tests) these are typically not available except for some asymptotic large sample size cases (Strobl *et al.*, 2017). The alternative then are permutation tests as discussed in Example 12.

Given the null distribution, the p-value is defined as the probability—given H_0—of observing a value of the test statistic that is the same or more extreme than what was actually observed. If our test statistic is non-negative (such as CMI), the p-value for an observed test statistic value $\widehat{I}_n$ is defined as $p = Pr(I_n \geq \widehat{I}_n \mid H_0)$. Choosing a *significance level* α, we reject the null hypothesis if $p < \alpha$.

There are two types of errors we can make. Rejecting H_0 when H_0 is true is called a type I error or false positive. Retaining H_0 when H_1 is true is called a type II error or false negative. We also call one minus the type II error rate of a test the true positive rate or detection rate.

If the test statistic has a continuous distribution, then under H_0 the p-value has a uniform $(0, 1)$ distribution (Wasserman, 2004). Therefore, if we reject H_0 when the p-value is less than α, the probability of a type I error is α. For a well-calibrated test under H_0, we thus expect to measure on average a false positive rate of α. If this is not the case, the test is ill-calibrated indicating that we got the null distribution wrong because certain assumptions, such as independent samples, are violated. In Examples 12 and 13, we discuss such cases.

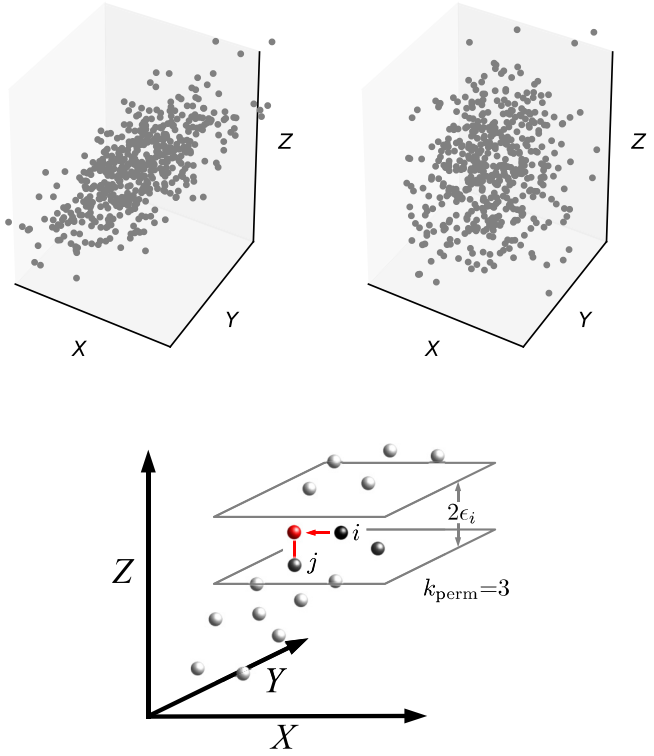

FIG. 12. Permutation approach to conditional independence testing. **(Top left)** Example sample drawn from a common driver scheme $X \leftarrow Z \rightarrow Y$. **(Top right)** Permuted sample with randomly shuffled data points $\mathbf{x}$, which destroys the associations between $\mathbf{x}$ and $\mathbf{y}$, but also between $\mathbf{x}$ and $\mathbf{z}$ leading to ill-calibrated tests. **(Bottom)** Schematic of local permutation scheme. Each sample point i's x-value is mapped randomly to one of its k_{perm}-nearest neighbors in subspace $\mathcal{Z}$ (see Runge, 2018) to preserve dependencies between $\mathbf{x}$ and $\mathbf{z}$.

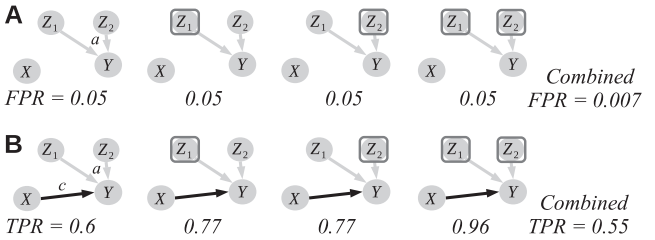

FIG. 13. The problem of sequential testing for $X \rightarrow Y$ conditional on other variables (gray boxes). **(A)** While the false positive rates of each individual test are as expected at $\alpha = 0.05$, the combined false positive rate of the sequence of tests is much lower. **(B)** Similarly, the combined true positive rate is lower than the minimal true positive rate among the individual tests. Gaussian noise model with coefficients $a = 1$, $c = 0.4$; rates estimated from 5000 realizations with sample size $n = 100$.

Example 12. (Permutation testing)

Permutation testing is straightforward in the bivariate independence test case. To create an estimate of the null distribution of a test statistic $\widehat{I}_n(\mathbf{x}; \mathbf{y})$, we can simply generate a large number of test statistics $\widehat{I}_n(\mathbf{x}^*; \mathbf{y})$ where $\mathbf{x}^*$ is a permuted version of $\mathbf{x}$.

But how to permute for *conditional* independence testing of $X \perp\!\!\!\perp Y \mid Z$? In the top left panel of Fig. 12, we illustrate an example scatterplot of a sample drawn from a common driver scheme $X \leftarrow Z \rightarrow Y$ where we have $X \perp\!\!\!\perp Y \mid Z$. If we now permute $\mathbf{x}$, we get the sample shown in the top right panel. Here, any association between $\mathbf{x}$ and $\mathbf{y}$ is indeed destroyed, but we also destroyed the association between $\mathbf{x}$ and $\mathbf{z}$. That is, for the permuted sample we have

$$\widehat{I}(\mathbf{x}^*; \mathbf{y}) \approx 0 \quad \text{and} \quad \widehat{I}(\mathbf{x}^*; \mathbf{z}) \approx 0 \tag{33}$$

but what we actually want to achieve is

$$\widehat{I}(\mathbf{x}^*; \mathbf{y} \mid \mathbf{z}) \approx 0 \tag{34}$$

$$\text{with} \quad \widehat{I}(\mathbf{x}^*; \mathbf{z}) \approx \widehat{I}(\mathbf{x}; \mathbf{z}) \tag{35}$$

in order to test the correct null hypothesis. The above global permutation scheme results in inflated false positives as, for example, shown in Runge (2018) and for FullCI and OCE in the numerical comparison studies in Fig. 14.

To achieve a test under the correct null hypothesis, we can use a *local permutation scheme* that preserves the associations between $\mathbf{x}$ and $\mathbf{z}$. Runge (2018) suggests such a scheme depicted in the bottom panel of Fig. 12 which only permutes those x_i and x_j where $z_i \approx z_j$. This scheme can be used for CMI conditional independence testing or also other test statistics. Other schemes are discussed in Doran *et al.* (2014) and Sen *et al.* (2017).

