A Mixed Noise and Constraint-based Approach to Causal Inference in Time Series

Karim Assaad[1,2], Emilie Devijver[1], Eric Gaussier[1], and Ali Ait-Bachir[2]

[1] Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG
{karim.assaad, emilie.devijver}@univ-grenoble-alpes.fr
eric.gaussier@imag.fr
[2] Coservit
ali.ait.bachir@coservit

Abstract. We address, in the context of time series, the problem of learning a summary causal graph from observations through a model with independent and additive noise. The main algorithm we propose is a hybrid method that combines the well-known constraint-based framework for causal graph discovery and the noise-based framework that gained much attention in recent years. Our method is divided into two steps. First, it uses a noise-based procedure to find the potential causes of each time series. Then, it uses a constraint-based approach to prune all unnecessary causes. A major contribution of this study is to extend the standard causation entropy measure to time series to handle lags bigger than one time step, and to rely on a lighter version of the faithfulness hypothesis, namely the *adjacency faithfulness*. Experiments conducted on both simulated and real-world time series show that our approach is fast and robust wrt to different causal structures and yields good results over all datasets, whereas previously proposed approaches tend to yield good results on only few datasets.

Keywords: Causal graph discovery · noise-based approach · constraint-based approach · time series.

1 Introduction

Identifying causal structure from observational data is an important but also challenging task in many applications. Most causal graph discovery algorithms assume that causal relations can be described within causal graphs, where arrows encode causal information. For time series, the true complete causal graph $\mathcal{G} = (V, E)$ with V the set of vertices and E the set of edges, is called a *full time causal graph* and represents a complete graph of the dynamic system, through infinite vertices. In practice, inferring an infinite graph is unfeasible, so most algorithms assume that causal relations are consistent throughout time, *i.e.* for two time series X^p and X^q, if X^p_{t-i} causes X^q_t, denoted $X^p_{t-i} \rightarrow X^q_t$, then $X^p_{t-i-j} \rightarrow X^q_{t-j}$ for all j. Under this assumption, and given the maximal lag τ between cause and effect that can be present in the system, the full time causal graph can be

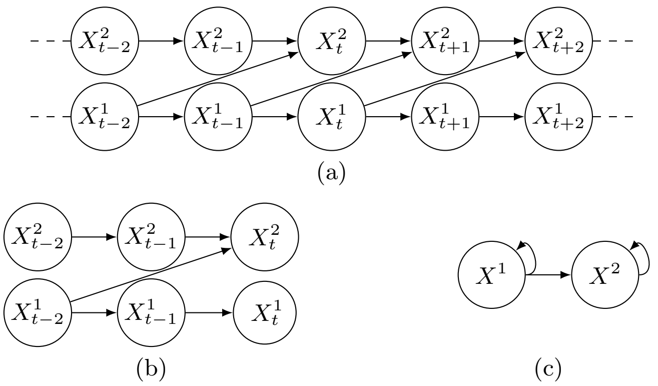


Fig. 1: Different causal graphs that one can infer from two time series X^1 and X^2: full time causal graph (a), window causal graph of size $\tau = 2$ (b) and summary causal graph (c).

contracted to give a finite graph which we call *window causal graph*, with $\tau + 1$ nodes for each time series [15]. However, sometimes, in practice, knowing only the causes of a given time series without necessarily knowing the time delay between the cause and the effect is all we need, so the true complete graph of the system is compressed even more via the so-called *summary graph* which represents causal relations between time series without referencing to lags [12]. Those notions are illustrated in Figure 1. Algorithms that detect causal relations can be classified according to the type of graph they look for.

Whatever the type of graph, the algorithms in question rely also on additional assumptions. The *Causal Markov Condition* states that in a causal graph, each node is independent of all other nodes given its parents, except its children [17]. The *causal sufficiency* states that there are no hidden common causes. One of the best known approach for causal discovery methods is the constraint-based approach that relies on conditional independencies and assume *faithfulness*, which states that the joint distribution P over V is faithful to the true causal Directed Acyclic Graph (DAG) $\mathcal{G}$ over V in the sense that every conditional independence statement satisfied by P is entailed by $\mathcal{G}$ [17].

Our contribution is two-fold. We introduce a new measure of dependence between two time series called the temporal causation entropy, which is an extension of the standard causation entropy measure [18] to time series to handle instantaneous relations and lags bigger than one. Based on it, we develop an algorithm to infer a summary causal graph from observational time series that is not limited to the Markov equivalent class even for instantaneous relations (which are common in practice due to the discretization of the time), that assumes causal Markov condition and a weaker version of faithfulness described in Section 3.1, and which is proved to be complete. The algorithm we propose is hybrid in the sense that it combines two different families of causal graph discovery methods: the noise-based family to find the potential causes of each time series and the constraint-based family to prune all unnecessary causes by looking at possible confounders and therefore end-up with only genuine cause. Remarkably, this is to our knowledge the first algorithm hybrid between constraint-based and noise-based methods for time series. Our evaluation, conducted on several datasets, illustrates the efficacy and efficiency of our approach.

