Necessary and Sufficient Conditions for Causal Feature Selection in Time Series with Latent Common Causes

Atalanti A. Mastakouri [1] **Bernhard, Schölkopf** [1 2] **Dominik, Janzing** [1]

Abstract

We study the identification of direct and indirect causes on time series with latent variables, and provide a constrained-based causal feature selection method, which we prove that is both sound and complete under some graph constraints. Our theory and estimation algorithm require only two conditional independence tests for each observed candidate time series to determine whether or not it is a cause of an observed target time series. Furthermore, our selection of the conditioning set is such that it improves signal to noise ratio. We apply our method on real data, and on a wide range of simulated experiments, which yield very low false positive and relatively low false negative rates.

1. Introduction

Causal feature selection in time series is a fundamental problem in several fields (i.e. biology, economics, climate research (Runge et al., 2019a)). Often the causes of a target time series need to be detected from a pool of candidate causes with latent confounders.

While Granger causality (Wiener, 1956; Granger, 1969; 1980) (see def. 3. in Appendix) has been the standard approach to causal analysis of time series since half a century, several issues caused by violations of its assumptions (causal sufficiency, no instantaneous effects) have been described in the literature (Peters et al., 2017). Several approaches addressing these problems have been proposed during the last decades (Hung et al., 2014; Guo et al., 2008). Nevertheless, causal inference in time series is still challenging without an efficient solution yet, despite the fact that the time order of variables provide information about the direction of some edges (Pearl, 2009; Spirtes et al., 1993). The discovery of the causal graph from data is largely based on the graphical criterion of d-separation formalizing the set of conditional independences (CI) to be expected, based on the causal Markov condition and causal faithfulness (Spirtes et al., 1993) (See definitions in App. Sec. 3).

Several authors showed how to derive d-separation based causal conclusions in time series beyond Granger's work. The majority of these works focuses on full graph discovery with conclusions up to Markov-equivalent classes. The remaining works focus on the problem of causal feature selection, which means the detection of direct and indirect causes of a given target time series. In the former group belong methods such as tsFCI (Entner & Hoyer, 2010) and SVAR-FCI (Malinsky & Spirtes, 2018), which are inspired by the FCI algorithm (Spirtes et al., 1993) and the work from (Eichler, 2007) and (Moneta et al., 2011) (see also (Runge, 2018; Runge et al., 2019a)). These methods do not assume causal sufficiency, and as such they need extensive CI testing. These methods are computationally intensive with exhaustive searching over all lags and conditioning sets. Another method of this first group is PCMCI ((Runge et al., 2019b)), which although it reaches lower rates of false positives compared to classical Granger causality, it still relies on the assumption of causal sufficiency. The most known method among those that focus on the causal feature selection (latter group) is seqICP (Pfister et al., 2019). However, seqICP requires sufficient interventions in the dataset, which should affect only the input and not the target. This requirement is hard to be met in problems where only observational data are available. Moreover, in the presence of hidden confounders, seqICP will detect only a subset of the ancestors of the target time series.

Here, we focus on the problem of causal feature selection in time series based on solely observational data, without assuming causal sufficiency. Under some connectivity assumptions, we construct conditions, which we prove to be sufficient for direct and indirect causes, and necessary for direct unconfounded causes, even in the presence of latent confounders. In contrast to other CI based methods, our method directly constructs the right conditioning set, without *searching* over a large set of possible combinations. It thus avoids statistical issues of multiple hypothesis testing. In contrast to seqICP, given our assumptions, we prove that

[1]Amazon Research Tuebingen, AWS Causality Group, Germany [2]Max Planck Institute for Intelligent Systems, Empirical Inference Department, Germany. Correspondence to: Atalanti A. Mastakouri <atalanti@amazon.de>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

our method will detect all the unconfounded direct causes of the target without requiring interventions in the dataset. We provide experimental results on simulated graphs of varying numbers of observed and hidden time series, density of edges, noise levels, and sample sizes. We show that our method leads to almost zero false positives and relatively low false negative rates, even in latent confounded environments, thus outperforming Granger causality among other methods. Finally, we achieve meaningful results even on experiments with real data where we cannot validate our graph assumptions. We call our method *SyPI* as it performs a **Sy**stematic **P**ath **I**solation for causal feature selection in time series.

2. Theory and Methods

We are given observations from a univariate target time series $Y := (Y_t)_{t\in\mathbb{Z}}$ whose causes we wish to find, and observations from a multivariate time series $\mathbf{X} := ((X_t^1,\dots,X_t^d))_{t\in\mathbb{Z}}$ of potential causes (candidates). Also, we allow an unobserved multivariate time series $\boldsymbol{U_t} := ((U_t^1,\dots,U_t^m))_{t\in\mathbb{Z}}$, which may act as common cause of the observed ones; as such, we do not assume *causal sufficiency*. We use $\boldsymbol{Q_t^i}, i,t \in \mathbb{Z}$ to refer to any node when we need not specify if it belongs to an observed or unobserved time series. [*] We introduce the following terminology to describe the causal relations among $\mathbf{X}, \mathbf{U}, Y$:

Terminology-Notation:

T1 *full time graph* is the infinite DAG having X_t^i, Y_t and U_t^j as nodes.

T2 *summary graph* is the directed graph with nodes $Q \in (X^1,...,X^d,U^1,...,U^d,Y)$ containing an arrow from Q^j to Q^k for $j\neq k$ whenever there is an arrow from Q_t^j to Q_s^k for $t\le s\in\mathbb{Z}$. (Peters et al., 2017)

T3 $Q_t^i \to Q_s^j$ for $t\le s\in\mathbb{Z}$ means a directed path that does not include any intermediate observed nodes in the full time graph (confounded or unconfounded).

T4 $Q_t^i \dashrightarrow Q_s^j$ for $t\le s\in\mathbb{Z}$ in the full time graph means a directed path from Q_t^i to Q_s^j.

T5 A *confounding path* between Q_t^i and Q_s^j in the full time graph is a path of the form $Q_t^i \dashleftarrow Q_{t'}^k \dashrightarrow Q_s^j$, $t'\le t,s\in\mathbb{Z}$ consisting of two directed paths and a common cause of Q_t^i and Q_s^j.

T6 A *confounded path* is an arbitrary path between two nodes Q_t^i, Q_s^j in the full-time graph that coexists with a confounding path between Q_t^i andQ_s^j.

T7 An *sg-unconfounded* (summary graph unconfounded) causal path is a causal path in the full time graph that does not appear as a confounded path in the summary graph .