Example 13. (Non-independent samples)

A basic assumption underlying many conditional independence tests is that the samples are independent and

identically distributed (*i.i.d.*). Unfortunately, time series are typically dependent in time. To take this dependence into account, one can either adapt the distribution under the null hypothesis, for example, in partial correlation *t*-tests by estimating the *effective degrees of freedom*, which is, however, difficult in a multivariate setting. Or one can modify the test statistic to explicitly account for autocorrelation (Runge *et al.*, 2018). Also, a permutation scheme needs to be adapted to preserve auto-dependencies, for example, by shuffling blocks of samples (Peifer *et al.*, 2005). For bivariate tests such an approach is again straightforward, but not for the multivariate case as analyzed in the autocorrelation comparison study in Sec. VII.

Example 14. (Sequential testing of causal links)

The preceding discussion concerned tests of an individual conditional independence relationship. Directly testing causal links via Definition 1, thus, gives us a well-calibrated test if all assumptions are fulfilled.

However, in iterative causal algorithms (Sec. V A) such as the PC algorithm, OCE, or the first step of PCMCI (PC_1) *multiple* tests on a particular link $X_{t-\tau} \to Y_t$ are conducted with different condition sets that are determined by the outcome of previous tests. If a link is found non-significant in *any* of these tests, this link is removed. Since the tests are not independent of each other (since they are typically based on the same data sample), it is almost impossible to derive a combined *p*-value of all these tests.

Figure 13 depicts an illustrative numerical example where the combined false positive rate is much lower than the 0.05 of each individual test and the true positive rate is lower than the minimal true positive rate among the individual tests. In summary, even though all assumptions may be valid, sequential testing makes a significance assessment difficult. These issues are further discussed in Tsamardinos and Brown (2008) and Strobl and Spirtes (2016) and references therein where False Discovery Rate (Benjamini and Hochberg, 1995) approaches are discussed.

As a side remark on the previously discussed OCE and PCMCI causal discovery approaches, in the backward-elimination stage [Eq. (26)] OCE tests only the significance of each of the parents $X^i_{t-\tau} \in \widehat{\mathcal{P}}^{\text{OCE}}(X^j_t)$ conditional on the remaining parents. PCMCI (Runge *et al.*, 2018), on the other hand, in the second MCI step tests *all* links again [Eq. (29)] which makes the MCI test slightly less dependent on the sequential testing issue of the condition-selection step PC_1 since parents that have been removed in PC_1 (false negatives) are tested again in the MCI test. In particular, the false positives are as expected as demonstrated in the comparison studies in Sec. VII.

D. Computational complexity

Application areas of causal discovery methods vary in the typical numbers of variables as well as available sample sizes n. Next to the properties discussed before, an important issue then is how a method scales with dimensionality and sample size. High-dimensionality arises from the number of included variables N and the maximum time lag $\tau_{\max}$ [see Fig. 1(c)] and has at least two consequences: (1) higher computational complexity leading to longer runtimes and (2) typically lower detection power. Independence tests may also become ill-calibrated in high-dimensions.

For directly testing causal links via Definition 1 (FullCI), the computational complexity depends on the complexity of a single high-dimensional conditional independence test. In the linear partial correlation case, OLS regression scales $\sim \mathcal{O}(n(N\tau_{\max})^2)$. FullCI estimated using nearest-neighbor estimators of CMI (Kraskov *et al.*, 2004; Frenzel and Pompe, 2007), on the other hand, will scale $\sim \mathcal{O}(n \log n)$ regarding time complexity while the complexity in $N\tau_{\max}$ will depend on algorithmic details such as using efficient KD-tree nearest-neighbor search procedures (Maneewongvatana and Mount, 1999).

The methods PC, OCE, or PCMCI (Sec. V A) based on a condition-selection step avoid high-dimensional conditional independence estimation by conducting more tests with lower dimensional conditioning sets. Their theoretical complexities are difficult to evaluate, for numerical evaluations see Sun *et al.* (2015) and Runge *et al.* (2018), but typically they scale polynomially in time.

The other major challenge with high dimensionality is detection power as analyzed in Example VII C.

VI. PERFORMANCE EVALUATION CRITERIA

How can causal discovery methods, such as those described in Sec. V be evaluated? Typically, we want to know which method performs best on data from the kind of system we wish to study. Ideally, we would like to compare different methods on a data sample where the underlying causal truth is known or evaluate methods by experimentally manipulating a system, i.e., actually performing the *do*-experiment (Pearl, 2000) mentioned in the introduction which forms the theoretical basis of the present concept of causality. Since both of these options are mostly not available, an alternative is to construct synthetic model data where the underlying ground truth is known. These can then be used to study the performance of causal methods for realistic finite sample situations.

A. Models

To evaluate causal methods on synthetic data, several aspects for constructing model systems are relevant:

1. **Model realism:** The model systems should mimic the domain-specific properties of real data in terms of nonlinearity, autocorrelation, spectral properties, noise structure (dynamical as well as observational), etc.
2. **Model diversity:** To avoid biased conclusions, a large number of different randomly selected connectivity structures should be tested [including link density as well as properties such as small-worldness (Watts and Strogatz, 1998)]. For example, the aforementioned forward-selection approach failed for the example shown in Fig. 11 but works for many other graphs. But also consistent methods may have biases for finite samples as studied in Runge *et al.* (2018).
3. **Model dimensionality:** As studied in Fig. 16, a method may perform well only for a small number of variables

and the performance for high-dimensional settings (large networks) can quickly degrade.

4. **Sample sizes:** The comparative performance of different models may vary widely for different sample sizes which needs to be studied if no uniform consistency results are available.

B. Metrics

The performance of a causal method on a single realization of a model does not allow for reliable conclusions. Therefore, each model needs to be evaluated from many realizations. Then the most straightforward evaluation metric is to measure false positive rates and true positive rates for a given α as shown in the comparison studies in Sec. VII. The number of realizations should be chosen high enough since the error of a false or true positive rate r for B realizations is given by $\sigma_r = \sqrt{r(1-r)/B}$. Alternative evaluation metrics that do not depend on a particular significance level but directly on the p-values are the Kullback-Leibler divergence to evaluate whether the p-values are uniformly distributed (to measure how well-calibrated a test is) and the Area Under the Power Curve (AUPC) to evaluate true positives.

Next to the true and false positives of a causal method for finite samples, another performance criterion is computational runtime, though this may strongly depend on a given implementation.

VII. COMPARISON STUDIES

In this section, we compare several common causal discovery methods in three numerical comparison studies highlighting the effect of dynamical noise in deterministic chaotic systems, autocorrelation, and high dimensionality.

A. Dynamical noise in deterministic chaotic systems

In Example 6, we studied a static example of determinism. Here, we evaluate the effect of dynamical noise in a system of coupled chaotic logistic maps:

$$\begin{aligned} Z_t &= Z_{t-1}(r - rZ_{t-1} + \sigma\eta_t^Z) \mod 1, \\ X_t &= X_{t-1}(r - rX_{t-1} - Z_{t-1} + \sigma\eta_t^X) \mod 1, \\ Y_t &= Y_{t-1}(r - rY_{t-1} - Z_{t-1} + \sigma\eta_t^Y) \mod 1, \end{aligned} \tag{36}$$

(37)

with uniformly distributed independent noise η and $r = 4$ leading to chaotic dynamics. Here, σ controls the amount of dynamical noise in the system. To evaluate true positive rates (correctly detecting $Z_{t-1} \to X_t$ and $Z_{t-1} \to Y_t$) and false positive rates (incorrect detections for any other variable pair, direction, or lag), 200 realizations with time series length $n = 150$ were generated.