The remainder of the paper is organized as follows: Section 2 describes related work. Section 3 then introduces the main causal discovery algorithm, called NBCB for Noise-Based / Constraint-Based approach. It relies on weak assumptions that are reminded first, and on the temporal causation entropy that is also introduced. The causal graph discovery algorithm we propose is illustrated and evaluated on several datasets in Section 4. Finally, Section 5 concludes the paper.

2 State of the Art

Granger Causality is one of the oldest methods proposed to detect causal relations between time series. However, this approach is known to handle a restricted version of causality that focuses on causal priorities as it assumes that the past of a cause is necessary and sufficient for optimally forecasting its effect [5]. The simplicity constitutes its advantage but also its limitations: for instance, it cannot deal with instantaneous effects. More recently, [11] exploited deep learning to learn causal relations between time series using an attention mechanism within convolutional networks. It infers a potential set of causes by analysing the estimated coefficients and then applies a validation step that is to some extent comparable to conditional Granger causality. However, in our experiments, we observe particularly bad results for TCDF.

Constraint-based approaches for time series are usually extended from causal graph discovery algorithm for nontemporal data. The main idea is to eliminate potential causes by finding conditional independencies in the data. The PC algorithm [17] is known to be the representative of this family of methods in case of i.i.d. data, which optimize the search of the smallest conditioning set needed to achieve separation between each pair of nodes. PCMCI [15] is an extension of PC for time series where a window causal graph is constructed, using temporal priority constraint to reduce the search space of the causal structure. oCSE [18] takes a different procedure compared to PC: instead of limiting as much as possible the size of its conditioning set, it conditions since the start on all potential causes which constitute the past of all available nodes. However, it limits its search for causal relations with a lag of one to find a summary graph . These methods usually assume causal Markov condition and faithfulness. Moreover, in general, graphs can only be recovered up to Markov equivalence[3] classes. In the context of time series, the notion of time can make such algorithms go beyond the Markov equivalence class but only for lagged relations [15], *i.e.* instantaneous relations are always limited to the Markov equivalence class.

[3] Two DAGs are Markov equivalent if and only if they have the same skeleton and the same v-structures [19]

Lastly, noise-based methods assume that the causal system can be defined by a set of equations that explain each variable by its direct causes and an additional noise. Causal relations are in this case discovered using footprints produced by the causal asymmetry in the data. For time series, the most well known algorithms in this family are tsLiNGAM [7], which is an extension of LiNGAM through autoregressive models, and TiMINo [12], which discovers a causal relationship by looking at independence between the noise and the potential causes. However, self-causation is not discovered within TiMINo, and the summary graph is assumed to be acyclic. These methods usually assume causal Markov condition and causal minimality, which is a weaker assumption than faithfulness. The main drawbacks are that such methods usually do not scale well [4] and might need a large sample size to achieve good performance [8].

The method we propose, as an hybrid method, takes benefit of the two approaches: it is not limited to a Markov equivalence class and provides a specific graph, scales better and needs a smaller sample size.

3 Causal graph discovery between two time series

Let us consider d univariate time series $X^1, \cdots, X^d$. Our goal is to find a summary causal graph between them, as represented in Fig. 1(c).

3.1 Assumptions

The faithfulness assumption is difficult to check in practice, and it has been debated for a long time. It assumes that there are no accidental conditional independence relations in the true distribution, that is, no conditional independence relations unless entailed by the true causal structure. The faithfulness assumption is mainly used in constraint-based methods, where it is used at two different stages, skeleton construction and edge orientation. As such, it can be decomposed into two assumptions, as proposed in [14], namely adjacency faithfulness and orientation faithfulness. As the orientation process we rely on differs from the one used in PC-like algorithms, we dispense here with the second assumption and solely rely on adjacency faithfulness, which is defined as follows:

Definition 1 (Adjacency faithfulness [14]). *For every $X^p, X^q \in V$, if X^p and X^q are adjacent in $\mathcal{G}$, then they are not conditionally independent given any subset of $V \backslash \{X^p, X^q\}$.*

As shown in [14], the relaxation of the faithfulness assumption still leads to provably correct skeletons.

Finally, the approach we propose discovers causal relations from time series under the causal Markov condition, the causal minimality condition (needed in the causal ordering, when using the noise-based method) and adjacency faithfulness (needed in the pruning step, when using the constraint-based method). We also assume that time series satisfy a causal ordering, meaning that we assume that the summary graph is acyclic. The inferred summary graph can however be cyclic, as opposed to [12], with loops between at least 3 time series.

3.2 Method

Our approach is a hybrid method which is decomposed into two parts. The first part, a noise-based approach, is described in Algorithm 1. It is based on a Gaussian process to map the past of the time series to the present, and a dependency measure between its input and its residuals to infer which time series potentially causes the other. The second part, a constraint-based approach, is described in Algorithm 2. It prunes the graph being constructed to remove spurious causes by considering the set of potential parents.The two parts are detailed below.

Causal ordering. The first step relies on noise-based approaches, which were initially introduced for i.i.d. data. However, they gained much attention in recent years [6, 9, 10, 1], and have also been extended for time series [12].