T8 v is a *lag for the ordered pair of a time series X^i and the target Y* (X^i,Y) if there exists a collider-free path X_t^i- - -Y_{t+v} that does not contain a link of this form $Q_{t'}^r \to Q_{t'+1}^r$, with t' arbitrary, for any $r\not\equiv i,j$, nor any duplicate node, and any node in this path does not belong to X^i, Y. See explanatory Figure 1.

T9 We say that a set of time series $(\mathbf{X},Y)$ have *single-lag dependencies* if all the $X^i\in\mathbf{X}$ have only one lag v for each pair X^i, Y. Otherwise we refer to *multiple-lag dependencies*.

Figure 1 shows some example graphs and the lags between the candidate and the target time series, based on the definition T8. The integers defined by the highlighted green path between X^i and Y in graphs (a) and (b) are example lags for the single-lag (a) and multi-lag graph (b) accordingly, while the path in (c) does not define a lag because it contains a link $Q_{t+1}^r \to Q_{t+2}^r$. If the links between the time series were direct links, then the correct lag for (X^i,Y) in (c) would be 2.

We now assume that the graph satisfies the following assumptions.

Assumptions:

A1 **Causal Markov condition** in the full time graph.

A2 **Causal Faithfulness** in the full time graph.

A3 **No backward arrows** in time $X_{t'}^i \not\to X_t^j, \forall t'>t$

A4 **Stationary** full time graph: the full time graph is invariant under a joint time shift of all variables

A5 The full time graph is **acyclic**.

A6 The **target** time series Y is a sink node.

A7 There is an arrow $X_{t-1}^i \to X_t^i, Y_{t-1}\to Y_t \forall i,t\in\mathbb{Z}$. Note that arrows $U_{t-1}^i \to U_t^i$ need not exit, we then call U memoryless.

A8 There are no arrows $Q_{t-s}^i \to Q_t^i$ for $s>1$.

A9 Every variable U^i that affects Y **directly** (no intermediate observed nodes in the path in the summary graph) or that is connected with an observed collider in the summary graph should be memoryless ($U_{t-1}^i \not\to U_t^i$) and should have single-lag dependencies with Y in the full time graph.[†]

[*] Since there can only be one target time series Y, by overloading the notation, we use Q to refer to $\mathbf{X}$ or $\boldsymbol{U}$ when we already refer to target's nodes by Y (Fig. 1).

[†] Note that this assumption is only required for the completeness of the algorithm against direct false negatives (Theorem 2). The violation of this assumption does not spoil Theorem 1a/1b. The existence of a **latent variable with memory** affecting the

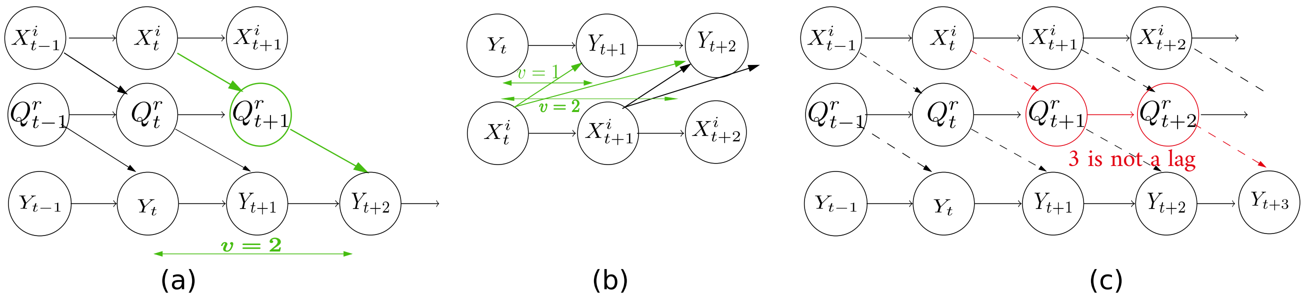

Figure 1. In (a) we have a single lag depedendency graph, and the integer 2 is the lag for (X^i, Y). (b) shows a multi-lag dependency graph where both integers 1 and 2 are lags for (X^i, Y). On the contrary, the red coloured path in (c) that corresponds to the integer 3 is not a lag, because it contains the link $Q^r_{t+1} \rightarrow Q^r_{t+2}$.

Note that the first five are usually standard assumptions of time series analysis and causal discovery, while assumptions A6 - A9 impose some restrictions on the connectivity of the graph. We further discuss about the assumptions in Section 5.

Below, we present three theorems for detection of causes in the full time graph. **Theorem 1a** provides **sufficient conditions for direct and indirect sg-unconfounded causes in single-lag dependency graphs**. **Theorem 1b provides sufficient conditions for direct and indirect causes in multi-lag dependency graphs. Theorem 2** provides **necessary conditions for identifying all the direct sg-unconfounded causes** of a target time series **in single-lag dependency graphs**, assuming the imposed graph constraints.

Intuition for proposed conditions in Theorems 1a/1b and 2: The idea is for each candidate time series X^j to *isolate* paths of the form $X^j_{t-1} \rightarrow X^j_t - - - Y_{t+w_j}, w_j \in \mathbb{Z}$, where no more than one observed node from each time series belong in '$- - -$', in the full time graph, and extract triplets $(X^j_{t-1}, X^j_t, Y_{t+w_j})$ as in (Mastakouri et al., 2019) (orange triplet Fig.2). This way we can exploit the fact that if there is a confounding path between X^j_t and Y_{t+w_j}, then X^j_t will be a collider that will unblock the path between X^j_{t-1} and Y_{t+w_j} when we condition on it. (Mastakouri et al., 2019) proposed sufficient conditions for causal feature selection in a DAG (no time-series) where a cause of a potential cause was known or could be assumed due to time-ordered pair of variables. Our goal here is to propose both necessary and sufficient conditions which will identify whether $X^j \dashrightarrow Y_{t+w_j}$ as in Fig. 2, or $X^j \dashleftarrow U \dashrightarrow Y_{t+w_j}$. To achieve that, we need for each candidate time series a conditioning set that will allow us to isolate the path of interest. As an example, Fig. 2 depicts with purple the nodes that will consist the conditioning set for the candidate X^j, as we propose in the Theorems below. Fig. 2 visualizes why time-series raise an additional challenge for identifying sg-unconfounded causal relations. While the influence of X^j on Y is unconfounded in the summary graph, the influence $X^j_t \rightarrow Y_{t+1} (\equiv Y_{t+w_j})$ is confounded in the full time graph due to its own past; for example X^j_t and Y_t are confounded by X^j_{t-1}.

