

Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets

Jakob Runge

German Aerospace Center
Institute of Data Science
07745 Jena, Germany

Abstract

We consider causal discovery from time series using conditional independence (CI) based network learning algorithms such as the PC algorithm. The PC algorithm is divided into a skeleton phase where adjacencies are determined based on efficiently selected CI tests and subsequent phases where links are oriented utilizing the Markov and Faithfulness assumptions. Here we show that autocorrelation makes the PC algorithm much less reliable with very low adjacency and orientation detection rates and inflated false positives. We propose a new algorithm, called PCMCI⁺ that extends the PCMCI method from [Runge et al., 2019b] to also include discovery of contemporaneous links. It separates the skeleton phase for lagged and contemporaneous conditioning sets and modifies the conditioning sets for the individual CI tests. We show that this algorithm now benefits from increasing autocorrelation and yields much more adjacency detection power and especially more orientation recall for contemporaneous links while controlling false positives and having much shorter runtimes. Numerical experiments indicate that the algorithm can be of considerable use in many application scenarios for dozens of variables and large time delays.

Spirtes and Zhang, 2016] conditional independence (CI) based network learning algorithms [Spirtes et al., 2000] form a main pillar for learning causal relations from data utilizing a number of assumptions. Here we focus on the PC algorithm [Spirtes and Glymour, 1991] as the main representative of conditional independence algorithms. Its advantages lie, firstly, in the flexibility of utilizing a wide and growing class of CI tests, from linear partial correlation (ParCorr) and non-parametric residual-based approaches [Ramsey, 2014, Runge et al., 2019b] to Kernel measures [Zhang et al., 2011], tests based on conditional mutual information [Runge, 2018b], and neural networks [Sen et al., 2017]. Secondly, the PC algorithm utilizes sparsity making it applicable also to large numbers of variables while autoregressive model-based approaches such as Granger causality strongly suffer from the curse of dimensionality [Runge et al., 2019b]. The PC algorithm is divided into a skeleton phase where adjacencies are determined based on CI tests and subsequent orientation phases where links are oriented based on the collider and further rules utilizing the Markov and Faithfulness assumptions [Spirtes et al., 2000].

Causal discovery in the time series case is partially less and partially more challenging [Runge et al., 2019a]. Obviously, time-order greatly helps in identifying causal directions for lagged links (causes precede effects), but properties such as non-stationarity [Malinsky and Spirtes, 2019] and especially autocorrelation can make the PC algorithm much less reliable. Here we show that autocorrelation, an ubiquitous property of time series, is especially detrimental leading to very low adjacency detection rates and subsequently low orientation recall for contemporaneous links. Due to the iterative nature of the PC algorithm, missed links then also lead to false positives, in addition to ill-calibrated CI tests [Runge, 2018a]. We propose a new algorithm, called PCMCI⁺ that extends the PCMCI method from [Runge et al., 2019b] to also include discovery of contemporaneous links. PCMCI⁺

1 Introduction

A number of frameworks address the problem of causal discovery from time series. Next to Bayesian score-based methods [Chickering, 2002], classical Granger causality [Granger, 1969], and more recent structural causal model frameworks [Peters et al., 2017,

is based on two central ideas: First, the skeleton phase is conducted separately for lagged and contemporaneous conditioning sets and the lagged phase uses much less tests leading to more power. Secondly, PCMC⁺ modifies the conditioning sets for the individual CI tests to make them well-calibrated under autocorrelation and increase detection power by utilizing the momentary conditional independence (MCI) approach [Runge et al., 2019b]. We show that this algorithm now benefits from increasing autocorrelation and yields much more adjacency detection power and especially more orientation recall for contemporaneous links while controlling false positives and having much shorter runtime.

2 Causal discovery for time series

2.1 Preliminaries

In the time series context we are interested in discovering *time series graphs* (e.g., [Runge, 2018a]) to represent the temporal dependency structure underlying complex dynamical systems. Consider an underlying time-dependent system $\mathbf{X}_t = (X_t^1, \dots, X_t^N)$ with

$$X_t^j = f_j \left(\mathcal{P}(X_t^j), \eta_t^j \right) \quad (1)$$

where f_j is some arbitrary measurable function with non-trivial dependencies on its arguments and η_t^j represents mutually ($i \neq j$) and serially ($t' \neq t$) independent dynamical noise. The nodes in a time series graph (example in Fig. 1) represent the variables X_t^j at different lag-times and the set of variables that X_t^j depends on defines the causal parents $\mathcal{P}(X_t^j) \subset \mathbf{X}_{t+1} = (\mathbf{X}_t, \mathbf{X}_{t-1}, \dots) \setminus \{X_t^j\}$. We denote *lagged parents* by $\mathcal{P}_t^-(X_t^j) = \mathcal{P}(X_t^j) \cap \mathbf{X}_t^-$. A causal link $X_{t-\tau}^i \rightarrow X_t^j$ exists if $X_{t-\tau}^i \in \mathcal{P}(X_t^j)$. For $\tau > 0$ we call them *lagged links* and for $\tau = 0$ we call $X_t^i \rightarrow X_t^j$ a contemporaneous link. The graph is actually infinite in time, but in practice only considered up to some maximum time lag τ_{\max} . Throughout this work we assume that the graph \mathcal{G} is *acyclic* and the process (1) is stable and thus *stationarity*, i.e., that all moments of the process are time invariant and the above definition for links at time t holds for links at every $t' \in \mathbb{Z}$. Then the variables $X_t^j \in \mathbf{X}_t$ together with their parents $\mathcal{P}(X_t^j)$ represent the time series graph \mathcal{G} . We define the set of adjacencies $\mathcal{A}(X_t^j)$ of a variable X_t^j to include all $X_{t-\tau}^i$ for $\tau \geq 0$ that have a (lagged or contemporaneous) link with X_t^j in \mathcal{G} . We define contemporaneous adjacencies as $\mathcal{A}_t(X_t^j) = \mathcal{A}(X_t^j) \cap \mathbf{X}_t$. A sequence of m contemporaneous links is called a *directed contemporaneous path* if for all $k \in \{1, \dots, m\}$ the link $X_{t+k-1}^i \rightarrow X_{t+k}^i$ occurs. We call X_t^i a *con-*

temporaneous ancestor of X_t^j if there is a directed contemporaneous path from X_t^i to X_t^j and we denote the set of all contemporaneous ancestors as $\mathcal{C}_t(X_t^j)$ (which excludes X_t^j itself). We denote separation in the graph by \bowtie , see [Runge, 2018a] for further notation details.

2.2 PC algorithm

The PC algorithm is the most wide-spread conditional independence-based causal discovery algorithm and utilizes the Markov and Faithfulness assumptions, as well as Causal Sufficiency (no unobserved common drivers) as formally defined in Sect. 3.2. It consists of three phases: First, a skeleton of adjacencies is learned based on iteratively testing which pairs of variables (at different time lags) are conditionally independent at some significance level α_{PC} (Algorithm 2 with the PC option). For lagged links, time-order automatically provides orientations, while for contemporaneous links a collider phase (Supplementary Algorithm S3) and rule phase (Supplementary Algorithm S4) determine the orientation of links. Nevertheless, conditional independence-based discovery algorithms can identify the contemporaneous graph structure only up to a Markov equivalence class represented as a completed partially directed acyclic graph (CPDAG). We denote links for which more than one orientation occurs in the Markov equivalence class by $X_t^i \circ - \circ X_t^j$. Here we consider a variant of PC that removes an unwanted dependence on the order of variables, called PC-stable [Colombo and Maathuis, 2014]. These modifications also include the *majority* or *conservative* [Ramsey et al., 2006] rules for handling ambiguous triples where separating sets are inconsistent, and conflicting links where different triples in the collider or orientation phase lead to opposite link orientations. Under the *conservative* rule the PC algorithm is consistent already under the weaker Adjacency Faithfulness condition. Our focus here is the causally sufficient case while extension of PC to latent variables in the time series case is treated in [Entner and Hoyer, 2010, Malinsky and Spirtes, 2019]. Another approach in the linear and causally sufficient case is to combine vector-autoregressive modeling to identify lagged links with the PC algorithm for the contemporaneous causal structure [Moneta et al., 2011].