In this paper, we focus on Additive Noise Models (ANMs), which are defined as follows:

$$X_t^q = f\left([Par(X_t^q)_{t'}]_{t-\tau\leq t'\leq t}\right) + \xi_t^q \tag{1}$$

where f is a potentially nonlinear function, $Par(X_t^q)_{t'}$ is the set of parents of X_t^q at time-point t', $(\xi_t^q)_{q,t}$ are jointly independent; futhermore, for each q, ξ_t^q are identically distributed in t and the finite dimensional distributions for the time series $(X^q)_{1\leq q\leq d}$ are absolutely continuous wrt a product measure. Note that this model allows instantaneous relations. ANMs belong to the Identifiable Functional Model Class (IFMOC) [13], even in case of non-faithful causal models, for which conditional independence-based methods, as constraint-based, usually fail [13].

Similarly to the bivariate case [6, 10], the independence between the signal and the residuals allows one to detect the most probable cause from a set of variables through the following principle.

Principle 1 (Multivariate additive noise principle) *Suppose we are given a joint distribution $P(X^1, \cdots, X^d)$. If it satisfies an identifiable Additive Noise Model such that $\{(X_{t-j}^p)_{1\leq p\neq q\leq d, 0\leq j\leq\tau}, (X_{t-j}^q)_{1\leq j\leq\tau}\} \rightarrow X^q$, then it is likely that $\{(X_{t-j}^p)_{1\leq p\neq q\leq d, 0\leq j\leq\tau}, (X_{t-j}^q)_{1\leq j\leq\tau}\}$ precedes X^q in the causal order.*

Similarly to [10], when considering a suitable regression estimator and a suitable dependency estimator, the true causal order will be inferred. If we consider the fully connected graph given by this causal ordering (an edge between each node and its parents), it leads to a graph that contains the real graph as all true causal relations are in the inferred graph.

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

$$f_q : \{(X_{t-j}^p)_{1\leq p\neq q\leq d, 0\leq j\leq\tau}, (X_{t-j}^q)_{1\leq j\leq\tau}\} \mapsto X_t^q$$

by a Gaussian Process and deduce the residuals

$$\hat{\xi}_t^q = X_t^q - \hat{f}\{(X_{t-j}^p)_{1\leq p\neq q\leq d, 0\leq j\leq\tau}, (X_{t-j}^q)_{1\leq j\leq\tau}\}.$$

```
Algorithm 1: NBCB Part I: noise-based approach to order causes
  Result: $\mathcal{G}$
  $X$ a $d$-dimensional time series, $\tau$ a window size;
  $\mathcal{G}$ an empty graph with nodes $\{X^1, \ldots, X^d\}$; $S = \{1, \ldots, d\}$;
  while $length(S) > 1$ do
      for $j \in S$ do
          Learn $\hat{f}^j : \{(X^p_{t-j})_{p \in S, p \neq q, 0 \leq j \leq \tau}, (X^q_{t-j})_{1 \leq j \leq \tau}\} \mapsto X^j_t$;
          Deduce $\hat{\xi}^j_t$ and compute;
          $c_j$ from Eq. (2)
      Choose $j^* = \text{argmin}\, c_j$;
      $S = S \backslash j^*$;
      for $s \in S$ do
          Add $X^s \rightarrow X^{j^*}$ in $\mathcal{G}$;
```

The last place in the causal ordering (which belongs to the most probable effect of all other time series) is given to the time series which yields the residuals that are more independent to the other time series. The dependency between the residuals and the input is estimated with

$$c_q = C\left(\{(X^p_{t-j})_{1 \leq p \neq q \leq d, 0 \leq j \leq \tau}, (X^q_{t-j})_{1 \leq j \leq \tau}\}, \hat{\xi}^q_t\right), \tag{2}$$

where C is a dependence measure[4].

However, this method is not capable to detect independence between two time series, and thus it is susceptible to treat indirect causes as direct causes. To remove indirect causes or detect independencies, we complement this procedure with a second step that prunes spurious relations from the graph. It necessitates an exact estimation of the lag between two time series (through a maximum window of size τ).

Since this procedure uses a regression function estimator, it is subject to the curse of dimensionality when d is large compared to n. So we also consider a pairwise version of the procedure which consists on estimating for each pair of time series X^q, X^p two regression functions

$$f^q : \{(X^p_{t-j})_{0 \leq j \leq \tau}, (X^q_{t-j})_{1 \leq j \leq \tau}\} \mapsto X^q_t,$$
$$f^p : \{(X^q_{t-j})_{0 \leq j \leq \tau}, (X^p_{t-j})_{1 \leq j \leq \tau}\} \mapsto X^p_t.$$

We then compare the dependency of the residuals of those two functions with their inputs, and as before the potential cause is the one that is mapped by the function that yields the higher dependency, i.e. we choose the causal direction that yields the best bivariate ANM. While one cannot prove that the inferred graph contains the real one, numerical experiments show good performances for this method.

[4] As motivated in [12], we use the partial correlation to measure the dependence, but one can use our procedure with any measure.

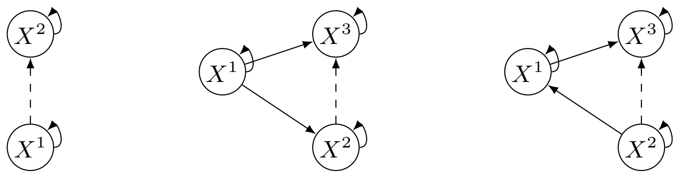


Fig. 2: Wrong causal relations potentially inferred in the first step of our algorithm. Dashed lines represents wrong causal relations. On the left, we show a spurious cause, whereas on the middle and on the right, we provide two indirect causes.