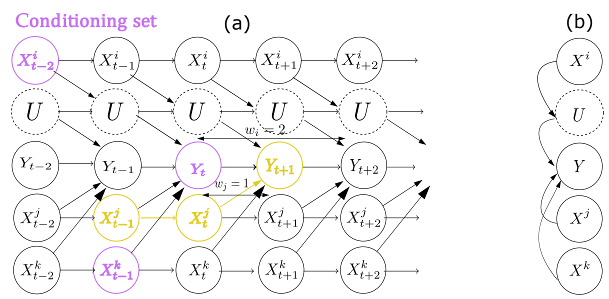

Figure 2. An example full-time graph (a) of 2 observed, 1 potentially hidden and 1 target time series. Identifying sg-unconfounded causal paths in time series is a challenge, as the past of each series introduces dependencies that are not visible in the summary graph (b).

Therefore we need to condition on $Y_t (\equiv Y_{t+w_j-1})$ to remove past dependencies. If no other time series were present, that would be sufficient. However, in the presence of other time series affecting the target Y, Y_{t+w_j-1} becomes a collider that unblocks dependencies. If, for example, we want to examine X^i as a candidate cause, we need first to condition on $Y_{t+w_i-1} \equiv Y_{t+1}$, which is the past of the Y_{t+w_i}. Following, we need to condition on one node from each time series $\mathbf{X} \setminus X^i$ that enter $Y_{t+w_i-1} \equiv Y_{t+1}$ (which is a collider) to avoid all the dependencies that might be created by conditioning on it. It is enough to condition only on these nodes for the following reason: If a node $X^{j \neq i}$ has a w_j lag-dependency with Y, then there is an (un)directed path from $X^j_{t+w_i-w_j-1}$ to Y_{t+w_i-1}. If this path is a confounding one, then conditioning on $X^j_{t+w_i-w_j-1}$ is not necessary, but also not harmful, because the future of this time se-

target time series Y **directly**, or of a **latent variable affecting directly the target with multiple lags** renders impossible the existence of a conditioning set that could d-separate the future of the target variable and the past of any other observed variable.

ries in the full graph is still independent of Y_{t+w_i}. This independence is forced by the fact that the $X^j_{t+w_i-w_j}$ is a collider because of the stationarity of graphs and this collider is by construction *not* in the conditioning set. If $X^j, j \neq i$ is connected with Y_{t+w_i-1} via a directed link (as in fig. 2), then conditioning on $X^j_{t+w_i-w_j-1}$ is necessary to block the parallel path created by its future values $X^j_{t+w_i-w_j-1} \rightarrow X^j_{t+w_i-w_j} \dashrightarrow Y_{t+v}$. Based on this idea of isolating the path of interest, we build the conditioning set as described in Theorem 1a/1b and its almost converse Theorem 2, where we prove the necessity and sufficiency of their conditions.

Theorem 1a. *[Sufficient conditions for a direct or indirect sg-unconfounded cause of Y in single-lag dependency graphs] Assuming A1-A5, A7 and A8 and single-lag dependency graphs, let w_i be the minimum lag (see T8) between X^i and Y. Further, let $w_{ij} := w_i - w_j$. Then, for every time series $X^i \in \boldsymbol{X}$ we define a conditioning set $\boldsymbol{S^i} = \{X^1_{t+w_{i1}-1}, X^2_{t+w_{i2}-1}, ..., X^{i-1}_{t+w_{i,i-1}-1}, X^{i+1}_{t+w_{i,i+1}-1}, ..., X^n_{t+w_{in}-1}\}$.*

If

$$X^i_t \not\perp\!\!\!\perp Y_{t+w_i} \mid \{\boldsymbol{S^i}, Y_{t+w_i-1}\} \tag{1}$$

and

$$X^i_{t-1} \perp\!\!\!\perp Y_{t+w_i} \mid \{\boldsymbol{S^i}, X^i_t, Y_{t+w_i-1}\} \tag{2}$$

are true, then

$$X^i_t \dashrightarrow Y_{t+w_i}$$

and the path between the two nodes is sg-unconfounded.

Proof. **(Proof by contradiction)**
We need to show that in single-lag dependency graphs, if $X^i_t \not\dashrightarrow Y_{t+w_i}$ or if the path $X^i_t \dashrightarrow Y_{t+w_i}$ is sg-confounded then at least either (1) or (2) is violated.

First assume that there is no directed path between X^i_t and Y_{t+w_i}: $X^i_t \not\dashrightarrow Y_{t+w_i}$. Then, there is a confounding path $X^i_t \dashleftarrow Q^j_{t'} \dashrightarrow Y_{t+w_i}, t' \leq t$ without any colliders. (Colliders cannot exist in the path by the definition of the lag T8.) In that case we will show that either condition 1 or 2 is violated. If all the existing confounding paths $X^i_t \dashleftarrow Q^j_{t'} \dashrightarrow Y_{t+w_i}, t' \leq t$ contain an observed confounder $Q^j_{t'} \equiv X^j_{t'} \in \{\boldsymbol{S^i}, Y_{t+w_i-1}\}$ (there can be only one confounder since in this case there are no colliders in the path), then condition 1 is violated, because we condition on $X^j_{t'}$ which d-separates X^i_t and Y_{t+w_i}. If in all the existing confounding paths the confounder node $Q^j_{t'} \notin \{\boldsymbol{S^i}, Y_{t+w_i-1}\}, t' \leq t$ but some observed non-collider node is in the path and this node belongs to $\{\boldsymbol{S^i}, Y_{t+w_i-1}\}$, then condition 1 is violated, because we condition on $\boldsymbol{S^i}$ which d-separates X^i_t and Y_{t+w_i}. If there is at least one confounding path and its confounder node does not belong in $\{\boldsymbol{S^i}, Y_{t+w_i-1}\}$ and no other observed (non-collider or descendant of collider) node which is in the path belongs in $\{\boldsymbol{S^i}, Y_{t+w_i-1}\}$ then condition 2 is violated for the following reasons: Let's name $p1 : X^i_t \dashleftarrow Q^j_{t'} \dashrightarrow Y_{t+w_i}, t' \leq t$. We know the existence of the path $p2 : X^i_{t-1} \rightarrow X^i_t$, due to A7.

(1I) If $p1$ and $p2$ have X^i_t in common, then X^i_t is a collider. Thus, adding X^i_t in the conditioning set would unblock the path between X^i_{t-1} and Y_{t+w_i}.

(1II) If $p1$ and $p2$ have X^i_{t-1} in common, that means X^i_{t-1} lies on $p1$. Thus X^i_t is not in the path from X^i_{t-1} to Y_{t+w_i} and hence adding X^i_t to the conditioning set could not d-separate X^i_{t-1} and Y_{t+w_i}.