2.3 The curse and blessing of autocorrelation

To illustrate the issue of autocorrelation, in Fig. 1 we consider a linear example with lagged and contemporaneous ground truth links shown for the PCMC⁺ case (right panel). The PC algorithm (Alg. 2) starts by testing all unconditional independencies. Here the coupled pairs (X^5, X^6) as well as (X^7, X^8) are independent of the

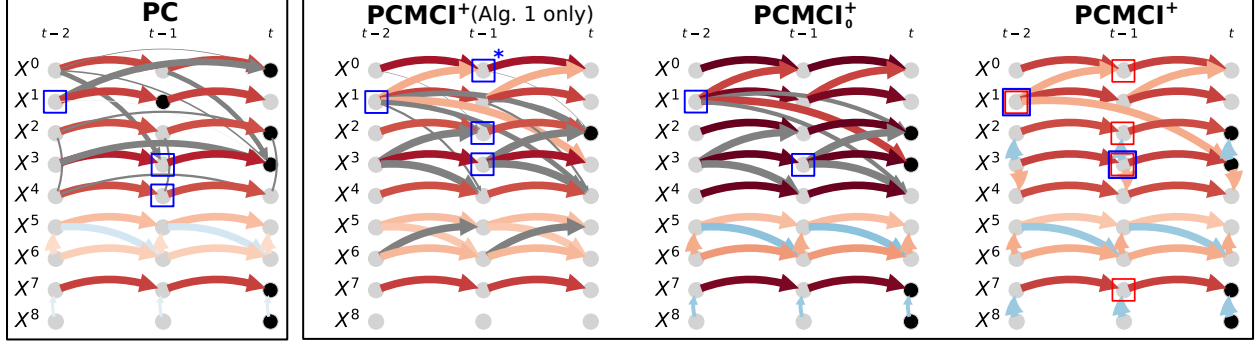


Figure 1: The curse and blessing of autocorrelation. Linear example of model (3) with lagged and contemporaneous ground truth links shown for the PCMCi⁺ case (right panel). All autodependency coefficients are 0.95 (except 0.475 for $X^{5,6}$) and all cross-coupling coefficients are 0.4 (\pm indicated by red/blue links). The panels show true and false link detection rates as the link width (if > 0.06) for true (color indicating partial correlation) and incorrect links (grey) for the PC algorithm, Algorithm 1, and the variants PCMCi⁺ and PCMCi₀⁺ as explained in the text (detection rates based on 500 realizations run at $\alpha_{PC} = 0.01$ for $T = 500$).

other variables and removed from each others adjacency sets which shows how PC exploits sparsity and reduces the estimation dimension compared to fitting a full model on the whole past as in the Granger causality framework [Runge et al., 2019b, Runge, 2018a]. But strong autocorrelation in the example will lead to many adjacencies among also the indirectly connected variables. In the next iteration, one-dimensional ($p = 1$) conditioning sets are tested for all remaining links. Here the PC algorithm misses the lagged link $X_{t-1}^1 \rightarrow X_t^0$ (black dots) due to the incorrect CI result $X_{t-1}^1 \perp\!\!\!\perp X_t^0 | X_{t-2}^1$ (condition marked by blue box). In a later stage this then leads to the false positive $X_{t-2}^0 \rightarrow X_t^0$ (grey link) since X_{t-1}^1 is not conditioned on. In a similar way the link $X_{t-2}^1 \rightarrow X_t^3$ is missed leading to the false positive $X_{t-1}^0 \rightarrow X_t^3$. This pattern of false negatives leading to false positives also occurs for contemporaneous links. Here $X_t^2 \circ - \circ X_t^3$ is removed when conditioning on $\mathcal{S} = (X_{t-1}^4, X_{t-1}^3)$ (blue boxes), which leads to the false positive autodependencies at lag 2 for X_t^2, X_t^4 , while the false autodependency at $X_{t-2}^3 \rightarrow X_t^3$ is due to missing $X_{t-2}^1 \rightarrow X_t^3$. This illustrates a cascade of false negative errors (missing links) leading to false positives.

What determines the removal of a true link? Detection power depends on sample size, the significance level α_{PC} , the CI test dimension ($p + 2$), and effect size, e.g., the absolute ParCorr (population) value, here denoted $I(X_{t-\tau}^i; X_t^j | \mathcal{S})$ for some conditioning set \mathcal{S} . In each p -iteration the sample size, α_{PC} , and the dimension are the same and a link will be removed if $I(X_{t-\tau}^i; X_t^j | \mathcal{S})$ is too small such that it falls below the significance threshold. Hence, $\min_{\mathcal{S}} [I(X_{t-\tau}^i; X_t^j | \mathcal{S})]$ determines whether a link is removed and the issue is that PC will iterate through all subsets of adjacencies such that the minimum can be-

come very small. Here $I(X_{t-1}^1; X_t^0 | X_{t-2}^1)$ is very small since X_{t-1}^1 shares much information with X_{t-2}^1 due to strong autocorrelation.

But autocorrelation can also be a blessing. The contemporaneously coupled pair (X^7, X^8) illustrates a case where autocorrelation helps to identify the orientation of the link. Without autocorrelation the output of PC would be an unoriented link to indicate the Markov equivalence class (can be addressed in structural causal model framework [Peters et al., 2017]). But, on the other hand, the detection rate here is very weak because the effect size (correlation) of $I(X_t^7; X_t^8)$ is very small due to high autocorrelation.

This illustrates the curse and blessing of autocorrelation. In summary, the PC algorithm yields many false negatives (low recall) and this then leads to false positives. False positives also occur due to ill-calibrated tests: Each individual CI test would need to account for autocorrelation, which is difficult to do in a complex multivariate and potentially nonlinear setting. The authors in [Runge, 2018a, Runge et al., 2019b] show that the CI tests then lead to inflated false positives.

As a side comment, the pair (X^5, X^6) depicts a feedback cycle. These often occur in real data and the example shows that time series graphs allow to resolve feedbacks in time while other graph modeling approaches could not represent a cyclic dependency. The orientation of the contemporaneous link $X_t^6 \rightarrow X_t^5$ is achieved via rule R1 in the orientation phase of PC (Supplementary Alg. S4).

3 PCMCI⁺

3.1 Algorithm

The goal of PCMCI⁺ is to overcome the problem of too many CI tests and to increase detection power and at the same time maintain well-calibrated tests. The approach is based on two central ideas, (1) separating the skeleton phase into a lagged conditioning phase with much less CI tests and (2) utilizing the momentary conditional independence test [Runge et al., 2019b] idea in the contemporaneous conditioning phase. Below we explain the reasoning behind.

Firstly, the goal of PC’s skeleton phase is to remove all those adjacencies that are due to indirect paths by conditioning on subsets \mathcal{S} of the nodes’ neighboring adjacencies in each iteration. Consider a variable X_t^j . If test lagged adjacencies from nodes $X_{t-\tau}^i$ by conditioning on the whole past, i.e., $\mathcal{S} = \mathbf{X}_t^- \setminus \{X_{t-\tau}^i\}$, the only indirect adjacencies remaining from variables in \mathbf{X}_t^- are due to paths through contemporaneous parents of X_t^j . This is in contrast to conditioning sets on contemporaneous adjacencies which can also induce paths through children of X_t^j . The reason for the PC algorithm to test *all* combinations of subsets \mathcal{S} is to avoid opening up such paths. However, conditioning on large-dimensional conditioning sets strongly affects detection power. The idea behind the lagged conditioning phase of PCMCI⁺ (Alg. 1) is to test all links $X_{t-\tau}^i \rightarrow X_t^j$ for $\tau > 0$ conditional on only the *strongest* p adjacencies in each p -iteration without going through all p -dimensional subsets of adjacencies. This speeds up the skeleton phase and, importantly, conducting fewer CI tests leads to much more detection power. We call the lagged adjacency set resulting from Alg. 1 $\hat{\mathcal{B}}_t^-(X_t^j)$. Lemma 1 proves that the only remaining indirect adjacencies in $\hat{\mathcal{B}}_t^-(X_t^j)$ are then due to paths passing through contemporaneous parents of X_t^j .

Secondly, in Alg. 2 the graph \mathcal{G} is initialized with all contemporaneous adjacencies plus all lagged adjacencies from $\hat{\mathcal{B}}_t^-(X_t^j)$ for all X_t^j . Algorithm 2 tests all (unordered lagged and ordered contemporaneous) adjacent links $(X_{t-\tau}^i, X_t^j)$ and iterates through contemporaneous conditions $\mathcal{S} \subseteq \mathcal{A}_t(X_t^j)$, but in addition each CI test is conditioned on $\hat{\mathcal{B}}_t^-(X_t^j)$ to block paths through lagged parents. There are two versions of CI tests now that both lead to asymptotically consistent algorithms:

$$\begin{aligned} \text{PCMCI}^+: X_{t-\tau}^i \perp\!\!\!\perp X_t^j | \mathcal{S}, \hat{\mathcal{B}}_t^-(X_t^j) \setminus \{X_{t-\tau}^i\}, \hat{\mathcal{B}}_{t-\tau}^-(X_{t-\tau}^i) \\ \text{PCMCI}_0^+: X_{t-\tau}^i \perp\!\!\!\perp X_t^j | \mathcal{S}, \hat{\mathcal{B}}_t^-(X_t^j) \setminus \{X_{t-\tau}^i\} \end{aligned} \quad (2)$$

Both versions are followed by the collider orientation phase (Alg. S3) and rule orientation phase (Alg. S4) which we defer to the Supplementary Material since they

are equivalent to the usual PC algorithm with the modification that the additional CI tests in the collider phase for the conservative and majority rule are also based on the tests 2.