We compare three methods from an information-theoretic framework (FullCI, OCE, PCMCI) with *convergent-cross mapping* (CCM, Sugihara *et al.*, 2012) (see also Arnhold *et al.*, 1999; Hirata *et al.*, 2016) as a nonlinear dynamics-inspired approach. The significance of CMIs in FullCI and OCE is tested with a nearest-neighbor CMI estimator (Kraskov *et al.*, 2004; Frenzel and Pompe, 2007; Vejmelka and Palus, 2008) in combination with a permutation test where $X^i_{t-\tau}$ is randomly shuffled. PCMCI is implemented with the CMIknn independence test (Runge, 2018) as discussed in Example 12. We also evaluate a variant (PCMCI$_0$) where the condition on the parents $\hat{\mathcal{P}}(X^i_{t-\tau})$ is dropped and the only difference to OCE is the condition-selection step. CCM reconstructs the variable's state-spaces using lagged coordinate embedding and concludes on $X \to Y$ if points on X can be well predicted using nearest neighbors in the state space of Y. Note that CCM and related works only use the time series of X and Y with the underlying assumption that the dynamics of Z can be reconstructed using delay embedding. All methods were evaluated at a significance level of 0.05. For implementation details, see Appendix B 1. In the top panel of Fig. 14, we depict the true causal graph.

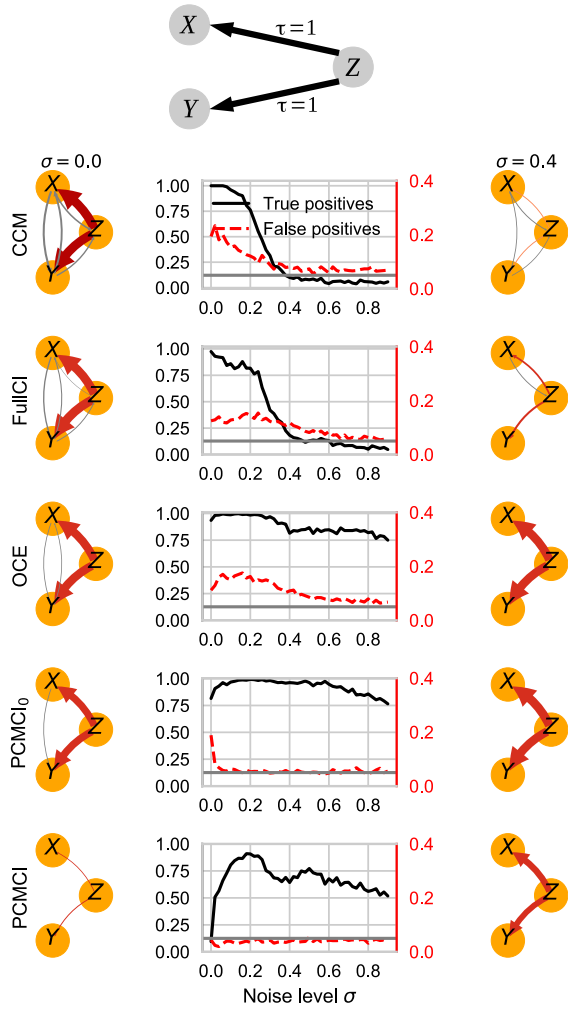

FIG. 14. Comparison of CCM, FullCI, OCE, and two versions of PCMCI on common driver system of three coupled chaotic logistic maps. The top panel shows the true graph structure. In the left and right graphs for noise levels $\sigma = 0$ and $\sigma = 0.4$, respectively, the width of arrows denotes the detection rate, gray edges depict false links (only false positive rates 0.08 shown). The center panels depict average true (black, left axis) and false positive rates (red, right axis) for different strengths σ of dynamical noise. The gray line marks the 5% significance threshold.

Figure 14 shows that in the purely deterministic regime with $\sigma = 0$, PCMCI has almost no power, while PCMCI$_0$ features a detection rate of 0.8 and FullCI, OCE, and CCM

almost always detect the couplings. On the other hand, CCM also has the highest number of false positives around 0.2 which exceeds the significance level indicating an ill-calibrated test. FullCI, OCE, and PCMCI$_0$ also do not control false positives well, while with PCMCI they are around the expected level.

For higher dynamical noise levels interesting behavior emerges: FullCI and CCM have continuously decreasing power (and slightly decreasing false positives) dropping to almost zero at $\sigma = 0.4$, while the power of OCE and both PCMCI versions steadily increases up to $\sigma = 0.2$ after which power decreases with a more pronounced decrease for PCMCI and PCMCI$_0$ having the highest power. False positives are above the 0.05 threshold for FullCI and OCE for a wide range of noise levels, while for PCMCI false positives are always well-controlled at the expected 0.05.

How can these results be understood? CCM attempts to reconstruct the attractor manifolds underlying X and Y. With more dynamical noise, this reconstruction becomes more difficult and CCM looses power. The fact that CCM does not control false positives well, especially in the deterministic regime deserved further study. FullCI suffers from the curse of dimensionality especially for higher noise levels.

The MCI test statistic for the link $Z_{t-1} \to X_t$ estimates $I^{\mathrm{MCI}}_{Z_{t-1}\to X_t} = I(Z_{t-1}; X_t \mid Z_{t-2}, X_{t-1})$. Now since for $\sigma = 0$ Z_{t-1} is a deterministic mapping of Z_{t-2}, in theory we have $I^{\mathrm{MCI}} = 0$ for the same reasons as in the simple deterministic model (19) above (and analogously for $Z_{t-1} \to Y_t$). In practice, we can only measure the entropies at some coarse-grained level (here determined by the CMI nearest-neighbor parameter) and the deterministic dependency is never exactly recovered leading to a non-zero MCI. In OCE and PCMCI$_0$, only the parents of X_t are included in the condition. The fact that FullCI, despite conditioning on the whole past, also detects the link deserves further study. Possible explanations are an ill-calibrated test or dynamical properties of the logistic-map system. In summary, purely deterministic dynamics here seem to generate too little momentary information which is necessary for information-theoretic coupling detection with MCI (see also Pompe and Runge, 2011).

Given at least some dynamical noise to suffice the Faithfulness condition and together with the other assumptions discussed in this paper, OCE and PCMCI provably converge to the true causal graph in the limit of infinite sample size (Runge *et al.*, 2018; Sun *et al.*, 2015, see also Sec. V B), while no such theoretical results is available for CCM. In the infinite sample limit also the inflated false positives of OCE due to time-dependent samples vanish. For finite samples, on the other hand, among other factors, consistency depends on how well-calibrated the significance test is. PCMCI here always yields expected levels of false positives. The inflated false positives for FullCI and OCE for a wide range of noise levels and for PCMCI$_0$ for very low dynamical noise is related to the way significance testing is implemented. In theory, OCE should have less false positives than the expected 0.05 due to the sequential testing problem (Sec. V C), but in our examples autocorrelation and the global permutation scheme leads to ill-calibrated tests with inflated false positives in each individual test, see Examples 12 and 13, leading to an overall higher false positive rate. The effect of autocorrelation is evaluated further in the next example.