In both cases condition 2 is violated.
Now, assume that there is a directed path $X^i_t \dashrightarrow Y_{t+w_i}$ but it is "sg-confounded" (there exist also a parallel confounding path $p3 : X^i_t \dashleftarrow Q^j_{t'} \dashrightarrow Y_{t+w_i}, t' \leq t$. Then, if $p3$ and $p2$ have X^i_t in common, then condition 2 is violated due to (1I). If $p3$ and $p2$ have X^i_{t-1} in common, then condition 2 is violated due to (1II). In all the above cases we show that if conditions 1 and 2 hold true in single-lag dependency graphs, then X^i_t is an "sg-unconfounded" direct or indirect cause of Y_{t+w_i}. □

Theorem 1b. *[Sufficient conditions for a (possibly confounded) direct or indirect cause of Y in multi-lag dependency graphs] Assuming A1-A5, A7 and A8, and allowing multi-lag dependency graphs, let w_i be the minimum lag (see T8) between X^i and Y. Further, let $w_{ij} := w_i - w_j$. Then, for every time series $X^i \in \boldsymbol{X}$ we define a conditioning set $\boldsymbol{S^i} = \{X^1_{t+w_{i1}-1}, X^2_{t+w_{i2}-1}, ..., X^{i-1}_{t+w_{i,i-1}-1}, X^{i+1}_{t+w_{i,i+1}-1}, ..., X^n_{t+w_{in}-1}\}$.*

If conditions 1 and 2 of Theorem 1a hold true for the pair X^i_t, Y_{t+w_i}, then

$$X^i_t \dashrightarrow Y_{t+w_i}$$

We can think of $\boldsymbol{S^i}$ as the set that contains only one node from each time series X^j and this node is the one that enters the node Y_{t+w_i-1} due to a directed or confounded path (if w_j exists then the node is the one at $t + w_{ij} - 1$).

Proof of Theorem 1b is provided in Sec. 6.2 of the Appendix, following similar logic with the proof of Theorem 1a.

Remark 1. *Theorem 1b conditions hold for any lag as defined in T8; not only for the minimum lag.*

Theorem 2. *[Necessary conditions for a direct sg-unconfounded cause of Y in single-lag graphs]*

Let the assumptions and the definitions of Theorem 1a hold, in addition to Assumptions A6 and A9.

If X^i_t is a direct, "sg-unconfounded" cause of Y_{t+w_i} ($X^i_t \rightarrow Y_{t+w_i}$), then cond. 1 and 2 of Theorem 1a hold.

Proof. **(Proof by contradiction)**
Assume that the direct path $X^i_t \rightarrow Y_{t+w_i}$ exists and it is unconfounded. Then, condition 1 is true. Now assume that condition 2 does not hold. This would mean that the set $\{\boldsymbol{S^i}, X^i_t, Y_{t+w_i-1}\}$ does not d-separate X^i_{t-1} and Y_{t+w_i}. Note that a path p is said to be *d-separated* by a set of nodes in Z if and only if p contains a chain or a fork such that the middle node is in Z, or if p contains a collider such that neither the middle node nor any of its descendants are in the Z. Hence, a violation of condition 2 would imply that (a) there is some middle node of a collider or descendant of a collider in $\{\boldsymbol{S^i}, X^i_t, Y_{t+w_i-1}\}$ and no non-collider node in this path belongs to this set, or (b) that there is a collider-free path between X^i_{t-1} and Y_{t+w_i} that does not contain any node in $\{\boldsymbol{S^i}, X^i_t, Y_{t+w_i-1}\}$.

(a) *There is some middle node of a collider or descendant of a collider in $\{\boldsymbol{S^i}, X^i_t, Y_{t+w_i-1}\}$ and no non-collider node in this path belongs to this set:*
(a1:) If there is at least one path $p1 : X^i_{t-1} \dashrightarrow Y_{t+w_i-1} \dashleftarrow Y_{t+w_i}$ where Y_{t+w_i-1} is a middle node of a collider and none of the non-collider nodes in the path belong to $\{\boldsymbol{S^i}, X^i_t\}$: Such a path could be formed only if in addition to X^i some $Q^j_{t'}$ directly caused Y. Then $p1 : X_{t-1} \dashrightarrow Y_{t+w_i-1} \dashleftarrow Q^j_{t'} \rightarrow Y_{t+w_i}, t' \leq t + w_i$. (Due to our assumption for single-lag dependencies (see T9) a path of the form $X_{t-1} \dashrightarrow Y_{t+w_i-1} \dashleftarrow X^i_s - - - Y_{t+w_i}$ could not exist). Then, due to stationarity of graphs the node $Q^j_{t'-1}$ will enter Y_{t+w_i-1}. If this $Q^j_{t'}$ is hidden ($Q^j_{t'} \equiv U^j_{t'}$), then due to A9 this time series will be memoryless ($U^j_{t'-1} \not\rightarrow U^j_{t'}$). Therefore, the collider Y_{t+w_i-1} in the conditioning set will not unblock any path between X^i_{t-1} and Y_{t+w_i} that could contain $U^j_s, s > t'$. If $Q^j_{t'}$ is observed ($Q^j_{t'} \equiv X^j, j \neq i$) then due to A7 the path $p1$ will be $X^i_{t-1} \dashrightarrow Y_{t+w_i-1} \dashleftarrow X^j_{t+w_{ij}-1} \rightarrow X^j_{t+w_{ij}} \rightarrow Y_{t+w_i}$. However, this path is always blocked by $X^j_{t+w_{ij}-1} \in \boldsymbol{S^i}$ due to the rule we use to construct $\boldsymbol{S^i}$. That means a non-collider node in the conditioning set will necessarily be in the path $p1$, which contradicts the original statement.

(a2:) If there is at least one path $p2 : X^i_{t-1} \dashrightarrow X^i_t \dashleftarrow Y_{t+w_i}$ where X^i_t is a middle node of a collider and none of the non-collider nodes in the path belongs to $\{\boldsymbol{S^i}, Y_{t+w_i-1}\}$: This could only mean that there is a confounder between the target Y_{t+w_i} and X^i_t. However this contradicts that $X^i_t \rightarrow Y_{t+w_i}$ is "sg-unconfounded".