We now discuss these two versions on the example in Fig. 1. Algorithm 1 now tests $X_{t-1}^1 \rightarrow X_t^0$ conditional on $\mathcal{S} = X_{t-1}^0$ for $p = 1$ and $\mathcal{S} = (X_{t-1}^0, X_{t-2}^1)$ for $p = 2$ as the two strongest adjacencies (as determined by the test statistic value, see pseudo-code). In both of these tests the effect size I is much larger than for the condition on $\mathcal{S} = (X_{t-2}^1)$ which lead to the removal of $X_{t-1}^1 \rightarrow X_t^0$ in the PC algorithm. In general, conditioning on the strongest adjacencies will yield larger effect sizes (increase the ‘causal signal-to-noise ratio’) and, hence, higher detection power. Below we provide a more rigorous treatment of effect size. In the example $\hat{\mathcal{B}}_t^-(X_t^2)$ is indicated as blue boxes in the second panel and contains lagged parents as well as adjacencies due to paths passing through contemporaneous parents of X_t^2 . One false positive, likely due to an ill-calibrated test caused by autocorrelation, is marked by a star. Based on these lagged adjacencies, PCMCI₀⁺ then recovers all lagged links (3rd panel), but it still misses contemporaneous adjacencies $X_t^2 \circ - X_t^3$ and $X_t^3 \circ - X_t^4$ and we also see strong lagged false positives from X^3 to X^2 and X^4 . What happened here? The problem are now tests on contemporaneous links: The CI test (2) for PCMCI₀⁺ in the $p = 0$ loop, like the original PC algorithm will test *ordered* pairs, hence first $X_t^2 \circ - X_t^3$ conditional on $\hat{\mathcal{B}}_t^-(X_t^3)$ and, if the link is not removed, $X_t^3 \circ - X_t^2$ conditional on $\hat{\mathcal{B}}_t^-(X_t^2)$. In general, a contemporaneous link is removed if $\min(I(X_t^i; X_t^j | \hat{\mathcal{B}}_t^-(X_t^j)), I(X_t^j; X_t^i | \hat{\mathcal{B}}_t^-(X_t^i)))$ is smaller than the significance threshold. Here $X_t^2 \circ - X_t^3$ is removed conditional on $\hat{\mathcal{B}}_t^-(X_t^3)$ (indicated by blue boxes in the panel) because $I(X_t^3; X_t^2 | \hat{\mathcal{B}}_t^-(X_t^3))$ is very small.

To fix this, we employ the central PCMCI⁺ idea to condition on *both* lagged adjacencies in the CI test (2) for PCMCI⁺ (see blue and red boxes in Fig. 1). At least for the initial phase $p = 0$ one can prove that the effect size of the PCMCI⁺ CI test is always larger than that of the PCMCI₀⁺ test (Thm. 4). For $p > 0$ either may be larger, but in our numerical experiments we found that this helps to increase effect size and detection power for contemporaneous links. PCMCI⁺ now recovers all lagged as well as contemporaneous links and also correctly unveils the lagged false positives that PCMCI₀⁺ obtains. Also the contemporaneous coupled pair (X^7, X^8) is now much better detected since $I(X_t^7; X_t^8 | X_{t-1}^7) > I(X_t^7; X_t^8)$. Another advantage, discussed in [Runge et al., 2019b] is that PCMCI⁺ CI tests are well-calibrated, in contrast to PCMCI₀⁺ tests, since the condition on both parents re-

moves autocorrelation effects. Note that for lagged links the effect size of PCMCI^+ is generally smaller than that of PCMCI_0^+ since the extra condition on $\widehat{\mathcal{B}}_{t-\tau}^-(X_{t-\tau}^i)$ always reduce the information. However, there is a trade off between allowing inflated false positives and detection power. In summary, the central PCMCI^+ idea is to increase effect size in individual CI tests to achieve higher detection power and at the same time maintain well-controlled false positives also for high autocorrelation. Correct adjacency information then leads to better orientation recall in Alg. S3,S4. The other advantage of PCMCI^+ compared to PC is much faster and, as our numerical examples show, also much less variable runtime.

The full algorithm is detailed in pseudo-code Algorithms 1,2,S3,S4 with differences to PC and PCMCI_0^+ indicated. Based on the PC stable variant, PCMCI^+ is fully order-independent. We here show the majority-rule implementation of the collider phase, the version without handling ambiguous triples and for the conservative rule are detailed in Supplementary Alg. S3. Note that the tests in the collider phase also use the CI tests (2). One can construct (rather conservative) p -values for the skeleton adjacencies $(X_{t-\tau}^i, X_t^j)$ by taking the maximum p -value over all CI tests conducted in Alg. 2. Correspondingly, we define a link strength corresponding to the test statistic value of the maximum p -value. The free parameter α_{PC} can be chosen based on cross-validation or the BIC score as discussed in [Colombo and Maathuis, 2014]. One can further speed up the algorithm by caching CI tests since these might sometimes be repeated. Further variants of PCMCI^+ include a version similar to PCMCI for the lagged case [Runge et al., 2019b] where Alg. 2 is initialized with a fully connected graph also for lagged links instead of restricting it to $\widehat{\mathcal{B}}_t^-(X_t^j)$. This would increase detection power for lagged links, but also lead to slower runtime.

The computational complexity of PCMCI^+ strongly depends on the network structure and the parameter α_{PC} . The sparser the causal dependencies, the faster the convergence. Compared to the original PC algorithm with worst-case exponential complexity, the complexity is much reduced since Algorithm 1 only has polynomial complexity [Runge et al., 2019b] and Algorithm 2 only iterates through contemporaneous conditioning sets, hence the worst-case exponential complexity only applies to N and not to $N\tau_{\max}$. PCMCI_0^+ is slightly faster than PCMCI^+ because it has slightly lower dimensional conditions, as further investigated in the numerical experiments.

3.2 Asymptotic consistency

The asymptotic consistency of PCMCI^+ (including that of PCMCI_0^+) follows relatively straightforward from that of the original PC algorithm. Note that we can add the time-order and stationarity assumptions to the standard assumptions of causal discovery. The consistency of network learning algorithms is separated into *soundness*, i.e., the returned graph has correct adjacencies, and *completeness*, i.e., the returned graph is also maximally informative (links are oriented as much as possible). We start with the following assumptions.

Assumptions 1 (Asymptotic case). *Throughout this paper we assume Causal Sufficiency, the Causal Markov Condition, the Adjacency Faithfulness Conditions, and consistent conditional independence tests (oracle). In the present time series context we also assume stationarity and time-order. Furthermore, we rule out selection variables [Spirtes et al., 2000, Zhang, 2008] and measurement error.*

Detailed definitions of these assumptions, adapted from [Spirtes et al., 2000] to the time series context, are given in Supplementary Sect. S1 and all proofs are given in Sect. S2. To prove the consistency of PCMCI^+ we start with the following lemma.

Lemma 1. *Under Assumptions 1 Algorithm 1 returns a set that always contains the parents of X_t^j and, at most, the lagged parents of all contemporaneous ancestors of X_t^j , i.e., $\widehat{\mathcal{B}}_t^-(X_t^j) = \mathcal{P}_t^-(X_t^j, \mathcal{C}_t(X_t^j))$.*

$\widehat{\mathcal{B}}_t^-(X_t^j)$ contains *all* lagged parents of all contemporaneous ancestors only if the weaker Adjacency Faithfulness assumption is replaced by standard Faithfulness.

This establishes that the conditions $\widehat{\mathcal{B}}_t^-(X_t^j)$ estimated in the first step of PCMCI^+ will suffice to block all lagged confounding paths that do not go through contemporaneous links. This enables to prove the soundness of Algorithm 2, even though Algorithm 2 is a variant of the PC algorithm that only iterates through contemporaneous conditioning sets.

Theorem 1 (Soundness of PCMCI^+). *Algorithm 2 returns the correct adjacencies under Assumptions 1, i.e., $\widehat{\mathcal{G}}^* = \mathcal{G}^*$, where the $*$ denotes the skeleton of the time series graph.*

The proof follows from the proof for the PC algorithm taking into account the slightly different conditioning sets.

To prove the completeness of PCMCI^+ , we start with the following observation.