Pruning using temporal causation entropy. Knowing the list of potential parents of each time series, as detected in the previous step, one way to prune the causes that are not genuine is to conduct conditional independence tests between time series. Indeed, suppose X^p is a potential cause of X^q but X^p and X^q are conditionally independent, as illustrated in Figure 2. Then, we can conclude that X^p is not a cause of X^q.

In order to capture the dependencies (and conditional dependencies) between two time series, one needs to take into account the lag between them, as the true causal relations might not be instantaneous. Several studies have acknowledged the importance of taking into account lags to measure (conditional) dependencies between time series [5, 18]. Causation entropy, introduced in [18], is an asymmetric measure that detects the uncertainty reduction of the future states of X^q as a result of knowing the past states of X^p given that the past of $X^{\mathbf{R}}$ is already known, where $\mathbf{R}$ is a subset of $\{1, \cdots, d\}$. However, it only considers causation with a lag of size one, whereas it can take any values in practice.

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

Definition 2 (Temporal causation entropy). *We first define the optimal lag γ_{pq} between time series X^p and X^q and (λ_{pq}, λ_{qp}) the optimal windows of time series X^p regarding X^q and of time series X^q regarding X^p respectively as:*

$$\begin{aligned}\gamma_{pq}, \lambda_{pq}, \lambda_{qp} = \underset{\gamma \geq 0, \lambda_1, \lambda_2}{\operatorname{argmax}} \; & h(X^q_{t:t+\lambda_2} \mid X^q_{t-1}, X^p_{t-\gamma-1})) \\ & - h(X^q_{t:t+\lambda_2} \mid X^p_{t-\gamma-1:t-\gamma+\lambda_1}, X^q_{t-1}),\end{aligned}$$

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

$$\begin{aligned}TCE(X^p \rightarrow X^q \mid X^{\boldsymbol{R}}) = \min_{\Gamma_{r_i} \geq 0,\, 1 \leq i \leq K} \; & h(X^q_{t:t+\lambda_{qp}} \mid (X^{r_i}_{t-\Gamma_{pq|r_i}})_{1 \leq i \leq K}, X^q_{t-1}, X^p_{t-\gamma_{pq}-1})) \\ & - h(X^q_{t:t+\lambda_{qp}} \mid (X^{r_i}_{t-\Gamma_{pq|r_i}})_{1 \leq i \leq K}, X^p_{t-\gamma_{pq}-1:t-\gamma_{pq}+\lambda_{pq}}, X^q_{t-1}),\end{aligned}$$

Algorithm 2: NBCB part II: constraint-based approach for pruning

```
Result: $\mathcal{G}$
$X$ $d$-dimensional time series, $\alpha$ a significance threshold, $\mathcal{G}$ a causal graph;
n = 0;
while there exists $X^q \in V$ such that $card(Par(X^q, \mathcal{G})) \geq n+1$ do
    $\mathbf{D} = list()$;
    for $X^q \in V$ such that $card(Par(X^q, \mathcal{G})) \geq n+1$ do
        for $X^p \in Par(X^q, \mathcal{G})$, $X^{\mathbf{R}} \subset Par(X^q, \mathcal{G}) \setminus \{X^p\}$ with $card(X^{\mathbf{R}}) = n$ do
            $y_{q,p,\mathbf{R}} = \mathrm{TCE}(X^p; X^q \mid X^{\mathbf{R}})$;
            append($\mathbf{D}$, $\{X^q, X^p, X^{\mathbf{R}}\}$);
    Sort $\mathbf{D}$ by increasing order of $y$;
    while $\boldsymbol{D}$ is not empty do
        $\{X^q, X^p, X^{\mathbf{R}}\} = \mathrm{pop}(\mathbf{D})$;
        if $X^p \in Par(X^q, \mathcal{G})$ and $X^{\mathbf{R}} \subset Par(X^q, \mathcal{G})$ then
            Compute $z$ the p-value given by Eq. (3);
            if $z > \alpha$ then
                Remove edge $X^p \rightarrow X^q$ from $\mathcal{G}$;
    n=n+1;
```

where $\Gamma_{pq|r_1}, \cdots, \Gamma_{pq|r_K}$ *are the lags between* $X^{\boldsymbol{R}}$ *and* X^q.

First, the lag between X^p and X^q is detected by maximizing the dependency between X^p and X^q. As we measure the amount of information brought by the observations of one variable on the observations of another variable, taking the maximum ensures that one does not miss any possible information contributing to relating the two time series. In a second step, we find the lags between (X^p, X^q) and $X^{\mathbf{R}}$ that minimize the conditional dependency between X^p and X^q conditioned on $X^{\mathbf{R}}$. Taking the minimum ensures that we search for the lags that break the maximal dependence. Following the temporal priority principle, which states that causes precede their effects in time, we also ensure while finding only nonnegative lags that X^p as well as the conditional variables should precede in time X^q. If $\gamma = 1$ and $\lambda_{pq} = \lambda_{qp} = 1$, then the temporal causation entropy is equivalent to causation entropy when the latter is conditioned on the past.