(a3:) If there is at least one path $p3 : X^i_{t-1} \dashrightarrow X^j_{t'} \dashleftarrow Y_{t+w_i}$ where $X^j_{t'} \in \boldsymbol{S^i}$ with $t' \leq t + w_i - 1$ is a middle node of a collider and no non-collider node in the path belongs to $\{\boldsymbol{S^i} \setminus X^j_{t'}, X^i_t, Y_{t+w_i-1}\}$: In this case, $t' \equiv t + w_{ij} - 1$ because $X^j_{t'} \in \boldsymbol{S^i}$. By construction of $\boldsymbol{S^i}$ all the observed nodes in $\mathbf{X} \setminus X^i$ that enter the node Y_{t+w_i-1} belong in $\boldsymbol{S^i}$. That means that $X^j_{t'}$ enters the node Y_{t+w_i-1}. Hence, in the path $p3$ Y_{t+w_i-1} will necessarily be a non-collider node which belongs to the conditioning set. This contradicts the original statement "and no non-collider node in the path belongs to $\{\boldsymbol{S^i} \setminus X^j_{t'}, X^i_t, Y_{t+w_i-1}\}$".

(a4:) If a descendant D of a collider G in the path $p4 : X^i_{t-1} \dashrightarrow G \dashleftarrow C \dashrightarrow Y_{t+w_i}$ belongs to the conditioning set $\{\boldsymbol{S^i}, X^i_t, Y_{t+w_i-1}\}$ and no non-collider node in the path belongs to it: Due to the single-lag dependencies assumption, $w_C \equiv w_i$ otherwise there are multiple-lag effects from C to Y. That means that, independent of C being hidden or not, the C in the collider path will enter the node Y_{t+w_i-1}. If $C \in \mathbf{X}$ then because C enters the node Y_{t+w_i-1}, $C \in \{\boldsymbol{S^i}, X^i_t, Y_{t+w_i-1}\}$. In the first case Y_{t+w_i-1} only and in the latter case also C are a non-collider variable in the path $p4$ that belongs to the conditioning set, which contradicts the statement of (a4). If the collider $G \in \mathbf{X}$, as explained in (a3) at least one non-collider variable in the path will belong in the conditioning set, which contradicts the statement (a4). Finally, if G and C are hidden, if $w_D \equiv w_C$ then the node Y_{t+w_i-1} is necessarily in the path as a pass-through node, which contradicts the statement (a4). If $w_D \not\equiv w_C$ then the single-lag assumption is violated.

(b) *There is a collider-free path between X^i_{t-1} and Y_{t+w_i} that does not contain any node in $\{\boldsymbol{S^i}, X^i_t, Y_{t+w_i-1}\}$*: Such a path would imply the existence of a hidden confounder between X^i_{t-1} and Y_{t+w_i} or the existence of a direct edge from X_{t-1} to Y_{t+w_i}. The former cannot exist because we know that X_t is an sg-unconfounded direct cause of Y_{t+w_i}. The latter would imply that there are multiple lags of direct dependency between X_t and Y_{t+w_i} which contradicts the assumption of single-lag dependencies.

Thus, whenever $X^i_t \rightarrow Y_{t+w_i}$ is an sg-unconfounded causal path, conditions 1 and 2 are necessary. □

Since it is unclear how to identify the lag in T8, we introduce the following lemmas for the detection of the minimum lag that we require in the theorems. We provide the proofs of the lemmas in Appendix Sec. 2.

Lemma 1. *If the paths between X^j and Y are directed then the minimum lag w_j as defined in T8 coincides with the minimum non-negative integer w'_j for which $X^j_t \not\perp\!\!\!\perp Y_{t+w'_j} \mid X^j_{past(t)}$. The only case where $w'_j \not\equiv w_j$ is when there is a confounding path between X^j and Y that contains a node from a third time series with memory. In this case $w'_j = 0$.*

Lemma 2. *Theorems 1a/1b and 2 are valid if the minimum lag w_j as defined in T8 is replaced with w'_j from lemma 1.*

Algorithm 1 *SyPI* Algorithm for Theorems 1a/1b, 2.

input $\mathbf{X}, Y$
output $causes\_of\_R$
  n_{vars} = shape$(\mathbf{X}, 1)$, causes_of_R= []
  $w = min\_lags(\mathbf{X}, Y)$
  for $i = 1$ **to** n_{vars} **do**
    $\mathbf{S_i} = \bigcup_{j=1, j\neq i}^{n_{\text{vars}}} \{X^j_{t+w[i]-w[j]-1}\}$
    pvalue1 $= cond\_ind\_test(X^i_t, Y_{t+w[i]}, [\mathbf{S_i}, Y_{t+w[i]-1}])$
    if pvalue1 $<$ threshold1 **then**
      pvalue2
      $= cond\_ind\_test(X^i_{t-1}, Y_{t+w[i]}, [\mathbf{S_i}, X^i_t, Y_{t+w[i]-1}])$
      if pvalue2 $>$ threshold2 **then**
        causes_of_R $=$ [causes_of_R, X^i_t]
      end if
    end if
  end for

Using Lemma 1 via lasso regression and the two conditions in Theorems 1a and 2 we build an algorithm to identify direct and indirect causes on time series. The input is a 2D array $\mathbf{X}$ (candidate time series) and a vector Y (target), and the output a set with indices of the time series that were identified as causes. The complexity of our algorithm is $\mathcal{O}(n)$ for n candidate time series, assuming constant execution time for the conditional independence test.

3. Experiments

3.1. Simulated experiments

To test our method, we build simulated full-time graphs, respecting the aforementioned assumptions. We sampled 100 random graphs for the following hyperparameters and their tested values: # samples $\in \{500, 1000, 2000, 3000\}$, # hidden variables $\in [0, 1, 2]$, # observed variables $\in [1, 2, 3, 4, 5, 6, 7, 8]$, Bernoulli($p$) existence of edge among candidate time series $\in \{0.1, 0.15, 0.2, 0.25\}$, Bernoulli($p$) existence of edge between candidate time series and target series $\in \{0.1, 0.2, 0.3\}$, and noise variance $\in \{10\%, 20\%, 30\%\}$. Although 10 time series (including hidden and target) are considered already many in causal discovery for statistical reasons ((Entner & Hoyer, 2010) ran up to 9, and (Moneta et al., 2011) up to 8 series), for proof of concept we also examined a combination of 20, and 30 time series with 5 hidden. We then calculate the false positive (FPR) and false negative rates (FNR) for the 100 random graphs. When constructing the time series, every time step is calculated as the weighted sum of the previous step of all the incoming time series, including the previous step of the current time series. The weights of the adjacent matrix between the time series are uniformly selected in the range $[0.7, 0.95]$ if they were not set to zero (we thus prevent too deterministic relationships or too weak edges, which would result in almost non-faithful distributions) ‡.