B. Autocorrelation

In Fig. 15, we evaluate the previously introduced causal algorithms (Sec. V A) FullCI, OCE, and PCMCI on autocorrelated data which is an ubiquitous feature in real world time series data. The full model setup is described in Appendix B 2. In short, here we only evaluate the false positive rates (for $c = 0$) and true positive rates (for $c \neq 0$) of the link $X_{t-1} \to Y_t$ (top panel of Fig. 15) where the autocorrelation a is varied for different numbers of common drivers D_Z and the coefficients b and σ_Z are chosen such that the unconditional dependence stays the same, and we only investigate the effect of autocorrelation. The time series length is $n = 150$ and 1000 realizations were run for each model setup to evaluate false positive rates and true positive rates at an $\alpha = 0.05$ significance level.

We compare partial correlation implementations (test statistic ρ) of the following tests: (1) FullCI directly tests Definition 1, $X_{t-1} \perp\!\!\!\perp Y_t \mid \mathbf{X}_t^{(t-1,\ldots,t-\tau_{\max})} \setminus \{X_{t-1}\}$ for $\tau_{\max} = 5$. For OCE and PCMCI, we assume that the condition-selection steps already picked the correct parent sets and only test the link $X_{t-1} \to Y_t$ in the second stages of OCE and PCMCI, respectively: (2) OCE [Eq. (26), equivalent to PCMCI$_0$] here tests $X_{t-1} \perp\!\!\!\perp Y_t \mid \mathcal{P}_{Y_t}$, where $\mathcal{P}_{Y_t}$ is given by the gray and blue boxes in the top panel of Fig. 15. (3) OCEpw with conditioning set as for OCE, but where all variables are pre-whitened beforehand as described in Appendix B 2. (4) OCEbs with conditioning set as for OCE, but where a block-shuffle test was used as described in Appendix B 2. (5) PCMCI tests $X_{t-1} \perp\!\!\!\perp Y_t \mid \mathcal{P}_{Y_t}, \mathcal{P}_{X_{t-1}}$ [Eq. (29)] as given by the gray, blue, and red boxes. For all approaches, the analytical null distribution of the partial correlation test statistic was used (t-test), except for the block-shuffle permutation test.

In the bottom four panels of Fig. 15, we depict results for $D_Z = 0$, that is, the bivariate case, and $D_Z = 4$, both for varying the autocorrelation strength a. In the bivariate case, all approaches well control the false positive rates except for OCE and (slightly better) OCEbs, which feature inflated false positive rates for very high autocorrelation. The reason is that when testing $\rho^{\mathrm{OCE}} = \rho(X_{t-1}; Y_t | Y_{t-1})$ with the t-test, we assume i.i.d.-data, but since X is autocorrelated, this is not the case. This false positive inflation is also seen in Fig. 14. Here, pre-whitening removes this autocorrelation and block-shuffling remedies it to some extent. PCMCI and FullCI both condition out autocorrelation and well-control false positives with constant true positive levels, independent of a, while the true positive level depends on a for OCE and its modifications, even though the coupling coefficient c is constant.

For $D_Z = 4$ false positive inflation becomes even more severe for OCE and here also pre-whitening does not help but leads to strongly increased false positive rates since univariate pre-whitening is not suitable for multivariate conditional independence testing. The PCMCI approach conditions on the parents of the lagged variable which helps to exclude autocorrelation as shown in Runge *et al.* (2018) and allows us to utilize analytical null distributions that assume i.i.d. data.

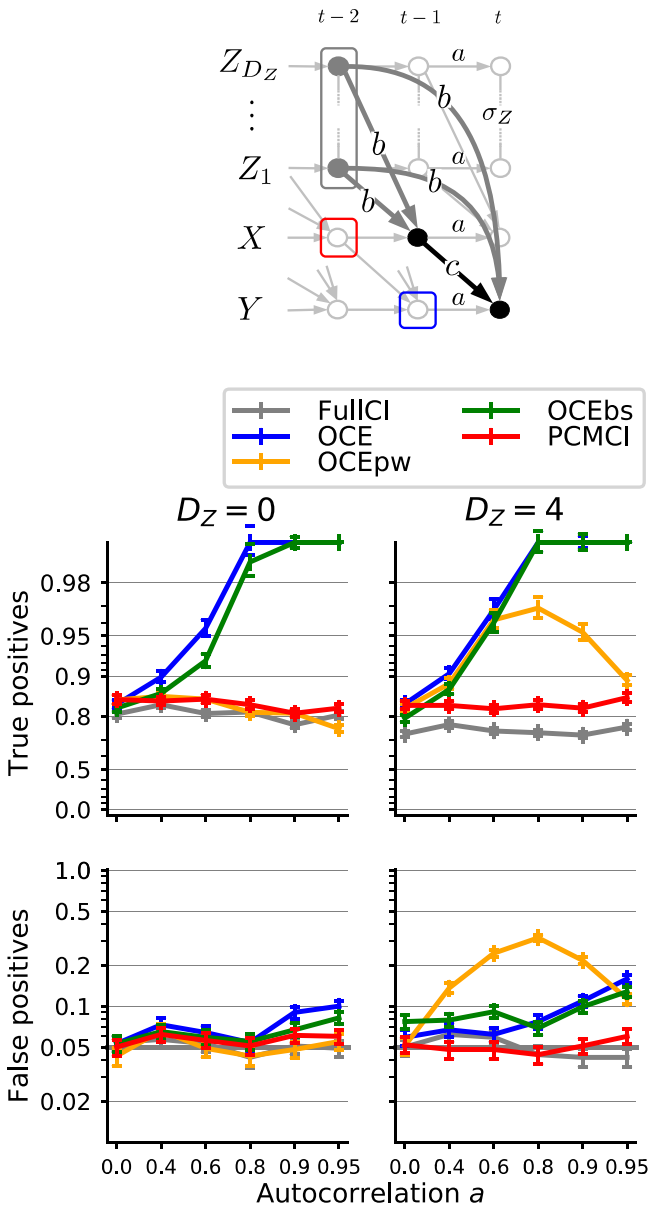

FIG. 15. Comparison of FullCI, three versions of OCE, and PCMCI under strong autocorrelation. **(Top panel)** Model time series graph with labels denoting linear coefficients. The full model setup is described in Appendix B 2. FullCI has the conditioning set $\mathbf{X}_t^- = (\mathbf{X}_{t-1}, \ldots, \mathbf{X}_{t-5})$, OCE has conditioning set depicted by gray and blue boxes, and PCMCI has conditioning set depicted by gray, blue, and red boxes. **(Bottom panel)** False positive rates and true positive rates are shown for two different dimensions D_Z and various autocorrelation strengths a.

C. Curse of dimensionality

In Fig. 16, we evaluate the causal algorithms (Sec. V A) for high-dimensional data using the same model as in Fig. 15 described in Appendix B 2. Here only FullCI, OCE, and PCMCI are compared, all of them again based on partial correlation.