In practice, the success of temporal causation entropy (and in fact, any entropy-based approaches) depends crucially on reliable estimation of the relevant entropies from data. This leads to two practical challenges. The first one is based on the fact that entropies must be estimated from finite time series data. To do so, we rely here on the k-NN estimator introduced in [3]. We denote by $\epsilon_{ik}/2$ the distance from

$$(X^p_{t-\gamma_{pq}:t-\gamma_{pq}+\lambda_{pq}}, X^q_{t:t+\lambda_{pq}}, ((X^{r_i}_{t-\Gamma_{pq|r_i}})_{1\leq i\leq K}, X^q_{t-1}, X^p_{t-\gamma_{pq}}))$$

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

$$(X^p_{t-\gamma_{pq}:t-\gamma_{pq}+\lambda_{pq}}, ((X^{r_i}_{t-\Gamma_{pq|r_i}})_{1\leq i\leq K}, X^q_{t-1}, X^p_{t-\gamma_{pq}})),$$

$$(X^q_{t:t+\lambda_{pq}}, ((X^{r_i}_{t-\Gamma_{pq|r_i}})_{1\leq i\leq K}, , X^q_{t-1}, X^p_{t-\gamma_{pq}}))$$

and

$$((X^{r_i}_{t-\Gamma_{pq|r_i}})_{1\leq i\leq K}, X^q_{t-1}, X^p_{t-\gamma_{pq}})$$

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

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

where ψ denotes the digamma function. The second problem is the following: to detect independence, we need a statistical test to check if the temporal causation entropy is equal to zero. We rely here on a permutation test:

Definition 3 (Permutation test of TCE). *Given X^p, X^q and $X^{\boldsymbol{R}}$, the p-value associated to the permutation test of TCE is given by:*

$$p = \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}_{\widehat{TCE}(b(X^p)\to X^q|X^{\boldsymbol{R}})\geq \widehat{TCE}(X^p\to X^q|X^{\boldsymbol{R}})}, \tag{3}$$

where $b(X^p)$ is a permuted version of X^p, $\mathbb{1}$ denotes the indicator function and B the maximum number of bootstrap sampling.

The method, detailed in Algorithm 2, can be summarized as follows. Starting with a fully directed graph (with one sided edges coming from a causal ordering), the first step consists in removing edges between nodes that are unconditionally independent: for each pair of nodes, a test of TCE is computed an edge is removed if the dependency, measured by TCE, is not significant given a threshold α. Once this is done, the algorithm checks, for the remaining oriented edges, whether two time series are conditionally independent or not given a set of parents of the arrow side node: in the first iteration the set of parents is of size one and then it gradually increases until either the edge between X^p and X^q is removed or all subsets of parents of X^q have been considered. Note that we make use of the same strategy as the one used in PC-stable [2], which consists in sorting time series according to their TCE scores and, when an independence is detected, removing all other occurrences of the time series. This leads to an order-independent procedure.

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

Theorem 4. *Given the true ordering of the causal process, Algorithm 2 is complete.*

Proof. Similarly to PC, Algorithm 2 prunes all unnecessary edges by removing edges that are conditionally independent given a subset S. Thanks to the causal order, the possible subsets space is reduced. By removing all links that are conditionally independent, by causal Markov condition, adjacency faithfulness and causal sufficiency, we are left with links that are directly causal and which are oriented wrt causal ordering.

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

3.3 Complexity analysis

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

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

where k represents the maximal degree of any vertex and each operation consists in conducting significance test to a conditional independence measure. Algorithms more adapted to time series, such as PCMCI [15], use the notion of time to reduce the number of tests. In those cases, the complexity would be divided by 2 (if instantaneous relations are not taken into account). NBCB is inferring a summary graph, which limits the number of decisions that need to be taken. NBCB's complexity in the worst case (when all relations are instantaneous) is bounded by:

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

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

4 Experiments

To illustrate the behavior of our method, we test it on several artificial and real datasets.

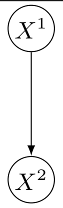
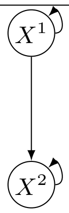
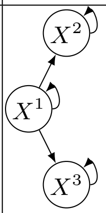
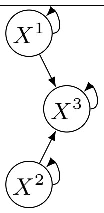
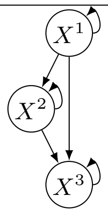
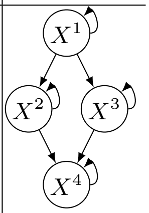
NBCB[5] and its pairwise version denoted pwNBCB are fitting a Gaussian Process with zero mean and squared exponential covariance function. The hyperparameters are automatically chosen by marginal likelihood optimization.

We compare NBCB with seven state-of-the-art methods: the constraint-based methods PCMCI[6] [15], where two variations of PCMCI are considered, varying

[5] Python code available at `https://github.com/kassaad/causal_discovery_for_time_series`

[6] Python code available at `https://github.com/jakobrunge/tigramite`

Table 1: Structures of simulated data.