The two CI tests are calculated with partial correlation, since our simulations are linear, but there is no restriction for non-linear systems (see extension in Sec.5). For the "lag" calculation step of SyPI, we use lasso in a bivariate form between each node in $\mathbf{X}$ in the summary graph and Y (for non-linear relationships this step can be replaced with a non-linear regressor). We fixed the lasso parameters ($\lambda = 0.001$ and cut-off threshold for the coefficients $= 0.1$) once before running the experiments, without re-adjusting them for the different types of graphs. While our method is sound for both single and multi-lag dependency graphs, it is complete only for the former type. Thus, we simulated the time series with single-lags for the main core of the experiments. For completeness, we tested the performance of SyPI even with multiple lags, which we present in App. Sec. 6.5.4. Moreover, we compared our method to Lasso-Granger (Arnold et al., 2007) for 2 hidden and 3, 4 and 5 observed time series. SyPI operates with two thresholds for the p values of the two tests, *threshold1* for rejecting independence in condition 1, and *threshold2* for accepting independence in condition 2. Lasso-Granger (Arnold et al., 2007) operates with one hyper-parameter: the regularizer λ. To ensure a fair comparison, we tuned the λ for Lasso-Granger (not SyPI) such as to allow it at least the same FNR as SyPI, for same type of graphs. We did not do the comparison based on matching FPR, because Lasso-Granger generates many FPs in the presence of hidden confounders. For all the experiments, we used *threshold1*$= 0.01$ and *threshold2*$= 0.2$ for SyPI. In addition, we produced ROC curves for the two methods (see App. Sec. 6.6).

Furthermore, we compared SyPI against seqICP (Pfister et al., 2019) and PCMCI (Runge et al., 2019b). We simulated 10 different combinations (2 to 6 observed and 1 to 2 hidden series) Finally, we ran 100 simulations for 5 observed, 2 hidden and 1 target time series (only one combination due to the very long computation time of tsFCI), sample size 2000, medium density and noise to compare SyPI against the tsFCI (Entner & Hoyer, 2010). For a fair comparison we used the same thresholds for all the statistical tests of these methods (*threshold1* = *threshold2* = $\alpha = 0.05$).

3.2. Experiments on real-data

We also examined the performance of SyPI on real data, where we have no guarantee that our assumptions hold true. We use the official recorded prices of dairy products in Eu-

‡For completeness, weights in the range $[0.2, 0.95]$ were also tested leading to some increase in FPR, as expected due to faithfulness violation.

rope (EU) (data provided, App. Sec. 6.4). The target of our analysis was the variable 'Butter'. According to the production pipeline described in (Soliman & Mashhour, 2011), the first material for 'Butter' is 'Raw Milk', and 'Butter' is not used as ingredient for the other dairy products in the list (sink node assumption). Therefore, we can hypothesize that the direct cause of 'Butter' prices is the 'Raw Milk', and that the rest (other cheese, 'WMP', 'SMP', 'Whey Powder') are not causing 'Butter'. We examine three countries, two of which provide data for 'Raw Milk' (Germany 'DE' (8 time series) and Ireland 'IE' (6 time series)), and one where these values are not provided (United Kingdom 'UK' (4 time series)). This last dataset was on purpose selected as this would be a good realistic scenario of a hidden confounder. In that case our method must not identify any cause. As we have extremely low sample sizes (<180) identifying dependencies is particularly hard. For that reason we set 0 threshold on our lag detector and the *threshold1* at 0.05 for accepting dependence in condition 1.

4. Results

4.1. Simulated graphs

We tested SyPI for varying edge-density, noise levels, sample sizes, and number of observed and hidden time series. Figures 7a-9h in App. Sec. 6.5.1 depict the FPR and FNR for all these combinations. Overall, SyPI yielded FPR below 1% for sample size > 500, independent of noise level, density, or size of the graphs. FNR for the direct causes (indicated with red) ranges between 12% for small and sparse graphs and 45% for very large and dense graphs. Fig. 3 shows the behaviour of our algorithm in moderately dense graphs, for 2000 sample size, 20% noise variance and varying number of hidden series. We see that the FPR is close to zero, independent of the number of hidden variables. Although the total FNR increases with the number of series, the FNR that corresponds to *direct* causes (dashed lines), remains below 40%. We focus on the missed *direct* causes because SyPI is complete only for the direct ones (see Th. 2). Edge-density does not seem to affect the rates as shown in App. Sec. 6.5.2. Finally, as explained on Sec. 3.1 we tested the behaviour of SyPI on much larger graphs, with 20 and 30 time series (5 hidden). The FPRwas $0.8\% \pm 2.9$ for the 20 series, and $0.8\% \pm 2.3$ for the 30. The FNR for the direct causes was accordingly $22.7\% \pm 20.3$ and $15.6\% \pm 18.7$.

4.2. Comparison against other methods

First, we compare our algorithm against the widely used Lasso-Granger method. Fig. 4 shows that even in such confounded graphs (2 hidden time series) SyPI yields almost zero FPR, for similar or even lower total FNR than Lasso-Granger, which yields up to 16% FPR. Moreover, Fig. 10 in the Appendix shows that the ROC curve of SyPI is above the ROC curve of Lasso-Granger for all operating points, indicating that SyPI outperforms the latter, as expected due to its robustness to hidden confounders. Figure 5 shows the comparison of SyPI with PCMCI and seqICP. As we can see, SyPI has the lowest FPR ($< 1.5\%$) compared to PCMCI and seqICP for all type of tested graphs, and lower both direct ($20 - 40\%$, dashed lines) and total (solid lines) FNR than seqICP, which yielded up to 12% FPR and around 95% FNR. This is not surprising, as with hidden confounders seqICP will detect only a subset of the ancestors AN(Y). PCMCI yielded up to 25% FPR and around 25% FNR. Finally, we compared our method against tsFCI (Entner & Hoyer, 2010) for one combination due to the very long execution time tsFCI required (5 observed, 2 hidden, 1 target) over 100 random graphs. SyPI yielded significantly smaller average FPR than tsFCI with comparable variance ($2.8\% \pm 8.5$) than tsFCI ($15.4\% \pm 22.9$), yet almost twice as large FNR ($25.1\% \pm 29.5$) than tsFCI ($13.3\% \pm 24.1$).