Lemma 2. *Due to time-order and the stationarity assumption, the considered triples in the collider phase (Al-*

Algorithm 1 (PCMCI⁺ / PCMCI₀⁺ lagged skeleton phase)

Require: Time series dataset $\mathbf{X} = (X^1, \dots, X^N)$, max. time lag τ_{\max} , significance threshold α_{PC} , CI test $\text{CI}(X, Y, \mathbf{Z})$ returning p -value and test statistic value I

- 1: **for all** X_t^j in \mathbf{X}_t **do**
 - 2: Initialize $\widehat{\mathcal{B}}_t^-(X_t^j) = \mathbf{X}_t^- = (\mathbf{X}_{t-1}, \dots, \mathbf{X}_{t-\tau_{\max}})$ and $I^{\min}(X_{t-\tau}^i, X_t^j) = \infty \ \forall X_{t-\tau}^i \in \widehat{\mathcal{B}}_t^-(X_t^j)$
 - 3: Let $p = 0$
 - 4: **while** any $X_{t-\tau}^i \in \widehat{\mathcal{B}}_t^-(X_t^j)$ satisfies $|\widehat{\mathcal{B}}_t^-(X_t^j) \setminus \{X_{t-\tau}^i\}| \geq p$ **do**
 - 5: **for all** $X_{t-\tau}^i$ in $\widehat{\mathcal{B}}_t^-(X_t^j)$ satisfying $|\widehat{\mathcal{B}}_t^-(X_t^j) \setminus \{X_{t-\tau}^i\}| \geq p$ **do**
 - 6: $\mathcal{S} =$ first p variables in $\widehat{\mathcal{B}}_t^-(X_t^j) \setminus \{X_{t-\tau}^i\}$
 - 7: $(p\text{-value}, I) \leftarrow \text{CI}(X_{t-\tau}^i, X_t^j, \mathcal{S})$
 - 8: $I^{\min}(X_{t-\tau}^i, X_t^j) = \min(|I|, I^{\min}(X_{t-\tau}^i, X_t^j))$
 - 9: **if** $p\text{-value} > \alpha_{\text{PC}}$ **then** mark $X_{t-\tau}^i$ for removal
 - 10: Remove non-significant entries and sort $\widehat{\mathcal{B}}_t^-(X_t^j)$ by $I^{\min}(X_{t-\tau}^i, X_t^j)$ from largest to smallest
 - 11: Let $p = p + 1$
 - 12: **return** $\widehat{\mathcal{B}}_t^-(X_t^j)$ for all X_t^j in \mathbf{X}_t
-

Algorithm 2 (PCMCI⁺ / PCMCI₀⁺ contemporaneous skeleton phase / PC full skeleton phase)

Require: Time series dataset $\mathbf{X} = (X^1, \dots, X^N)$, max. time lag τ_{\max} , significance threshold α_{PC} , $\text{CI}(X, Y, \mathbf{Z})$, PCMCI⁺ / PCMCI₀⁺: $\widehat{\mathcal{B}}_t^-(X_t^j)$ for all X_t^j in \mathbf{X}_t

- 1: PCMCI⁺ / PCMCI₀⁺: Form time series graph \mathcal{G} with lagged links from $\widehat{\mathcal{B}}_t^-(X_t^j)$ for all X_t^j in \mathbf{X}_t and fully connect all contemporaneous variables, i.e., add $X_t^i \circ \circ X_t^j$ for all $X_t^i \neq X_t^j \in \mathbf{X}_t$
 - 2: PCMCI⁺ / PCMCI₀⁺: Initialize contemporaneous adjacencies $\widehat{\mathcal{A}}(X_t^j) := \widehat{\mathcal{A}}_t(X_t^j) = \{X_t^i \neq X_t^j \in \mathbf{X}_t : X_t^i \circ \circ X_t^j \text{ in } \mathcal{G}\}$
 - 3: PC: Initialize full adjacencies $\widehat{\mathcal{A}}(X_t^j)$ for all (lagged and contemporaneous) links in \mathcal{G}
 - 4: Initialize $I^{\min}(X_{t-\tau}^i, X_t^j) = \infty$ for all links in \mathcal{G}
 - 5: Let $p = 0$
 - 6: **while** any adjacent, ordered pairs $(X_{t-\tau}^i, X_t^j)$ for $\tau \geq 0$ in \mathcal{G} satisfy $|\widehat{\mathcal{A}}(X_t^j) \setminus \{X_{t-\tau}^i\}| \geq p$ **do**
 - 7: Select new adjacent, ordered pair $(X_{t-\tau}^i, X_t^j)$ for $\tau \geq 0$ satisfying $|\widehat{\mathcal{A}}(X_t^j) \setminus \{X_{t-\tau}^i\}| \geq p$
 - 8: **while** $(X_{t-\tau}^i, X_t^j)$ are adjacent in \mathcal{G} and not all $\mathcal{S} \subseteq \widehat{\mathcal{A}}(X_t^j) \setminus \{X_{t-\tau}^i\}$ with $|\mathcal{S}| = p$ have been considered **do**
 - 9: Choose new $\mathcal{S} \subseteq \widehat{\mathcal{A}}(X_t^j) \setminus \{X_{t-\tau}^i\}$ with $|\mathcal{S}| = p$
 - 10: PCMCI⁺: Set $\mathbf{Z} = (\mathcal{S}, \widehat{\mathcal{B}}_t^-(X_t^j) \setminus \{X_{t-\tau}^i\}, \widehat{\mathcal{B}}_{t-\tau}^-(X_{t-\tau}^i))$
 - 11: PCMCI₀⁺: Set $\mathbf{Z} = (\mathcal{S}, \widehat{\mathcal{B}}_t^-(X_t^j) \setminus \{X_{t-\tau}^i\})$
 - 12: PC: Set $\mathbf{Z} = \mathcal{S}$
 - 13: $(p\text{-value}, I) \leftarrow \text{CI}(X_{t-\tau}^i, X_t^j, \mathbf{Z})$
 - 14: $I^{\min}(X_{t-\tau}^i, X_t^j) = \min(|I|, I^{\min}(X_{t-\tau}^i, X_t^j))$
 - 15: **if** $p\text{-value} > \alpha_{\text{PC}}$ **then**
 - 16: Delete link $X_{t-\tau}^i \rightarrow X_t^j$ for $\tau > 0$ (or $X_t^i \circ \circ X_t^j$ for $\tau = 0$) from \mathcal{G}
 - 17: Store (unordered) sepset $(X_{t-\tau}^i, X_t^j) = \mathcal{S}$
 - 18: Let $p = p + 1$
 - 19: Re-compute $\widehat{\mathcal{A}}(X_t^j)$ from \mathcal{G} and sort by $I^{\min}(X_{t-\tau}^i, X_t^j)$ from largest to smallest
 - 20: **return** \mathcal{G} , sepset
-

gorithm S3) and rule orientation phase (Algorithm S4) can be restricted as follows: In the collider orientation phase only unshielded triples $X_{t-\tau}^i \rightarrow X_t^k \circ \circ X_t^j$ (for $\tau > 0$) or $X_t^i \circ \circ X_t^k \circ \circ X_t^j$ (for $\tau = 0$) in \mathcal{G} where $(X_{t-\tau}^i, X_t^j)$ are not adjacent are relevant. For

orientation rule R1 triples $X_{t-\tau}^i \rightarrow X_t^k \circ \circ X_t^j$ where $(X_{t-\tau}^i, X_t^j)$ are not adjacent, for orientation rule R2 triples $X_t^i \rightarrow X_t^k \rightarrow X_t^j$ with $X_t^i \circ \circ X_t^j$, and for orientation rule R3 pairs of triples $X_t^i \circ \circ X_t^k \rightarrow X_t^j$ and $X_t^i \circ \circ X_t^l \rightarrow X_t^j$ where (X_t^k, X_t^l) are not adjacent and

$X_t^i \circ - \circ X_t^j$ are relevant. These restrictions imply that only contemporaneous separating sets are relevant for the collider orientation phase.

With this lemma the completeness of PCMCI^+ (and PCMCI_0^+) follows straightforwardly from the completeness proof of the original PC algorithm [Meek, 1995, Spirtes et al., 2000].

Theorem 2 (PCMCI^+ is complete). *PCMCI⁺ (Algorithms 1,2,S3,S4) when used with the conservative rule for orienting colliders in Algorithm S3 returns the correct CPDAG under Assumptions 1. Under standard Faithfulness also PCMCI⁺ without no rule or the majority rule is complete.*

3.3 Order independence

Also the proof of order-independence follows straightforwardly from the proof in [Colombo and Maathuis, 2014].

Theorem 3 (Order independence). *Under Assumptions 1 PCMCI⁺ with the conservative or majority rule in Algorithm S3 is independent of the order of variables (X^1, \dots, X^N) .*

Of course, order independence does not apply to time-order. Order independence also trivially holds for PCMCI [Runge et al., 2019b] under the additional assumption of no contemporaneous causal links.

3.4 Effect size and false positive control

The toy example showed that a major problem of PCMCI_0^+ (and also PC) is lack of detection power for contemporaneous links. A main factor of statistical detection power is effect size, i.e., the population value of the test statistic considered (e.g., absolute partial correlation). In the following, we will base our argument in an information-theoretic framework and consider the conditional mutual information as a general test statistic, denoted I . In Algorithm 2 PCMCI_0^+ will test a contemporaneous dependency $X_t^i \circ - \circ X_t^j$ first with the test statistic $I(X_t^i; X_t^j | \widehat{\mathcal{B}}_t^-(X_t^j))$, and, if that test was positive, secondly with $I(X_t^i; X_t^j | \widehat{\mathcal{B}}_t^-(X_t^i))$. If either of these tests finds (conditional) independence, the adjacency is removed. Therefore, the minimum test statistic value determines the relevant effect size. On the other hand, PCMCI^+ treats both cases symmetrically since the test statistic is always $I(X_t^i; X_t^j | \widehat{\mathcal{B}}_t^-(X_t^j), \widehat{\mathcal{B}}_t^-(X_t^i))$.