Pair	Pair-sc	Fork	V-structure	Mediator	Diamond
					

the measure of independence between the mutual information for PCMCI-MI and the linear partial correlation for PCMCI-PC, and oCSE[7] [18]; the noise-based methods TiMINo[8] [12] with the linear time series model and VarLiNGAM[9] [7] where the regularization parameter in the adaptive Lasso is selected using BIC; the multivariate version of Granger Causality denoted GC[10] [5] and the Neural Network based method TCDF[11] [11] with default hyperparameters as introduced in the original paper. For all the methods, the best time lag is determined with the Akaike Information Criterion, the window size is set to $\tau = 5$ and the significant threshold for hypothesis testing to $\alpha = 0.05$.

In the different experimental settings, we compare the results wrt the *F1-score* denoted F1 of the orientations in the graph obtained without considering self causes, as it is treated differently depending on the methods.

4.1 Simulated data

We first test our method on simulated data generated from five different causal structures (pair, fork, V-structure, mediator, diamond) presented in Table 1. We distinguish pairs when the time series are self caused (Pair-sc) or not (Pair). For each benchmark, we generate randomly 10 data sets with 1000 observations. The data generating process is the following: for all q, $X_0^q = 0$ and for all $t > 0$,

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

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

[7] Python code available at `https://github.com/kassaad/causal_discovery_for_time_series`

[8] R code available at `http://web.math.ku.dk/~peters/code.html`

[9] Python code available at `https://github.com/cdt15/lingam`

[10] Matlab code available at `https://github.com/SacklerCentre/MVGC1`

[11] Python code available at `https://github.com/M-Nauta/TCDF`

Table 2: Results obtained on the simulated data for the different structures with 1000 observations. We report the mean and the standard deviation of the F1 score. The best results are in bold.

	Pair	Pair-sc	V-struct	Fork	Mediator	Diamond
NBCB	0.7 ± 0.46	0.7 ± 0.46	0.67 ± 0.28	0.67 ± 0.38	0.66 ± 0.32	0.71 ± 0.16
pwNBCB	0.7 ± 0.46	0.7 ± 0.46	0.75 ± 0.18	0.67 ± 0.38	0.7 ± 0.30	0.83 ± 0.12
PCMCI-PC	0.57 ± 0.47	0.6 ± 0.49	0.61 ± 0.33	0.53 ± 0.39	0.75 ± 0.24	0.63 ± 0.26
PCMCI-MI	0.9 ± 0.16	0.8 ± 0.4	0.67 ± 0.37	0.78 ± 0.17	0.84 ± 0.09	0.82 ± 0.16
oCSE	$\mathbf{1.0} \pm 0.0$	$\mathbf{1.0} \pm 0.0$	$\mathbf{0.90} \pm 0.16$	$\mathbf{0.8} \pm 0.12$	$\mathbf{0.95} \pm 0.08$	$\mathbf{0.88} \pm 0.09$
TiMINo	0.57 ± 0.49	0.5 ± 0.5	0.65 ± 0.37	0.52 ± 0.44	0.80 ± 0.19	0.60 ± 0.25
VarLiNGAM	0.54 ± 0.49	0.5 ± 0.5	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.03 ± 0.08
GC	0.67 ± 0.44	0.4 ± 0.49	0.37 ± 0.25	0.44 ± 0.38	0.83 ± 0.22	0.66 ± 0.26
TCDF	0.0 ± 0.0	0.1 ± 0.3	0.13 ± 0.26	0.26 ± 0.32	0.05 ± 0.15	0.16 ± 0.19

are considered: all the coefficients are random, or some coefficients are fixed to not be faithful to the true causal graph. Results are summarized in Table 2 for faithful data and in Table 3 for unfaithful data.

From Table 2, one can note that methods from the constraint-based family consistently outperform all other methods. However, our proposed algorithm is able to compete against pure constraint-based approaches. Specifically for the fork structure, the pairwise version outperforms all other methods, and for the others structures it performs better than most of the methods. When pwNBCB outperforms NBCB, one can expect that a larger sample size would improve the performance for NBCB, as the bivariate analysis is seen as a lower dimensional proxy of the full regression model. Furthermore, the similarity of results of our methods obtained by the F1 scores regarding to the structures illustrates the stability of our method. VarLINGAM performs particularly bad, but all the assumptions are violated in this design (Gaussian noise, non linearity). TCDF has also bad performances, whereas Granger Causality is surprisingly good, particularly for Mediator.

In Table 3 we consider two unfaithful datasets. The first one is a mediator, where $a^{13} = -a^{12}a^{23}$, without self cause, and all relations are instantaneous. Following [20], the second dataset is a linear unfaithful diamond without self causes, where we set the coefficient $a^{34} = -a^{12}a^{23}/a^{13}$ and all relations are instantaneous. From Table 3, we can see that PCMCI and oCSE perform poorly for unfaithful data, as expected. VarLINGAM has still bad results, again due to the simulation process. Our proposed algorithm comes out as one of the best algorithms in terms of performance, and there is an improvement with the full NBCB instead of its pairwise version.