4.3. Experiments on real data

We applied SyPI on the dairy-product prices for 'DE', 'IE' and 'UK'. SyPI successfully identified 'Raw Milk' as the direct cause of 'Butter' in the 'IE' dataset, correctly rejecting the remaining 4 nodes (100% TPR, 100% TNR). In 'DE', 'Raw Milk' was correctly identified with only one false positive ('Edam'); the remaining 6 nodes were rejected yielding 100% TPR and 84% TNR. Most importantly, in the 'UK' dataset where no measurements for 'Raw Milk' were provided (hidden confounder), SyPI correctly did not identify any cause (100% TNR). Finally, in (Mastakouri & Schölkopf, 2020) the SyPI method which we present here, was applied on Covid-19 infections cases yielding meaningful results on the causal tracking of the pandemic in Germany, on large graphs with noisy, confounded data.

5. Discussion

5.1. Efficient conditioning set

In contrast to other approaches, and due to the narrower goal of our method, SyPI does not search over a large set of possible combinations to identify the right conditioning sets. Instead, for each potential cause X^i, it directly constructs its 'separating set' for the nodes X^i_{t-1} and Y_{t+w_i} (cond. 2), from a pre-processing step that identifies the nodes that enter Y_{t+w_i-1} ($\mathbf{S^i}$). The resulting set $\{\mathbf{S^i}, Y_{t+w_i-1}, X^i_t\}$ contains therefore covariates that enter the outcome node Y_{t+w_i}, and not the potential cause X^i_{t-1}. Adjustment sets that include parents of the potential cause node are considered inefficient in terms of asymptotic variance of the causal effect estimate, as they can reduce the variance of the *cause* if they are strongly correlated with it, and thus reduce the signal (Henckel et al., 2019).

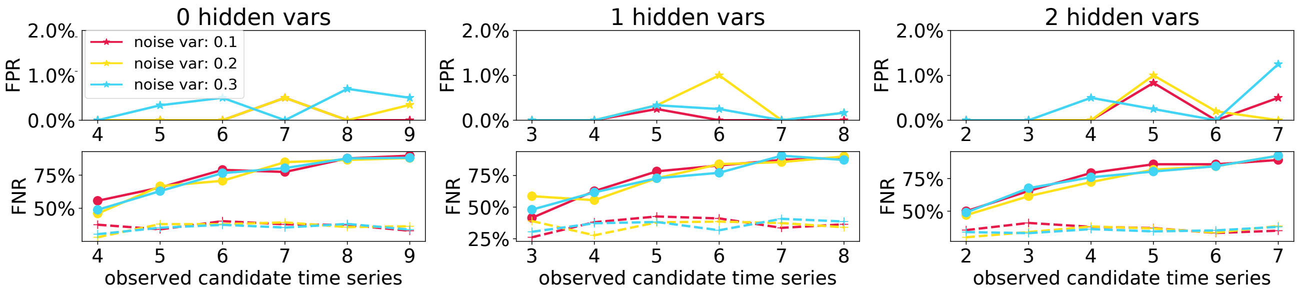

Figure 3. FPR and FNR for varying number of hidden (columns) and observed series (x-axis), noise variance and sample size 2000, for medium density. FPR is very low ($< 1.2\%$) for any number of hidden series. Although the total FNR increases with the graph size, the FNR for the direct causes (dashed lines), for which our method is complete, remains $< 40\%$.

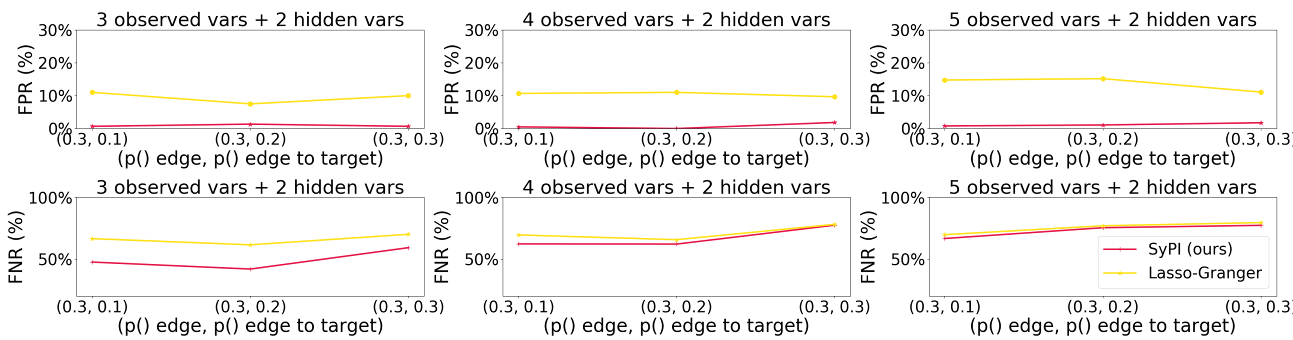

Figure 4. SyPI vs Lasso-Granger, for sample size 2000, 2 hidden series, 20% noise variance, for varying number of observed time series (columns) and edges density (x-axis). As we see, *SyPI* performs with significantly lower FPR ($< 1\%$) than Lasso-Granger, for similar or even lower FNR (direct + indirect). In contrast, Lasso-Granger reaches up to 16% FPR. Not tuning λ led to even larger FPR for Lasso-Granger.

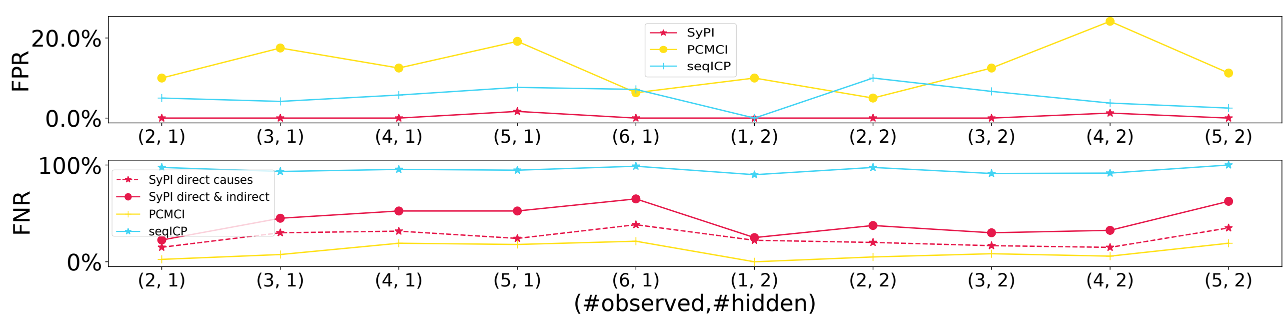

Figure 5. Comparison of SyPI against seqICP and PCMCI, for ten types (# observed, # hidden time series) of graphs. FPR and FNR are reported over 20 random graphs of each type. Our method SyPI has the lowest FPR ($< 1.5\%$) and direct-FNR $20 - 40\%$ (dash line). SeqICP yielded 12% FPR and 95% FNR. This is not surprising, as with hidden confounders seqICP will detect only a subset of $AN(Y)$. PCMCI yielded 25% FPR and 25% FNR for $a = 0.05$.