Theorem 4 (Effect size of MCI tests for $p = 0$). *Under Assumptions 1 the PCMCI⁺ CI tests in Algorithm 2 for $p = 0$ for contemporaneous true links $X_t^i \rightarrow X_t^j \in \mathcal{G}$ have an effect size that is always greater than that of the*

PCMCI₀⁺ CI tests, i.e., $I(X_t^i; X_t^j | \widehat{\mathcal{B}}_t^-(X_t^j), \widehat{\mathcal{B}}_t^-(X_t^i)) > \min(I(X_t^i; X_t^j | \widehat{\mathcal{B}}_t^-(X_t^j)), I(X_t^i; X_t^j | \widehat{\mathcal{B}}_t^-(X_t^i)))$.

For lagged links $X_{t-\tau}^i \rightarrow X_t^j$ PCMCI_0^+ will only conduct tests based on the lagged conditions of X_t^j . Then it holds that $I(X_{t-\tau}^i; X_t^j | \widehat{\mathcal{B}}_t^-(X_t^j) \setminus \{X_{t-\tau}^i\}, \widehat{\mathcal{B}}_{t-\tau}^-(X_{t-\tau}^i)) \leq I(X_{t-\tau}^i; X_t^j | \widehat{\mathcal{B}}_t^-(X_t^j) \setminus \{X_{t-\tau}^i\})$, i.e., the effect size of the PCMCI^+ tests is always smaller (or equal) than that of the PCMCI_0^+ tests (see [Runge et al., 2012]) which explains slightly higher detection power for lagged links found in the experiments below. For $p > 0$ in Algorithm 2 the conditioning on contemporaneous neighbors \mathcal{S} can lead to different effect sizes, but only if both $X_{t-\tau}^i$ and X_t^j cause \mathcal{S} since this opens a collider path through \mathcal{S} . Hence, while greater effect size for PCMCI^+ does not always obtain, it is difficult to optimize without knowing the graph and we found the PCMCI^+ to outperform PCMCI_0^+ for contemporaneous links in our numerical experiments, as also shown in the toy example.

While this result regards detection power, in the following we give a mathematical intuition why the CI tests in MCI tests are better calibrated and control false positives below the expected significance level. Lemma 1 implies that even though Algorithm 1 does not aim to estimate the contemporaneous parents, it still yields a set of conditions that shields X_t^j from the ‘infinite’ past \mathbf{X}_t^- , either by blocking the parents of X_t^j or by blocking indirect contemporaneous paths through contemporaneous ancestors of X_t^j . Blocking paths from the infinite past, we conjecture, is key to achieve well-calibrated CI tests in Algorithm 2. The authors in [Runge et al., 2019b] showed that under certain model assumptions the MCI tests reduce to CI tests among the noise terms η from model (1), which are assumed to be i.i.d. and help to achieve well-calibrated CI tests. In the Supplement we give numerical evidence that PCMCI_0^+ and PC have inflated false positive for high autocorrelation, while PCMCI^+ well controls false positives, but a formal proof of correct false positive control is beyond the scope of this paper.

4 Numerical experiments

We compare PC, PCMCI_0^+ , and PCMCI^+ with CI tests based on linear partial correlation (ParCorr) and PC and PCMCI^+ for the GPDC test [Runge et al., 2019b] that is based on Gaussian process regression and a *distance correlation* test on the residuals, which is suitable for a large class of nonlinear dependencies with additive noise. We model five typical properties of time series from complex systems: contemporaneous and time lagged causal de-

dependencies, strong autocorrelation, large number of variables, and nonlinearity. Consider the following additive variant of model (1):

$$X_t^j = a_j X_{t-1}^j + \sum_i c_i f_i(X_{t-\tau_i}^i) + \eta_t^j \quad (3)$$

for $j \in \{1, \dots, N\}$. Autocorrelations a_j are uniformly drawn from $[\max(0, a - 0.3), a]$ for some a as indicated in Fig. 2 and $\eta^j \sim \mathcal{N}(0, 1)$ is *iid* Gaussian noise. In addition to autodependency links, for each model we randomly choose $L = N$ ($L = 1$ for $N = 2$) cross-links whose functional dependencies are linear for ParCorr experiments and, in addition, $f_i(x) = (1 + 5xe^{-x^2/20})x$ for GPDC experiments. Coefficients c_i are drawn uniformly from $\pm[0.1, 0.5]$. 30% of the links are contemporaneous ($\tau_i = 0$) and the remaining τ_i are drawn from $[1, \tau_{\max}]$. We only consider stationary models. With $L = N$ links in each model, we have an average cross-in-degree of 1 for all network sizes (plus an auto-dependency) implying that models become sparser for larger N .

In Fig. 2 we evaluate performance as follows: True and false positive rates for adjacencies (based on the skeleton output of Algorithm 2) are distinguished between lagged cross-links ($i \neq j$), contemporaneous, and autodependency links. Due to time order, lagged links (and autodependencies) are automatically oriented. For orientation performance of contemporaneous links we have to take into account that the output of the present PC and PCMCI versions contains unoriented and conflicting links that are counted separately. Orientation precision is measured as the fraction of correctly oriented contemporaneous links, and recall as the fraction of true contemporaneous links detected. We further show the fraction of unoriented links and of conflicting links among all detected contemporaneous adjacencies. All metrics are computed across all estimated graphs from 500 realizations of model (3) at time series length T . The average (and std.) runtime estimates per graph estimate were evaluated on Intel Xeon E5-2695 v4 (Broadwell) at 2.1GHz.

4.1 Effect of autocorrelation

Figure 2A shows results for varying autocorrelation a (on x -axis) for linear experiments and the ParCorr CI tests for $N = 5$, $T = 500$, $\alpha_{PC} = 0.01$, and $\tau_{\max} = 5$ in the majority-rule variants of PC, PCMCI_0^+ and PCMCI^+ . Adjacency detection rates of PCMCI^+ for contemporaneous and lagged links are stable even under high autocorrelation with only a slight decrease for lagged links. PC has similar rates for small a , but quickly loses power for increasing autocorrelation. PCMCI_0^+ has slightly higher lagged true positives, but similarly to PC loses power for contemporaneous links. False positives are

well-controlled for PCMCI^+ due to the MCI tests while PC and PCMCI_0^+ show inflated rates for lagged links for high autocorrelation. These results for the skeleton phase translate into higher contemporaneous orientation recall for PCMCI^+ compared to PC and PCMCI_0^+ while both have similar high precision that increases with autocorrelation. Recall monotonously increases with autocorrelation only for PCMCI^+ while for very high autocorrelation PC and PCMCI_0^+ have very low recall. PCMCI^+ benefits from higher autocorrelation since more lagged triples involving autodependencies can be utilized in the collider and rule orientation phases. Slightly more unoriented links for PCMCI^+ are mainly due to slightly more contemporaneous false positives (still controlled below α_{PC}). PCMCI^+ and PCMCI_0^+ show almost no conflicts while PC's conflicts increase with autocorrelation until low power reduces them again. Finally, runtimes diverge strongly for higher autocorrelation with the runtime of PC also being much more variable (std. of graph estimate runtime indicated by errorbars). Figure 2C depicts results for nonlinear experiments and the GPDC CI tests for $N = 3$ and other parameters as before. The results are similar to the linear case but here the false positive inflation for PC is even more severe (PCMCI_0^+ not considered here, similar results as for ParCorr). In the Supplement we also show results for further N, T, α_{PC} that support these findings. In general, false positives become more severe for PC with small N . For larger N PC has even lower adjacency power and slightly fewer false positives, but the latter is mostly due to sparser models. For $N = 2$ and no autocorrelation there is no orientation recall for any method, as expected. The runtime difference between PC and PCMCI^+ becomes even more pronounced for larger N .

4.2 Large number of variables

Figure 2B shows results for varying number of variables N (on x -axis) for linear experiments and the ParCorr CI tests for $a = 0.9$, $T = 500$, $\alpha_{PC} = 0.01$, and $\tau_{\max} = 5$ in the majority-rule variants of PC, PCMCI^+ and PCMCI_0^+ . PCMCI^+ has robustly high detection power with a slight decrease for lagged links. PCMCI_0^+ has slightly more power for lagged links, but strongly decreasing power for contemporaneous links, and PC displays low power for both types of links. False positives are well controlled only for PCMCI^+ , while both PCMCI_0^+ and PC display false positives for small N where model connectivity is denser and false negatives are more likely leading to false positives. For high N PC has false positives only regarding autodependencies since the large number of tests removed many adjacencies and outweighs false positives due to ill-calibrated tests. Orientation precision is decreasing for