Figure 3 provides an empirical illustration of the algorithmic complexity. We compare NBCB to `oCSE`, `PCMCI-MI` and `TiMINo` on four structures (v-structure, fork, mediator, diamond), sorted according to their number of nodes, their maximal out-degree and their maximal in-degree. The time is given in seconds. As one can note, `NBCB` is always faster than `oCSE` and `PCMCI-MI`, the difference be-

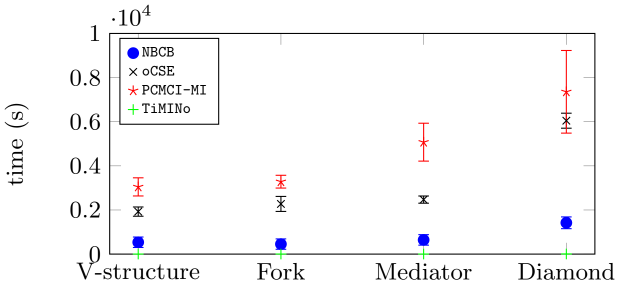


Fig. 3: Time computation (in second) for NBCB, oCSE and PCMCI-MI for four basic causal structures (V-structure, Fork, Mediator, Diamond). We report the mean and the standard deviation.

ing more important when the structure to be inferred is more complex. TiMINo is faster than NBCB, because the vector autoregressive model has been used instead of Gaussian process (there were not available in the online package) and the pruning in TiMINo relies on the noise-based approach, so less steps have to be considered. This illustrates again the trade-off between constraint-based and noise-based approaches.

4.2 Real data

Three different real datasets are considered in this study. We detail the performance of each method in the following paragraphs, but the results are summarized in Table 4.

Temperature This bivariate time series[12] of length 168 is about indoor X^{in} and outdoor X^{out} measurements. We expect that there is the following causal

[12] Data is available at https://webdav.tuebingen.mpg.de/cause-effect/

Table 3: Results obtained on the unfaithful simulated data for the different structures with 1000 observations. We report the mean and the standard deviation of the F1 score. The best results are in bold.

	unfaith. Mediator	unfaith. Diamond
NBCB	0.56 ± 0.26	$\mathbf{0.5} \pm 0.31$
pwNBCB	0.46 ± 0.23	0.39 ± 0.22
PCMCI-PC	0.21 ± 0.21	0.19 ± 0.16
PCMCI-MI	0.05 ± 0.15	0.20 ± 0.22
oCSE	0.05 ± 0.15	0.08 ± 0.16
TiMINo	$\mathbf{0.64} \pm 0.08$	0.49 ± 0.03
VarLiNGAM	0.0 ± 0.0	0.02 ± 0.06
GC	0.12 ± 0.27	0.14 ± 0.23
TCDF	0.4 ± 0.22	0.33 ± 0.17

Table 4: Results for real datasets. We report the mean and the standard deviation of the F1 score.

	Temperature	Diary	FMRI
NBCB	1	0.8	0.40 ± 0.21
pwNBCB	1	0.8	0.39 ± 0.21
PCMCI-PC	1	0.5	0.29 ± 0.19
PCMCI-MI	1	0.5	0.22 ± 0.18
oCSE	1	0.8	0.16 ± 0.20
TiMINo	0	0.0	0.32 ± 0.11
VarLiNGAM	0	0.0	0.49 ± 0.28
GC	0.66	0.33	0.24 ± 0.18
TCDF	0	0.0	0.07 ± 0.13

link: $X^{\text{out}} \rightarrow X^{\text{in}}$. VarLiNGAM wrongly infers no causal relation, Granger infers a bidirected arrow and TiMINo remains undecided. PCMCI-PC, PCMCI-MI, oCSE, NBCB and NBCBk correctly infer $X^{\text{out}} \rightarrow X^{\text{in}}$.

Diary This dataset[13] provides10 years (from 09/2008 to 12/2018) of monthly prices for milk X^m, butter X^b and cheddar cheese X^c, so the three time series are of length 124. We expect that the price of milk is a common cause of the price of butter and the price of cheddar cheese: $X^b \leftarrow X^m \rightarrow X^c$. VarLiNGAM wrongly infers X^b as common cause of X^m and X^c, Granger wrongly infers $X^m \leftrightarrow X^b \rightarrow X^c \rightarrow X^m$ and TiMINo only infers one wrong causal relation $X^c \rightarrow X^m$. TCDF infers no causal relation. PCMCI-PC and PCMCI-MI wrongly infer the causal chain $X^c \rightarrow X^m \rightarrow X^b$. oCSE, NBCB and pwNBCB correctly infer the causal relations but also add a wrong causal $X^c \rightarrow X^b$.

FMRI The last real-world dataset benchmark is about FMRI[14] (Functional Magnetic Resonance Imaging) that contains BOLD (Blood-oxygen-level dependent) datasets[16] for 28 different underlying brain networks. It measures the neural activity of different regions of interest in the brain based on the change of blood flow. There are 50 regions in total, each with its own associated time series. Since not all existing methods can handle 50 time series, datasets with more than 10 time series are excluded. In total we are left with 26 datasets containing between 5 and 10 brain regions. NBCB and VarLINGAM clearly outperforms other methods. All other methods are comparable, except TCDF which performs very poorly. Interestingly, PCMCI-PC performs better than PCMCI-MI, and VarLINGAM outperforms TiMINo which suggests the existence of linear causal relations.