As shown in Fig. 16, FullCI severely suffers from the curse of dimensionality and the OLS-solver even becomes ill-conditioned for $D_Z = 32$ since then the estimation dimension exceeds the sample size. For OCE and PCMCI, we again assume that the condition-selection algorithms selected the correct set of parents. Then the dimensionality of OCE and

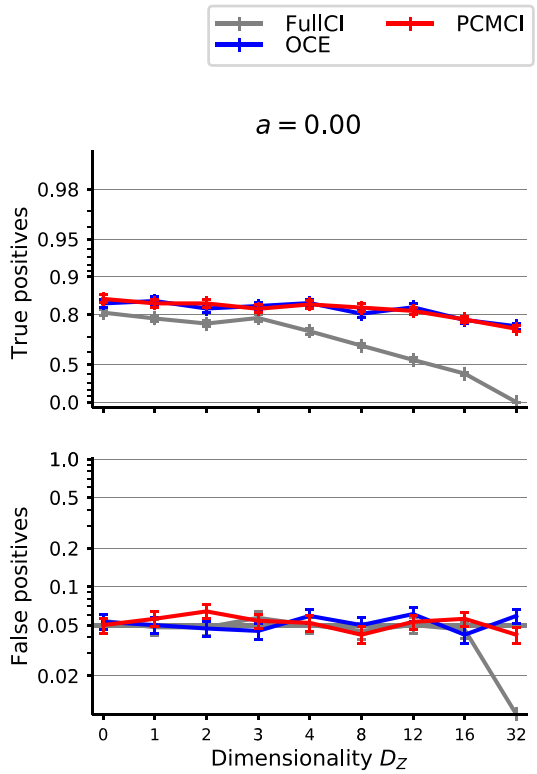

FIG. 16. Comparison of FullCI, OCE, and PCMCI under high dimensionality. The model setup is the same as in Fig. 15. Shown are false positive rates and true positive rates for different dimensions D_Z and no autocorrelation ($a = 0$) in the model detailed in Appendix B 2.

PCMCI increases only slightly and the power stays at higher levels.

The problem becomes even more severe for nonparametric tests such as multivariate transfer entropy. Another alternative to OLS partial correlation estimation are regularization techniques such as ridge regression (Hoerl *et al.*, 1970; Tibshirani, 1996; Tikhonov, 1963), but these come with other difficulties, for example regarding significance testing. These issues are further analyzed in Runge *et al.* (2018).

VIII. DISCUSSION AND CONCLUSIONS

The long preceding list of theoretical assumptions in Sec. IV may make causal discovery seem a daunting task. In most real data application, we will not have all common drivers measured, hence violating the causal sufficiency assumption. Then also the Markov assumption may be violated in many cases due to time-aggregation. A conclusion on the *existence* of a causal link, thus, rests on a number of partially strong assumptions. So what can we learn from an estimated causal graph? Let us consider what the assumptions on the *absence* of a causal link are.

Remark 1. *Let* $\tilde{\mathbf{X}}$ *be measurements of a stochastic process* $\mathbf{X}$. *Assuming Faithfulness* (*Definition* 5) *and that all variables in* $\tilde{\mathbf{X}}$ *are measured without error we have that for* $\tilde{X}, \tilde{Y}, \tilde{\mathbf{Z}} \in \tilde{\mathbf{X}}$ *with* $\tilde{X}, \tilde{Y} \notin \tilde{\mathbf{Z}}$ if

$$\tilde{X}_{t-\tau} \perp\!\!\!\perp \tilde{Y}_t \mid \tilde{\mathbf{Z}} \quad \textit{for any subset } \tilde{\mathbf{Z}} \in \tilde{\mathbf{X}}_t^- \Rightarrow X_{t-\tau} \not\to Y_t, \tag{38}$$

that is, if independence is measured given any subset of conditions, then there is no direct causal link between $X_{t-\tau}$ and Y_t in $\mathcal{G}$.

Note that for the practical estimation from finite data, we also need to assume that all dependencies among lagged variables in $\mathbf{X}$ can be modeled with the conditional independence test statistic, that is, no false negative errors occur. Nevertheless, Remark 1 rests on much weaker assumptions than the existence of a causal link since it does not require Causal Sufficiency or Causal the Markov Condition. The proof follows almost directly from the Faithfulness assumption.

Proof. Firstly, since we assume an error-free measurement process, we have that $\tilde{\mathbf{X}} = \mathbf{X}$. Then $\tilde{X}_{t-\tau} \perp\!\!\!\perp \tilde{Y}_t \mid \tilde{\mathbf{Z}} \Rightarrow X_{t-\tau} \perp\!\!\!\perp Y_t \mid \mathbf{Z} \Rightarrow X_{t-\tau} \bowtie Y_t \mid \mathbf{Z}$. The last relation is the Faithfulness assumption. Separation implies in particular that no *direct* link $X_{t-\tau} \to Y_t$ exists in $\mathcal{G}$. □

The second set of assumptions important for causal discovery are the assumptions underlying significance testing (Sec. V C). Failing to properly take into account autocorrelation or too simple permutation schemes imply ill-calibrated significance tests leading to inflated false positives beyond those expected by the significance level (see the comparison studies in Sec. VII). Next to the theoretical causal assumptions, statistical reliability of reconstructed networks is an important aspect for drawing causal conclusions.

This paper is intended to recapitulate the main concepts of time-lagged causal discovery from observational time series data and accessibly illustrate important challenges. But many more challenges exist, for example, we have not considered selection bias or issues with the definition of variables as elaborated on in Spirtes *et al.* (2000). We also have not discussed the topic of determining causal effects (Pearl, 2000) (causal quantification) or mediation (VanderWeele, 2015; Runge *et al.*, 2015a, 2015b) as opposed the pure existence or absence of causal links presented here.

Our focus was on time series which make the causal discovery problem easier in some aspects (e.g., time order can be exploited), but more difficult in other aspects, especially regarding statistical testing. We have briefly mentioned the recent works based on different sets of assumptions in the framework of structural causal modeling (Peters *et al.*, 2017), which do not require time-order. Also, many more techniques and insights from the conditional independence framework (Spirtes *et al.*, 2000) can be utilized in the time series case. An important conclusion is that causal discovery is a very active area of research in many fields, from mathematics, computer science, and physics to applied sciences, and methodological progress can greatly benefit from more interdisciplinary exchange.

ACKNOWLEDGMENTS

J.R. thanks C. Glymour and F. Eberhardt for comments and D. Sejdinovic, K. Zhang, and J. Peters for helpful discussions. Special thanks to C. Linstead for help with high-performance computing. J.R. received funding from a postdoctoral award by the James S. McDonnell Foundation and gratefully acknowledges the European Regional Development Fund (ERDF), the German Federal Ministry of Education and Research, and the Land Brandenburg for supporting this project by providing resources on the high-performance computer system at the Potsdam Institute for Climate Impact Research. Software is available online under https://github.com/jakobrunge/tigramite.