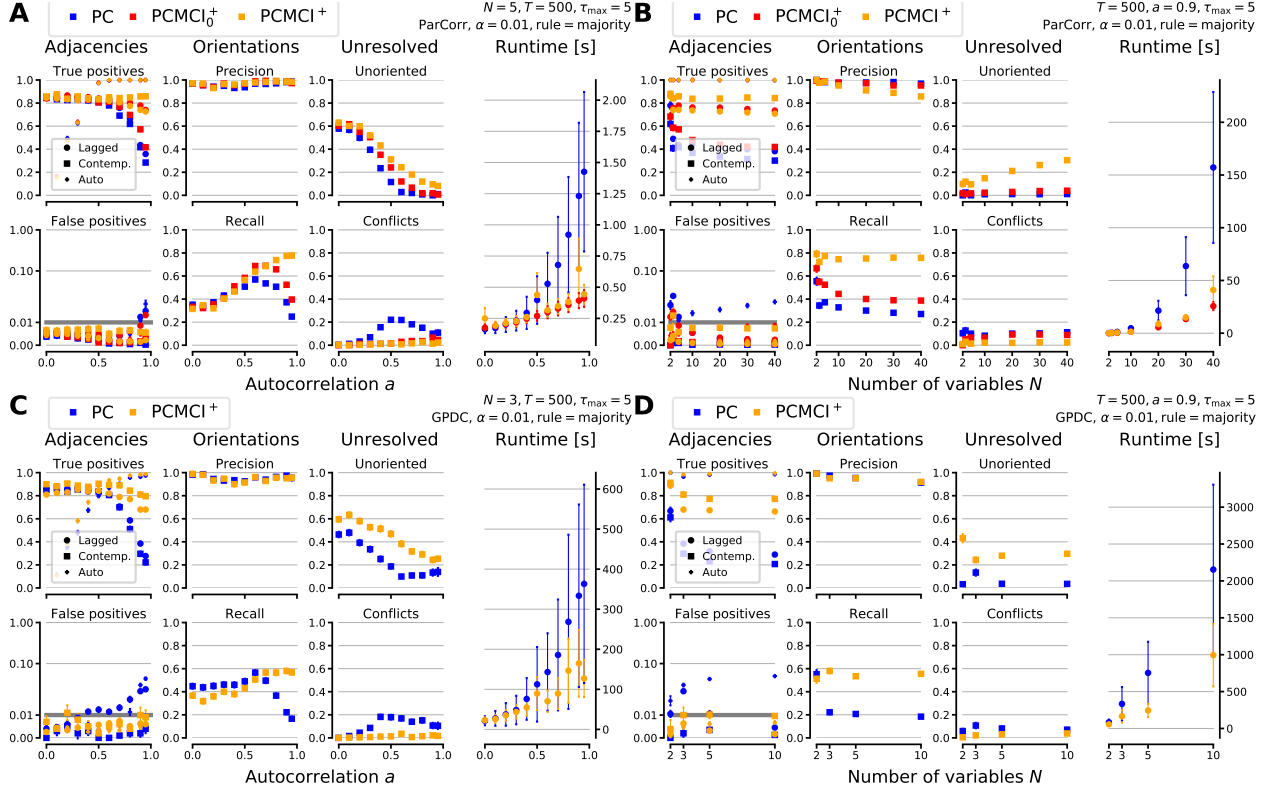


Figure 2: Results of numerical experiments. (A) Linear autocorrelation experiments for ParCorr test. (B) Linear high-dimensionality experiments for ParCorr test. (C,D) Same as above for nonlinear experiments with GPDC test.

all three methods with a higher decrease for PCMCI^+ , but PCMCI^+ has twice as much recall compared to PC and PCMCI_0^+ and is almost not affected by higher N . PCMCI^+ leaves more links unoriented (those are mainly contemporaneous false positives) for higher N , while the other two methods yield more conflicting information. Runtime is increasing at a much smaller rate for PCMCI^+ and PCMCI_0^+ compared to PC, which also has a very high runtime variability across the different model realizations. For $N = 40$ PC is on average 8 times slower than PCMCI^+ .

Figure 2D depicts results for nonlinear experiments and the GPDC CI tests. Here the loss in adjacency and orientation detection power with N is even more pronounced while precision is similarly high for PC and PCMCI^+ . Note that runtime for GPDC compared to ParCorr is orders of magnitude longer. These results are robust also for other T, α_{PC} (see Supplement). For low to no autocorrelation, where orientation recall is generally low, all methods perform almost identical.

In the Supplement we also show results for varying the maximum time lag τ_{max} for fixed N where we find that adjacency detections and orientation recall generally decrease with larger τ_{max} . Further, PC and PCMCI_0^+ show

inflated false positives for very small τ_{max} since then the effect of too many tests that counteract the ill-calibrated CI tests is diminished. Runtime scales approximately linearly with τ_{max} .

5 Conclusion

We have proposed a new conditional independence-based causal discovery algorithm for time series that yields much higher recall, well-controlled false positives, and faster runtime than the original PC algorithm in the case of highly autocorrelated time series, while maintaining the same performance in low autocorrelation settings. The algorithm well exploits sparsity in high-dimensional settings and can flexibly be combined with different conditional independence tests. PCMCI^+ is available as part of the *tigramite* Python package at <https://github.com/jakobrunge/tigramite>.

Acknowledgements

I thank Andreas Gerhardus for helpful comments.

References

- [Chickering, 2002] Chickering, D. M. (2002). Learning Equivalence Classes of Bayesian-Network Structures. *J. Mach. Learn. Res.*, 2:445–498.
- [Colombo and Maathuis, 2014] Colombo, D. and Maathuis, M. H. (2014). Order-Independent Constraint-Based Causal Structure Learning. *J. Mach. Learn. Res.*, 15:3921–3962.
- [Entner and Hoyer, 2010] Entner, D. and Hoyer, P. O. (2010). On causal discovery from time series data using FCI. In *Proc. Fifth Eur. Work. Probabilistic Graph. Model.*, pages 121–128.
- [Granger, 1969] Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438.
- [Malinsky and Spirtes, 2019] Malinsky, D. and Spirtes, P. (2019). Learning the structure of a nonstationary vector autoregression. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2986–2994.
- [Meek, 1995] Meek, C. (1995). Causal inference and causal explanation with background knowledge. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 403–410.
- [Moneta et al., 2011] Moneta, A., Chlaß, N., Entner, D., and Hoyer, P. (2011). Causal search in structural vector autoregressive models. In *NIPS Mini-Symposium on Causality in Time Series*, pages 95–114.
- [Peters et al., 2017] Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of causal inference: foundations and learning algorithms*. MIT Press, Cambridge, MA.
- [Ramsey et al., 2006] Ramsey, J., Spirtes, P., and Zhang, J. (2006). Adjacency-faithfulness and conservative causal inference. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 401–408.
- [Ramsey, 2014] Ramsey, J. D. (2014). A Scalable Conditional Independence Test for Nonlinear, Non-Gaussian Data. <https://arxiv.org/abs/1401.5031>.
- [Runge, 2018a] Runge, J. (2018a). Causal network reconstruction from time series: From theoretical assumptions to practical estimation. *Chaos An Interdiscip. J. Nonlinear Sci.*, 28(7):075310.
- [Runge, 2018b] Runge, J. (2018b). Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In Storkey, A. & Perez-Cruz, F., editor, *Proc. 21st Int. Conf. Artif. Intell. Stat.* Playa Blanca, Lanzarote, Canary Islands: PMLR.
- [Runge et al., 2019a] Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., Glymour, C., Kretschmer, M., Mahecha, M. D., Muñoz-Marí, J., van Nes, E. H., Peters, J., Quax, R., Reichstein, M., Scheffer, M., Schölkopf, B., Spirtes, P., Sugihara, G., Sun, J., Zhang, K., and Zscheischler, J. (2019a). Inferring causation from time series in earth system sciences. *Nature Communications*, 10(1):2553.
- [Runge et al., 2012] Runge, J., Heitzig, J., Marwan, N., and Kurths, J. (2012). Quantifying causal coupling strength: A lag-specific measure for multivariate time series related to transfer entropy. *Physical Review E*, 86(6):061121.
- [Runge et al., 2019b] Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., and Sejdinovic, D. (2019b). Detecting causal associations in large nonlinear time series datasets. *Science Advances*, eaau4996(5).
- [Sen et al., 2017] Sen, R., Suresh, A. T., Shanmugam, K., Dimakis, A. G., and Shakkottai, S. (2017). Model-Powered Conditional Independence Test. In *Proc. 30th Conf. Adv. Neural Inf. Process. Syst.*, pages 2955–2965.
- [Spirtes and Glymour, 1991] Spirtes, P. and Glymour, C. (1991). An Algorithm for Fast Recovery of Sparse Causal Graphs. *Soc. Sci. Comput. Rev.*, 9(1):62–72.
- [Spirtes et al., 2000] Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction, and Search*. MIT Press, Boston.
- [Spirtes and Zhang, 2016] Spirtes, P. and Zhang, K. (2016). Causal discovery and inference: concepts and recent methodological advances. *Appl. Informatics*, 3(1):3.
- [Zhang, 2008] Zhang, J. (2008). On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artif. Intell.*, 172(16-17):1873–1896.
- [Zhang et al., 2011] Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. (2011). Kernel-based Conditional Independence Test and Application in Causal

Discovery. In *Proc. 27th Conf. Uncertain. Artif. Intell.*, pages 804–813.

Supplementary Material

S1 Definitions

Definition 1 (Causal Sufficiency). A set \mathbf{X}_{t+1}^- of variables of a process is causally sufficient if and only if every common cause of any two or more variables in \mathbf{X}_{t+1}^- is in \mathbf{X}_{t+1}^- (or it is constant across time).