[13] Data is available at `http://future.aae.wisc.edu`

[14] Original data is available at`https://www.fmrib.ox.ac.uk/datasets/netsim/index.html`, a preprocessed version is available at `https://github.com/M-Nauta/TCDF/tree/master/data/fMRI`

5 Conclusion

We have addressed in this study the problem of learning a summary causal graph on time series without being restricted to the Markov equivalent class even in the case of instantaneous relations. To do so, we followed a hybrid strategy. First we used a noise-based method to find the causal ordering between the time series under the assumption of additive noise models. Second, we used a constraint-based method to prune unnecessary parents and therefore ending up with an oriented causal graph. The second step heavily relies on a new temporal causation entropy measure that generalizes the causation entropy by removing the restriction of one time lag. Experiments conducted on different benchmark datasets and involving previous state-of-the-art proposals showed that the algorithm we have introduced outperforms previous proposals. In particular, we have illustrated and compared the behavior of our algorithm robustness wrt to different causal structures which yielded good results over all datasets, particularly on real ones.

In the future, we would like to test the method on large datasets, increasing both the number of time series d and the number of timepoints n. In particular, it would be interesting to study the quality of estimation in the regime $d >> n$.

Acknowledgements

This work has been partially supported by MIAI@Grenoble Alpes (ANR-19-P3IA-0003).

References

1. Assaad, K., Devijver, E., Gaussier, E., Ait-Bachir, A.: Scaling causal inference in additive noise models. In: Le, T.D., Li, J., Zhang, K., Cui, E.K.P., Hyvärinen, A. (eds.) Proceedings of Machine Learning Research. Proceedings of Machine Learning Research, vol. 104, pp. 22–33. PMLR, Anchorage, Alaska, USA (05 Aug 2019)
2. Colombo, D., Maathuis, M.H.: Order-independent constraint-based causal structure learning. Journal of Machine Learning Research **15**(116), 3921–3962 (2014)
3. Frenzel, S., Pompe, B.: Partial mutual information for coupling analysis of multivariate time series. Physical review letters **99**, 204101 (2007)
4. Glymour, C., Zhang, K., Spirtes, P.: Review of causal discovery methods based on graphical models. Frontiers in Genetics **10**, 524 (2019)
5. Granger, C.W.J.: Time series analysis, cointegration, and applications. The American Economic Review **94**(3), 421–425 (2004)
6. Hoyer, P.O., Janzing, D., Mooij, J.M., Peters, J., Schölkopf, B.: Nonlinear causal discovery with additive noise models. In: Advances in Neural Information Processing Systems 21. ACM Press (2009)
7. Hyvärinen, A., Shimizu, S., Hoyer, P.O.: Causal modelling combining instantaneous and lagged effects: An identifiable model based on non-gaussianity. In: Proceedings of the 25th International Conference on Machine Learning. pp. 424–431. ICML '08, ACM, New York, NY, USA (2008)

8. Malinsky, D., Danks, D.: Causal discovery algorithms: A practical guide. Philosophy Compass **13**(1) (2018)
9. Mooij, J., Janzing, D., Peters, J., Schölkopf, B.: Regression by dependence minimization and its application to causal inference in additive noise models. In: Proceedings of the 26th International Conference on Machine Learning. pp. 745–752. Max-Planck-Gesellschaft, ACM Press, New York, NY, USA (2009)
10. Mooij, J.M., Peters, J., Janzing, D., Zscheischler, J., Schölkopf, B.: Distinguishing cause from effect using observational data: Methods and benchmarks. Journal of Machine Learning Research **17**(1), 1103–1204 (2016)
11. Nauta, M., Bucur, D., Seifert, C.: Causal discovery with attention-based convolutional neural networks. Machine Learning and Knowledge Extraction **1**(1), 312–340 (2019)
12. Peters, J., Janzing, D., Schölkopf, B.: Causal inference on time series using restricted structural equation models. In: Advances in Neural Information Processing Systems 26. pp. 154–162 (2013)
13. Peters, J., Mooij, J.M., Janzing, D., Schölkopf, B.: Identifiability of causal graphs using functional models. In: Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence. p. 589–598. UAI'11, AUAI Press, Arlington, Virginia, USA (2011)
14. Ramsey, J., Spirtes, P., Zhang, J.: Adjacency-faithfulness and conservative causal inference. In: Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence. p. 401–408. UAI'06, AUAI Press, Arlington, Virginia, USA (2006)
15. Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., Sejdinovic, D.: Detecting and quantifying causal associations in large nonlinear time series datasets. Science Advances **5**(11) (2019)
16. Smith, S.M., Miller, K.L., Khorshidi, G.S., Webster, M.A., Beckmann, C.F., Nichols, T.E., Ramsey, J., Woolrich, M.W.: Network modelling methods for fmri. NeuroImage **54**, 875–891 (2011)
17. Spirtes, P., Glymour, C., Scheines, R.: Causation, Prediction, and Search. MIT press, 2nd edn. (2001)
18. Sun, J., Taylor, D., Bollt, E.: Causal network inference by optimal causation entropy. SIAM Journal on Applied Dynamical Systems **14**(1), 73–106 (2015)
19. Verma, T., Pearl, J.: Equivalence and synthesis of causal models. In: Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence. pp. 255–270. UAI '90, Elsevier Science Inc., New York, NY, USA (1991)
20. Zhalama, Zhang, J., Mayer, W.: Weakening faithfulness: some heuristic causal discovery algorithms. International Journal of Data Science and Analytics **3**, 93–104 (2016)