APPENDIX A: INCONSISTENT CAUSAL ALGORITHM

For the example graph shown in Fig. 11, consider the following decomposition [chain rule (Cover and Thomas, 2006), dropping t]:

$$I(X, Z_1, Z_2; Y) = I(X; Y) + I(Z_1, Z_2; Y \mid X). \tag{A1}$$

Alternatively, one can decompose the same MI as

$$I(X, Z_1, Z_2; Y) = I(Z_1; Y) + I(Z_2; Y \mid Z_1) + \underbrace{I(X; Y \mid Z_1, Z_2)}_{=0}, \tag{A2}$$

where the last term vanishes because (Z_1, Z_2) separates X and Y in the graph (Markov condition). From these two equations, it follows that if

$$I(Z_1, Z_2; Y \mid X) < I(Z_2; Y \mid Z_1) \tag{A3}$$

$$\Longrightarrow\ I(Z_1; Y) < I(X; Y). \tag{A4}$$

Hence, the wrong parent X has higher MI and would be selected with a pure forward-selection scheme.

APPENDIX B: IMPLEMENTATION DETAILS FOR NUMERICAL EXAMPLES

1. Dynamical noise model

CCM was estimated with embedding dimension $E = 2$, and the surrogate test ebisuzaki with 500 surrogates using the R-package rEDM. CCM requires two criteria (Sugihara *et al.*, 2012), both a significant CCM value at library size n and an increasing CCM value over increasing library length. As a p-value of CCM, we thus take $\max(p_n, p_{conv})$, where p_n is the p-value of CCM at library size n and p_{conv} is the p-value for the hypothesis of an increasing linear trend. OCE was estimated with threshold parameter $\alpha_{\mathrm{OCE}} = 0.1$ and $\tau_{\max} = 2$ in the forward step and with CMI nearest-neighbor parameter $k_{CMI} = 15$ and $B = 500$ permutation surrogates. FullCI was also estimated with CMI parameter $k_{CMI} = 15$ and $B = 500$ permutation surrogates. PCMCI was implemented with $\alpha_{\mathrm{PC}} = 0.1, \tau_{\max} = 2$ and CMIknn parameters $k_{CMI} = 15, k_{perm} = 5$ and $B = 500$ permutation surrogates (Runge, 2018). PCMCI was run without restricting the number of parents p_X in the MCI step, while for PCMCI_0 only the parents of the non-lagged variable were included. For each noise level σ, we ran the four methods on 200 time series realizations of the model. We compute as true positives the average rates at which the links $Z_{t-1} \to X_t$ and $Z_{t-1} \to Y_t$ were detected from the 200 realizations at an $\alpha = 0.05$ significance level. We calculate as false positives the average rates for $i \neq j \in \{X, Y, Z\}$ and $\tau \in \{1, 2\}$ where there is no link. Since we use an $\alpha =$

0.05 significance level, a well-calibrated test should yield 5% false positives. The edge color and width of the graphs in Fig. 14 corresponds to the lag with maximum CMI/CCM value.

2. Model for examples on autocorrelation and high-dimensionality

The model time series graph is depicted in Fig. 15. In the model setup, all links correspond to linear dependencies. The autocorrelation a is the same for all variables $X, Y, Z_1, \ldots, Z_{D_Z}$. The coupling coefficient c of the link $X_{t-1} \to Y_t$ is zero to test false positive rates and nonzero to test true positive rates. For nonzero c its value is chosen such that in the corresponding bivariate model without autocorrelation ($D_Z = 0, a = 0$), we obtain a constant mutual information $I(X_{t-1}; Y_t) = 0.03$ *nats* for the linear coupling with partial correlation. On the other hand, the common driver forcing coefficient $b = b(D_Z, a)$ and the covariance among the drivers $\sigma_Z = b/2$ are chosen such that in the full model for every pair ($D_Z > 0, a \geq 0$) with $c = 0$ we have $I(X_{t-1}; Y_t) = 0.4$ *nats*, a relatively strong forcing corresponding to a correlation of ≈ 0.7.

This setup guarantees that the unconditional dependence stays the same and we only investigate the effect of increasing autocorrelation and higher dimensionality. We test additive Gaussian noise terms $\eta \sim \mathcal{N}(0, 1)$. For $D_Z = 0$ the model setup corresponds to two autocorrelated processes without a common driver forcing. The sample lengths are $n = 150$ for partial correlation. 1000 realizations were evaluated to assess false and true positives.

3. Pre-whitening and block-shuffling

We also tested the OCE test using a pre-whitening (OCEpw) and a block-shuffle permutation test (OCEbs). For the pre-whitening test, we preprocessed all N time series by estimating the univariate lag-1 autocorrelation coefficients $\hat{a}_i = \rho(X^i_{t-1}; X^i_t)$ and regressing out the AR(1) autocorrelation part of the signals:

$$\tilde{X}^i_t = X^i_t - \hat{a}_i X^i_{t-1} \quad \forall t \text{ and } i = 1, \ldots, N. \tag{B1}$$

Then the OCE test is applied to these residuals $\tilde{\mathbf{X}}$.

Another remedy is a block-shuffle permutation test, which is based on a block-shuffle surrogate test following Peifer *et al.* (2005) and Mader *et al.* (2013). For the test statistic T, an ensemble of $M = 500$ values of $I(X^*_{t-\tau}; Y_t \mid \ldots)$ is generated where $X^*_{t-\tau}$ is a block-shuffled surrogate of $X_{t-\tau}$, i.e., with blocks of the original time series permuted. As an optimal block-length, we use the method described in Peifer *et al.* (2005) and Mader *et al.* (2013) for non-overlapping blocks. The optimal block-length formula Eq. (6) in Mader *et al.* (2013) involves the decay rate of the envelope of the autocorrelation function $\gamma(\tau)$. The latter was estimated up to a maximum delay of 5% of the samples, and the envelope was estimated using the Hilbert transform. Then a function $C\phi^\tau$ was fit to the envelope with constant C to obtain the decay rate ϕ. The block length was limited to a maximum of 10% of the sample length. Finally, the estimated values are sorted, and a p-value is obtained as the fraction of surrogates with values greater than or equal to the estimated value.

Arnhold, J., Grassberger, P., Lehnertz, K., and Elger, C., "A robust method for detecting interdependences: Application to intracranially recorded EEG," Physica D **134**, 419–430 (1999). arXiv:9907013 [chao-dyn].

Barnett, L., Barrett, A. B., and Seth, A. K., "Granger causality andtransfer entropy are equivalent for Gaussian variables," Phys. Rev. Lett. **103**, 238701 (2009). arXiv:0910.4514

Barnett, L. and Seth, A. K., "Granger causality for state space models," Phys. Rev. E **91**, 040101 (2015).

Benjamini, Y. and Hochberg, Y., "Controlling the false discovery rate: A practical and powerful approach to multiple testing," J. R. Stat. Soc. Ser. B **57**, 289–300 (1995).

Breitung, J. and Swanson, N. R., "Temporal aggregation and spurious instantaneous causality in multiple time series models," J. Time Ser. Anal. **23**, 651–665 (2002).

Chalupka, K., Perona, P., and Eberhardt, F., "Fast conditional independence test for vector variables with large sample sizes," (2018). arXiv:1804.02747v1

Colombo, D. and Maathuis, M. H., "Order-independent constraint-based causal structure learning," J. Mach. Learn. Res. **15**, 3921–3962 (2014).

Cover, T. M. and Thomas, J. A., *Elements of Information Theory* (John Wiley & Sons, Hoboken, 2006).