Definition 2 (Causal Markov Condition). The joint distribution of a process \mathbf{X} whose causal structure can be represented in a time series graph \mathcal{G} fulfills the Causal Markov Condition iff for all $X_t^j \in \mathbf{X}_t$ every non-descendant of X_t^j in \mathcal{G} is independent of X_t^j given the parents $\mathcal{P}(X_t^j)$. In particular, $\mathbf{X}_t^- \setminus \mathcal{P}(X_t^j) \perp\!\!\!\perp X_t^j \mid \mathcal{P}(X_t^j)$ since all variables in \mathbf{X}_t^- are non-descendants of X_t^j by time-order.

Definition 3 (Adjacency and standard faithfulness Condition). The joint distribution of a process \mathbf{X} whose causal structure can be represented in a time series graph \mathcal{G} fulfills the Adjacency Faithfulness Condition iff for all disjoint $X_{t-\tau}^i, X_t^j, \mathcal{S} \in \mathbf{X}_{t+1}^-$ with $\tau \geq 0$

$$\begin{aligned} X_{t-\tau}^i \perp\!\!\!\perp X_t^j \mid \mathcal{S} &\Rightarrow X_{t-\tau}^i \rightarrow X_t^j \notin \mathcal{G} \\ X_{t-\tau}^i \rightarrow X_t^j \in \mathcal{G} &\Rightarrow X_{t-\tau}^i \not\perp\!\!\!\perp X_t^j \mid \mathcal{S} \text{ (contrapositive)} \end{aligned}$$

Furthermore, the variables fulfill the (standard) Faithfulness Condition iff

$$\begin{aligned} X_{t-\tau}^i \perp\!\!\!\perp X_t^j \mid \mathcal{S} &\Rightarrow X_{t-\tau}^i \not\propto X_t^j \mid \mathcal{S} \\ X_{t-\tau}^i \not\perp\!\!\!\perp X_t^j \mid \mathcal{S} &\Rightarrow X_{t-\tau}^i \not\propto X_t^j \mid \mathcal{S} \text{ (contrapositive)} \end{aligned}$$

S2 Proofs

S2.1 Proof of Lemma 1

We first state the following Lemma: Algorithm 1 returns a superset of lagged parents under Assumptions 1, i.e., $\mathcal{P}_t^-(X_t^j) \subseteq \widehat{\mathcal{B}}_t^-(X_t^j)$ for all X_t^j in \mathbf{X}_t .

Proof. The lemma states $\mathcal{P}_t^-(X_t^j) \subseteq \widehat{\mathcal{B}}_t^-(X_t^j)$ for all X_t^j in \mathbf{X}_t under Assumptions 1. We need to show that for arbitrary $(X_{t-\tau}^i, X_t^j)$ with $\tau > 0$ we have $X_{t-\tau}^i \notin \widehat{\mathcal{B}}_t^-(X_t^j) \Rightarrow X_{t-\tau}^i \notin \mathcal{P}_t^-(X_t^j)$. Algorithm 1 removes $X_{t-\tau}^i$ from $\widehat{\mathcal{B}}_t^-(X_t^j)$ iff $X_{t-\tau}^i \perp\!\!\!\perp X_t^j \mid \mathcal{S}$ for some $\mathcal{S} \subseteq \widehat{\mathcal{B}}_t^-(X_t^j) \setminus \{X_{t-\tau}^i\}$ in the iterative CI tests. Then Adjacency Faithfulness directly implies that $X_{t-\tau}^i$ is not adjacent to X_t^j and in particular $X_{t-\tau}^i \notin \mathcal{P}_t^-(X_t^j)$. \square

With this step we can prove Lemma 1.

Proof. The lemma states that under Assumptions 1 with Adjacency Faithfulness replaced by standard Faithfulness Algorithm 1 for all $X_t^j \in \mathbf{X}_t$ $\widehat{\mathcal{B}}_t^-(X_t^j) =$

$\mathcal{P}_t^-(X_t^j, \mathcal{C}_t(X_t^j))$ where $\mathcal{C}_t(X_t^j)$ denotes the contemporaneous ancestors of X_t^j . We need to show that for arbitrary $X_{t-\tau}^i, X_t^j \in \mathbf{X}_{t+1}^-$ with $\tau > 0$: (1) $X_{t-\tau}^i \notin \widehat{\mathcal{B}}_t^-(X_t^j) \Rightarrow X_{t-\tau}^i \notin \mathcal{P}_t^-(X_t^j, \mathcal{C}_t(X_t^j))$ and (2) $X_{t-\tau}^i \in \widehat{\mathcal{B}}_t^-(X_t^j) \Rightarrow X_{t-\tau}^i \in \mathcal{P}_t^-(X_t^j, \mathcal{C}_t(X_t^j))$.

Ad 1) Algorithm 1 removes $X_{t-\tau}^i$ from $\widehat{\mathcal{B}}_t^-(X_t^j)$ iff $X_{t-\tau}^i \perp\!\!\!\perp X_t^j \mid \mathcal{S}$ for some $\mathcal{S} \subseteq \widehat{\mathcal{B}}_t^-(X_t^j) \setminus \{X_{t-\tau}^i\}$ in the iterative CI tests. Then standard Faithfulness implies that $X_{t-\tau}^i \propto X_t^j \mid \mathcal{S}$ and in particular $X_{t-\tau}^i \notin \mathcal{P}_t^-(X_t^j)$, as proven already in Lemma 1 under the weaker Adjacency Faithfulness Condition. To show that $X_{t-\tau}^i \notin \mathcal{P}_t^-(\mathcal{C}_t(X_t^j))$ we note that $\mathcal{S} \subseteq \widehat{\mathcal{B}}_t^-(X_t^j) \setminus \{X_{t-\tau}^i\}$ does not include any contemporaneous conditions and, hence, all contemporaneous directed paths from contemporaneous ancestors of X_t^j are open and also paths from parents of those ancestors are open. If $X_{t-\tau}^i \in \mathcal{P}_t^-(\mathcal{C}_t(X_t^j))$, by the contraposition of standard Faithfulness we should observe $X_{t-\tau}^i \not\perp\!\!\!\perp X_t^j \mid \mathcal{S}$. Then the fact that on the contrary we observe $X_{t-\tau}^i \perp\!\!\!\perp X_t^j \mid \mathcal{S}$ implies that $X_{t-\tau}^i \notin \mathcal{P}_t^-(\mathcal{C}_t(X_t^j))$.

Ad 2) Now we have $X_{t-\tau}^i \in \widehat{\mathcal{B}}_t^-(X_t^j)$ which implies that $X_{t-\tau}^i \not\perp\!\!\!\perp X_t^j \mid \widehat{\mathcal{B}}_t^-(X_t^j) \setminus \{X_{t-\tau}^i\}$ in last iteration step of Algorithm 1. By (1), $\widehat{\mathcal{B}}_t^-(X_t^j)$ is a superset of $\mathcal{P}_t^-(X_t^j, \mathcal{C}_t(X_t^j))$. Define the lagged extra conditions as $W_t^- = \widehat{\mathcal{B}}_t^-(X_t^j) \setminus \{\mathcal{P}_t^-(X_t^j, \mathcal{C}_t(X_t^j)), X_{t-\tau}^i\}$. Since W_t^- is lagged, it is a non-descendant of X_t^j or any $X_t^k \in \mathcal{C}_t(X_t^j)$. We now proceed by a proof by contradiction. Suppose to the contrary that $X_{t-\tau}^i \notin \mathcal{P}_t^-(X_t^j, \mathcal{C}_t(X_t^j))$. The Causal Markov Condition applies to both $X_{t-\tau}^i$ and W_t^- and implies that $(X_{t-\tau}^i, W_t^-) \perp\!\!\!\perp X_t^j \mid \mathcal{P}_t^-(X_t^j, \mathcal{C}_t(X_t^j))$. From the weak union property of conditional independence we get $X_{t-\tau}^i \perp\!\!\!\perp X_t^j \mid \mathcal{P}_t^-(X_t^j, \mathcal{C}_t(X_t^j)), W_t^-$ which is equivalent to $X_{t-\tau}^i \perp\!\!\!\perp X_t^j \mid \widehat{\mathcal{B}}_t^-(X_t^j) \setminus \{X_{t-\tau}^i\}$, contrary to the assumption, hence $X_{t-\tau}^i \in \mathcal{P}_t^-(X_t^j, \mathcal{C}_t(X_t^j))$. \square

S2.2 Proof of Theorem 1

Proof. The theorem states that under Assumptions 1 $\widehat{\mathcal{G}}^* = \mathcal{G}^*$, where the $*$ denotes the skeleton of the time series graph. We denote the two types of skeleton links \rightarrow and $\circ\!\!\!\circ$ here generically as \star and can assume $\tau \geq 0$. We need to show that for arbitrary $X_{t-\tau}^i, X_t^j \in \mathbf{X}_{t+1}^-$: (1) $X_{t-\tau}^i \star X_t^j \notin \widehat{\mathcal{G}}^* \Rightarrow X_{t-\tau}^i \star X_t^j \notin \mathcal{G}^*$ and (2) $X_{t-\tau}^i \star X_t^j \in \mathcal{G}^* \Rightarrow X_{t-\tau}^i \star X_t^j \in \widehat{\mathcal{G}}^*$.