Instead, adding nodes that explain variance in the *outcome* node -as we do here- can contribute to a better SNR for the dependences under consideration.

5.2. Non-linear systems & Multiple-lags

SyPI can be used for both linear and non-linear relations among the time series. For the linear case, a partial correlation test is sufficient to examine the conditional dependencies, while in the non-linear case KCI (Zhang et al., 2012), KCIPT (Doran et al., 2014) or FCIT (Chalupka et al., 2018) could be used. Although SyPI is robust against FPs in "multiple-lags" graphs (see App. Fig. 11), Theorem 2 conditions are necessary only for "single-lags" (see T9). We could allow for "multiple-lags" if we were willing to condition on larger sets of nodes, which would significantly affect the statistical outcome. Right now, we require *at most* one node from each observed time series for the conditioning set. In a naive approach, n coexisting lags would require n nodes from each series to be added in the conditioning set. We further discuss future directions on multiple-lags in App. Sec. 6.6

5.3. Graph assumptions

Assumptions A1-A4 are often made in most constrained-based methods on time series. In addition, A7, A8 assure that X are time series with dependency from their previous time step. Therefore, although our assumptions seem many, they boil down to the graphical constraints that are required to avoid the problem that auto-lag hidden confounders create by inducing infinite-lag associations. This is a well known issue in which also (Malinsky & Spirtes, 2018) don't find causal relationships as stated in there. The graph simplification we impose by A9 aims to avoid this problem. This is a trade-off that we do not consider extreme, given the hardness of the problem of hidden confounding and the very few CI tests that we require. Finally the assumption A6 was added in order to be able to handle instantaneous effects with only the two tests that we require. This assumption could be replaced by a lighter one if we assumed that Y has no descendants that belong in its set of candidate causes (as shown in (Mastakouri & Schölkopf, 2020)). An alternative future step could be to expand the method to check both directions between target and candidates (candidate feature $\rightarrow$ target and candidate feature $\leftarrow$ target). This would of course result in twice as many tests, but it could replace A6.

5.4. Conclusion

We presented a causal feature selection method for time series that is build on only two CI-based conditions, which we proved that are necessary and sufficient for a time series to causally influence a target one, even in the possible presence of latent confounders, subject to some connectivity assumptions that seemed hard to avoid. The proposed algorithm scales linearly with the number of time series and requires a well defined conditioning set that contributes to the SNR. Our experiments on real data yielded meaningful results and on simulations particularly low FPR.

References

Arnold, A., Liu, Y., and Abe, N. Temporal causal modeling with Graphical Granger Methods. pp. 66–75, 2007.

Chalupka, K., Perona, P., and Eberhardt, F. Fast conditional independence test for vector variables with large sample sizes. ArXiv, abs/1804.02747, 2018.

Doran, G., Muandet, K., Zhang, K., and Schölkopf, B. A permutation-based kernel conditional independence test. In Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence, pp. 132–141, 2014.

Eichler, M. Causal inference from time series: What can be learned from Granger causality. In Proceedings of the 13th International Congress of Logic, Methodology and Philosophy of Science, pp. 1–12. King's College Publications London, 2007.

Entner, D. and Hoyer, P. O. On causal discovery from time series data using FCI. Probabilistic graphical models, pp. 121–128, 2010.

EU. European union prices of dairy products. https://ec.europa.eu/info/food-farming-fisheries/farming/facts-and-figures/markets/prices/price-monitoring-sector/.

Granger, C. W. J. Investigating causal relations by econometric models and crossspectral methods. Econometrica, 37:424–438, 1969.

Granger, C. W. J. Testing for causality, a personal viewpoint., volume 2. 1980.

Guo, S., Seth, A. K., Kendrick, K. M., Zhou, C., and Feng, J. Partial granger causality-Eliminating exogenous inputs and latent variables. Journal of Neuroscience Methods, 172(1):79 – 93, 2008.

Henckel, L., Perković, E., and Maathuis, M. H. Graphical criteria for efficient total effect estimation via adjustment in causal linear models. arXiv, 2019.

Hung, Y.-C., Tseng, N.-F., and Balakrishnan, N. Trimmed granger causality between two groups of time series. Electron. J. Statist., 8(2):1940–1972, 2014.

Malinsky, D. and Spirtes, P. Causal structure learning from multivariate time series in settings with unmeasured confounding. In Proceedings of 2018 ACM SIGKDD Workshop on Causal Disocvery, volume 92 of Proceedings of Machine Learning Research, pp. 23–47, 2018.

Mastakouri, A., Schölkopf, B., and Janzing, D. Selecting causal brain features with a single conditional independence test per feature. In Advances in Neural Information Processing Systems 32, 2019.

Mastakouri, A. A. and Schölkopf, B. Causal analysis of covid-19 spread in germany. arXiv preprint arXiv:2007.11896, 2020.

Moneta, A., Chlaß, N., Entner, D., and Hoyer, P. Causal search in structural vector autoregressive models. In NIPS Mini-Symposium on Causality in Time Series, pp. 95–114. PMLR, 2011.

Pearl, J. Causality. Cambridge University Press, 2nd edition, 2009.

Peters, J., Janzing, D., and Schölkopf, B. Elements of Causal Inference - Foundations and Learning Algorithms. Adaptive Computation and Machine Learning Series. The MIT Press, Cambridge, MA, USA, 2017.

Pfister, N., Bühlmann, P., and Peters, J. Invariant causal prediction for sequential data. Journal of the American Statistical Association, 114(527):1264–1276, 2019.

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

Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., Glymour, C., Kretschmer, M., Mahecha, M. D., Muñoz-Marí, J., et al. Inferring causation from time series in earth system sciences. Nature communications, 10(1):1–13, 2019a.

Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., and Sejdinovic, D. Detecting and quantifying causal associations in large nonlinear time series datasets. Science Advances, 5(11):eaau4996, 2019b.

Soliman, I. and Mashhour, A. Dairy marketing system performance in egypt. 01 2011.

Spirtes, P., Glymour, C., and Scheines, R. Causation, Prediction, and Search. 1993.

Wiener, N. The theory of prediction, Modern mathematics for the engineer, volume 8. 1956.

Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. Kernel-based conditional independence test and application in causal discovery. UAI, 2012.