Daniusis, P., Janzing, D., Mooij, J., Zscheischler, J., Steudel, B., Zhang, K., and B. Schölkopf, "Inferring deterministic causal relations," in *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, edited by P. Grünwald and P. Spirtes (UAI, 2010), pp. 143–150. arXiv:1203.3475

Doran, G., Muandet, K., Zhang, K., and B. Schölkopf, "A permutation-based kernel conditional independence test," in *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, edited by N. L. Zhang and J. Tian (UAI, 2014), pp. 132–141.

Eichler, M., "Graphical modelling of multivariate time series," Probab. Theory Relat. Fields **153**, 233–268 (2012). arXiv:0610654 [math].

Eichler, M., "Causal inference with multiple time series: Principles and problems," Philos. Trans. R. Soc. A: Math. Phys. Eng. Sci. **371**, 20110613–20110613 (2013).

Entner, D. and Hoyer, P. O., "On causal discovery from time series data using FCI," in *Proceedings of the Fifth European Workshop on Probabilistic Graphical Models*, edited by P. Myllymaeki, T. Roos and T. Jaakkola (HIIT Publications, 2010), pp. 121–128.

Frenzel, S. and Pompe, B., "Partial mutual information for coupling analysis of multivariate time series," Phys. Rev. Lett. **99**, 204101 (2007).

Gao, S., Steeg, G. V., and Galstyan, A., "Efficient estimation of mutual information for strongly dependent variables," in *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics, San Diego, CA* (2015), Vol. 38, pp. 277–286.

Geiger, P., Zhang, K., Gong, M., Janzing, D., and B. Schölkopf, "Causal inference by identification of vector autoregressive processes with hidden components," 1917–1925 (2014). arXiv:1411.3972

Gong, M., Zhang, K., Schoelkopf, B., Tao, D., and Geiger, P., "Discovering temporal causal relations from subsampled data," in *Proceedings of the 32nd International Conference on Machine Learning* (Lille, France, 2015), Vol. 37, pp. 1898–1906.

Granger, C. W. J., "Investigating causal relations by econometric models and cross-spectral methods," Econometrica **37**, 424–438 (1969).

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., B. Schölkopf, and Smola, A. J., "A kernel statistical test of independence," in *Advances in Neural Information Processing Systems 20:21st Annual Conference on Neural Information Processing Systems 2007*, edited by J. C. Platt, D. Koller, Y. Singer and S. T. Roweis (2008), pp. 585–592.

Hirata, Y., J. M. Amigó, Matsuzaka, Y., Yokota, R., Mushiake, H., and Aihara, K., "Detecting causality by combined use of multiple methods: Climate and brain examples," PLoS ONE **11**, e0158572 (2016).

Hlavácková-Schindler, K., Palus, M., Vejmelka, M., and Bhattacharya, J., "Causality detection based on information-theoretic approaches in time series analysis," Phys. Rep. **441**, 1–46 (2007).

Hoerl, A. E., Kennard, R. W., and Ridge, Regression, "Biased estimation for nonorthogonal problems," Technometrics **12**, 55–67 (1970). arXiv:9809069v1 [arXiv:gr-qc].

Hyttinen, A., Eberhardt, F., and M. Järvisalo, "Constraint-based causal discovery: Conflict resolution with answer set programming," in *Proceedings*

of the 30th Conference on Uncertainty in Artificial Intelligence (UAI, 2014), pp. 340–349.
Hyttinen, A., Plis, S., M. Järvisalo, Eberhardt, F., and Danks, D., "Causal discovery from subsampled time series data by constraint optimization," J. Mach. Learn. Res. **52**, 216–227 (2016). arXiv:1602.07970
James, R. G., Barnett, N., and Crutchfield, J. P., "Information flows? A critique of transfer entropies," Phys. Rev. Lett. **116**, 238701 (2016). arXiv:1512.06479
Janzing, D., Mooij, J., Zhang, K., Lemeire, J., Zscheischler, J., Daniusis, P., Steudel, B., and B. Schölkopf, "Information-geometric approach to inferring causal directions," Artif. Intell. **182–183**, 1–31 (2012).
Kalisch, M., "Estimating high-dimensional directed acyclic graphs with the PC-algorithm," J. Mach. Learn. Res. **8**, 613–636 (2007).
Kantz, H. and Schreiber, T., *Nonlinear Time Series Analysis* (Cambridge University Press, Cambridge, 2003), pp. 27–43.
Kraskov, A., Stögbauer, H., and Grassberger, P., "Estimating mutual information," Phys. Rev. E **69**, 16 (2004). arXiv:0305641 [cond-mat].
Lauritzen, S. L., *Graphical Models* (Oxford University Press, Oxford, 1996).
Lopez-Paz, D., Muandet, K., B. Schölkopf, and Tolstikhin, I., "Towards a learning theory of cause-effect inference," in *Proceedings of the 32nd International Conference on Machine Learning* (Lille, France, 2015), Vol. 37. arXiv:1502.02398
Lord, W. M., Sun, J., and Bollt, E. M., "Geometric k-nearest neighbor estimation of entropy and mutual information," Chaos **28**, 033114 (2018). arXiv:1711.00748
Mader, M., Mader, W., Sommerlade, L., Timmer, J., and Schelter, B., "Block-bootstrapping for noisy data," J. Neurosci. Methods **219**, 285–91 (2013).
Maneewongvatana, S. and Mount, D., "It'sokay to be skinny, if your friends are fat," in *Center for Geometric Computing 4th Annual Workshop on Computational Geometry* (1999), pp. 1–8.
Pearl, J., *Causality: Models, Reasoning, and Inference* (Cambridge University Press, Cambridge, 2000).
Peifer, M., Schelter, B., Guschlbauer, B., Hellwig, B., C. H. Lücking, and Timmer, J., "On studentising and block length selection for the bootstrap on time series," Biom. J. **47**, 346–357 (2005).
Peters, J., P. Bühlmann, and Meinshausen, N., "Causal inference using invariant prediction: Identification and confidence intervals," J. R. Stat. Soc.: Ser. B (Stat. Method.) 1–42 (2016). arXiv:1501.01332
Peters, J., Janzing, D., and B. Schölkopf, "Causal inference on time series using restricted structural equation models," in *NIPS* (Curran Associates, Inc., 2013), pp. 154–162. arXiv:1207.5136
Peters, J., Janzing, D., and B. Schölkopf, *Elements of Causal Inference: Foundations and Learning Algorithms, Number December* (MIT Press, Cambridge, MA, 2017), pp. 1214–1216.
Philander, S., "ElNiño and La Niña," J. Atmos. Sci. **42**, 2652–2662 (1985).
Póczos, B. and Schneider, J., "Nonparametric estimation of conditional information and divergences," in *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics XX* (2012), Vol. 22, pp. 914–923.
Pompe, B. and Runge, J., "Momentary information transfer as a coupling measure of time series," Phys. Rev. E: Stat. Nonlin. Soft Matter Phys. **83**, 1–12 (2011).
Ramb, R., Eichler, M., Ing, A., Thiel, M., Weiller, C., Grebogi, C., Schwarzbauer, C., Timmer, J., and Schelter, B., "The impact of latent confounders in directed network analysis in neuroscience," Philos. Trans. R. Soc. A: Math. Phys. Eng. Sci. **371**, 20110612–20110612 (2013).
Rasmussen, C. and Williams, C., *Gaussian Processes for Machine Learning* (MIT Press, Cambridge, MA, USA, 2006).
Robins, J. M., Scheines, R., Spirtes, P., and Wasserman, L., "Uniform consistency in causal inference," Biometrika **90**, 491–515 (2003).
Runge, J., "Quantifying information transfer and mediation along causal pathways in complex systems," Phys. Rev. E **92**, 062829 (2015).
Runge, J., "Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information," in *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS)* (Lanzarote, Spain, 2018).
Runge, J., Donner, R. V., and Kurths, J., "Optimal model-free prediction from multivariate time series," Phys. Rev. E **91**, 052909 (2015a). arXiv:1506.05822v1
Runge, J., Heitzig, J., Marwan, N., and Kurths, J., "Quantifying causal coupling strength: Alag-specific measure for multivariate time series related to transfer entropy," Phys. Rev. E **86**, 061121 (2012a). arXiv:1210.2748
Runge, J., Heitzig, J., Petoukhov, V., and Kurths, J., "Escaping the curse of dimensionality in estimating multivariate transfer entropy," Phys. Rev. Lett. **108**, 258701 (2012b).
Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., and Sejdinovic, D., "Detecting causal associations in large nonlinear time series datasets," http://arxiv.org/abs/1702.07007v2 (2018).
Runge, J., Petoukhov, V., Donges, J. F., Hlinka, J., Jajcay, N., Vejmelka, M., Hartman, D., Marwan, N., Palus, M., and Kurths, J., "Identifying causal gateways and mediators in complex spatio-temporal systems," Nat. Commun. **6**, 8502 (2015b).
Runge, J., Petoukhov, V., and Kurths, J., "Quantifying the strength and delay of climatic interactions: The ambiguities of cross correlation and a Novel Measure Based on Graphical Models," J. Clim. **27**, 720–739 (2014).
Scheines, R. and Ramsey, J., "Measurement error and causal discovery," in *CEUR Workshop Proceedings* (NIH Public Access, 2016), Vol. 1792, pp. 1–7.
Schreiber, T., "Measuring information transfer," Phys. Rev. Lett. **85**, 461–464 (2000). arXiv:0001042v1 [nlin].
Sen, R., Suresh, A. T., Shanmugam, K., Dimakis, A. G., and Shakkottai, S., "Model-powered conditional independence test," in *Advances in Neural Information Processing Systems* (Curran Associates, Inc. 2017), Vol. 30, pp. 2955–2965. arXiv:1709.06138
Smirnov, D. A., "Spurious causalities with transfer entropy," Phys. Rev. E **87**, 1–12 (2013).
Spirtes, P. and Glymour, C., "An algorithm for fast recovery of sparse causal graphs," Soc. Sci. Comput. Rev. **9**, 62–72 (1991)
Spirtes, P., Glymour, C., and Scheines, R., *Causation, Prediction, and Search* (The MIT Press, Boston, 2000).
Spirtes, P. and Zhang, K., "Causal discovery and inference: Concepts and recent methodological advances," Appl. Inform. **3**, 3 (2016).
Staniek, M. and Lehnertz, K., "Symbolic transfer entropy," Phys. Rev. Lett. **100**, 1–4 (2008).
Storch, H. V. and Zwiers, F. W., *Journal of the American Statistical Association* (Cambridge University Press, Cambridge, 1999), Vol. 95.
Strobl, E. V. and Spirtes, P. L., "Estimating and controlling the false discovery rate for the PC algorithm using edge-specific p-values," (2016). arXiv:1607.03975v1
Strobl, E. V., Zhang, K., and Visweswaran, S., "Approximate kernel-based conditional independence tests for fast non-parametric causal discovery," http://arxiv.org/abs/1702.03877 (2017). arXiv:1702.03877
Sugihara, G., May, R., Ye, H., Hsieh, C.-h., Deyle, E., Fogarty, M., and Munch, S., "Detecting causality in complex ecosystems," Science **338**, 496–500 (2012).
Sun, J. and Bollt, E., "Causation entropy identifies indirect influences, dominance of neighbors and anticipatory couplings," Physica D **267**, 49–57 (2014).
Sun, J., Taylor, D., and Bollt, E. M., "Causal network inference by optimal causation entropy," SIAM J. Appl. Dyn. Syst. **14**, 73–106 (2015). arXiv:1401.7574
Székely, G. J., Rizzo, M. L., and Bakirov, N. K., "Measuring and testing dependence by correlation of distances," Ann. Stat. **35**, 2769–2794 (2007). arXiv:0803.4101
Tibshirani, R., "Regression shrinkage and selection via the lasso," J. R. Stat. Soc. Ser. B Stat. Methodol. **58**, 267–288 (1996). arXiv:11/73273 [13697412].
Tikhonov, A., "Regularization of incorrectly posed problems," Sov. Math. Doki. **4**, 1624–1627 (1963).
Tsamardinos, I. and Brown, L. E., "Bounding the false discovery rate in local Bayesian network learning," in *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence* (AAAI Press, 2008), pp. 1100–1105.
Tsamardinos, I., Brown, L. E., and Aliferis, C. F., "The max-min hill-climbing Bayesian network structure learning algorithm," Mach. Learn. **65**, 31–78 (2006).
Uhler, C., Raskutti, G., P. Bühlmann, and Yu, B., "Geometry of the faithfulness assumption in causal inference," Ann. Stat. **41**, 436–463 (2013).
VanderWeele, T., *Explanation in Causal Inference: Methods for Mediation and Interaction* (Oxford University Press, Oxford, 2015).
Vejmelka, M. and Palus, M., "Inferring the directionality of coupling with conditional mutual information," Phys. Rev. E **77**, 026214 (2008).
Wang, X., Pan, W., Hu, W., Tian, Y., and Zhang, H., "Conditional distance correlation," J. Am. Stat. Assoc. **110**, 1726–1734 (2015).

Wasserman, L., *All of Statistics : A Concise Course in Statistical Inference* (Springer Berlin Heidelberg, New York, 2004). arXiv:1011.1669v3

Watts, D. J. and Strogatz, S. H., "Collective dynamics of "small-world" networks," Nature **393**, 440–442 (1998).

Wibral, M., Pampu, N., Priesemann, V., F. Siebenhühner, Seiwert, H., Lindner, M., Lizier, J. T., and Vicente, R., "Measuring information-transfer delays," PloS ONE **8**, e55809 (2013).

Wiener, N., *Modern Mathematics for Engineers* (McGraw-Hill, New York, 1956).

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

Zhang, K. and A. Hyvärinen, "On the identifiability of the post-nonlinear causal model," in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence* (AUAI Press Arlington, VA, 2009), pp. 647–655. arXiv:1309.2178

Zhang, K., Peters, J., Janzing, D., and B. Schölkopf, "Kernel-based conditional independence test and application in causal discovery," *27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)* (AUAI Press Arlington, VA, 2011), pp. 804–813. arXiv:1202.3775