Ad (1): Algorithm 2 deletes a link $X_{t-\tau}^i \star X_t^j$ from $\widehat{\mathcal{G}}^*$ iff $X_{t-\tau}^i \perp\!\!\!\perp X_t^j \mid \mathcal{S}, \widehat{\mathcal{B}}_t^-(X_t^j) \setminus \{X_{t-\tau}^i\}, \widehat{\mathcal{B}}_{t-\tau}^-(X_{t-\tau}^i)$ for some $\mathcal{S} \subseteq \widehat{\mathcal{A}}_t(X_t^j)$ in the iterative CI tests with $\widehat{\mathcal{B}}_t^-(X_t^j)$ estimated in Algorithm 1. $\widehat{\mathcal{A}}_t(X_t^j)$ denotes the

contemporaneous adjacencies. Then Adjacency Faithfulness directly implies that $X_{t-\tau}^i$ is not adjacent to X_t^j : $X_{t-\tau}^i \star \star X_t^j \notin \mathcal{G}^*$.

Ad (2): By Lemma 1 we know that $\widehat{\mathcal{B}}_t^-(X_t^j)$ is a superset of the lagged parents of X_t^j . Denote the lagged, extra conditions occurring in the CI tests of Algorithm 2 as $W_t^- = (\widehat{\mathcal{B}}_t^-(X_t^j) \setminus \{X_{t-\tau}^i\}, \widehat{\mathcal{B}}_{t-\tau}^-(X_{t-\tau}^i)) \setminus \mathcal{P}(X_t^j)$. W_t^- does not contain parents of X_t^j and by the assumption also $X_{t-\tau}^i$ is not a parent of X_t^j . We further assume that for $\tau = 0$ X_t^i is also not a descendant of X_t^j since that case is covered if we exchange X_t^i and X_t^j . Then the Causal Markov Condition implies $(X_{t-\tau}^i, W_t^-) \perp\!\!\!\perp X_t^j \mid \mathcal{P}(X_t^j)$. By the weak union property of conditional independence this leads to $X_{t-\tau}^i \perp\!\!\!\perp X_t^j \mid \mathcal{P}(X_t^j), W_t^-$ which is equivalent to $X_{t-\tau}^i \perp\!\!\!\perp X_t^j \mid \mathcal{P}(X_t^j), \widehat{\mathcal{B}}_t^-(X_t^j) \setminus \{X_{t-\tau}^i\}, \widehat{\mathcal{B}}_{t-\tau}^-(X_{t-\tau}^i)$. Now Algorithm 2 iteratively tests $X_{t-\tau}^i \perp\!\!\!\perp X_t^j \mid \mathcal{S}, \widehat{\mathcal{B}}_t^-(X_t^j) \setminus \{X_{t-\tau}^i\}, \widehat{\mathcal{B}}_{t-\tau}^-(X_{t-\tau}^i)$ for all $\mathcal{S} \subseteq \widehat{\mathcal{A}}_t(X_t^j)$. By the first part of this proof, the estimated contemporaneous adjacencies are always a superset of the true contemporaneous adjacencies, i.e., $\mathcal{A}_t(X_t^j) \subseteq \widehat{\mathcal{A}}_t(X_t^j)$, and by Lemma 1 $\widehat{\mathcal{B}}_t^-(X_t^j)$ is a superset of the lagged parents. Hence, at some iteration step $\mathcal{S} = \mathcal{P}_t(X_t^j)$ and Algorithm 2 will find $X_{t-\tau}^i \perp\!\!\!\perp X_t^j \mid \mathcal{P}(X_t^j), \widehat{\mathcal{B}}_t^-(X_t^j) \setminus \{X_{t-\tau}^i\}, \widehat{\mathcal{B}}_{t-\tau}^-(X_{t-\tau}^i)$ and remove $X_{t-\tau}^i \star \star X_t^j$ from $\widehat{\mathcal{G}}^*$. \square

For empty conditioning sets \mathcal{S} ($p = 0$), Alg. 2 is equivalent to the MCI algorithm [Runge et al., 2019b] with the slight change that the latter is initialized with a fully connected (lagged) graph, which has no effect asymptotically. In [Runge et al., 2019b] the authors prove the consistency of PCMCi assuming no contemporaneous causal links under the standard Faithfulness Condition. The proof above implies that PCMCi is already consistent under the weaker Adjacency Faithfulness Condition.

S2.3 Proof of Lemma 2

Proof. Time order and stationarity can be used to constrain the four cases as follows. Let us first consider a generic triple $X_{t_i}^i \star \star X_{t_k}^k \star \star X_{t_j}^j$. By stationarity we can fix $t = t_j$, and we consider only cases with $t_i, t_k \leq t$.

The possible triples in the collider phase of the original PC algorithm are $X_{t_i}^i \star \star X_{t_k}^k \star \star X_t^j$ where $(X_{t_i}^i, X_t^j)$ are not adjacent. For $t_k < t$ the time-order constraint automatically orients $X_{t_k}^k \rightarrow X_t^j$ and hence $X_{t_k}^k$ is a parent of X_t^j and must always be in the separating set that makes $X_{t_i}^i$ and X_t^j independent. Hence we only need to consider $t_k = t$ and can set $\tau = t - t_i$, leaving the two cases of unshielded triples $X_{t-\tau}^i \rightarrow X_{t-\tau}^k \circ \circ X_t^j$ (for $\tau > 0$) or $X_{t-\tau}^i \circ \circ X_{t-\tau}^k \circ \circ X_t^j$ (for $\tau = 0$) in \mathcal{G} where $(X_{t-\tau}^i, X_t^j)$ are

not adjacent. Since $X_{t-\tau}^k$ is contemporaneous to X_t^j , this restriction implies that only contemporaneous separating sets are relevant for the collider orientation phase.

For rule R1 in the orientation phase the original PC algorithm considers the remaining triples with $X_{t-\tau}^i \rightarrow X_t^j$ that were not oriented by the collider phase (or by time order). This leaves $X_{t-\tau}^i \rightarrow X_{t-\tau}^k \circ \circ X_t^j$ where $\tau \geq 0$.

For rule R2 the original PC algorithm considers $X_{t_i}^i \rightarrow X_{t_k}^k \rightarrow X_t^j$ with $X_{t_i}^i \circ \circ X_t^j$. The latter type of link leads to $t_i = t$ and time order restricts the triples to $X_t^i \rightarrow X_t^k \rightarrow X_t^j$ with $X_t^i \circ \circ X_t^j$.

For rule R3 the original PC algorithm considers $X_{t_i}^i \circ \circ X_{t_k}^k \rightarrow X_t^j$ and $X_{t_i}^i \circ \circ X_{t_l}^l \rightarrow X_t^j$ where $(X_{t_k}^k, X_{t_l}^l)$ are not adjacent and $X_{t_i}^i \circ \circ X_t^j$. The latter constraint leads to $t_i = t$ and $X_{t_i}^i \circ \circ X_{t_k}^k$ and $X_{t_i}^i \circ \circ X_{t_l}^l$ imply $t_k = t_l = t$. Hence we only need to check triples $X_t^i \circ \circ X_t^k \rightarrow X_t^j$ and $X_t^i \circ \circ X_t^l \rightarrow X_t^j$ where (X_t^i, X_t^l) are not adjacent and $X_t^i \circ \circ X_t^j$. \square

S2.4 Proof of Theorem 2

Proof. We first consider the case under Assumptions 1 with Adjacency Faithfulness and PCMCi⁺ in conjunction with the conservative collider orientation rule in Algorithm S3. We need to show that all separating sets estimated in Algorithm S3 during the conservative orientation rule are correct. From the soundness (Theorem 1) and correctness of the separating sets follows the correctness of the collider orientation phase and the rule orientation phase which implies the completeness.

By Lemma 2 we only need to prove that in Algorithm S3 for unshielded triples $X_{t-\tau}^i \rightarrow X_{t-\tau}^k \circ \circ X_t^j$ (for $\tau > 0$) or $X_{t-\tau}^i \circ \circ X_{t-\tau}^k \circ \circ X_t^j$ (for $\tau = 0$) the separating sets among subsets of contemporaneous neighbors of X_t^j and, if $\tau = 0$, of $X_{t_i}^i$, are correct. Algorithm S3 tests $X_{t-\tau}^i \perp\!\!\!\perp X_t^j \mid \mathcal{S}, \widehat{\mathcal{B}}_t^-(X_t^j) \setminus \{X_{t-\tau}^i\}, \widehat{\mathcal{B}}_{t-\tau}^-(X_{t-\tau}^i)$ for all $\mathcal{S} \subseteq \widehat{\mathcal{A}}_t(X_t^j) \setminus \{X_{t-\tau}^i\}$ and for all $\mathcal{S} \subseteq \widehat{\mathcal{A}}_t(X_t^i) \setminus \{X_t^j\}$ (if $\tau = 0$). Since PCMCi⁺ is sound, all adjacency information is correct and since all CI tests are assumed correct, all information on contemporaneous separating sets is valid. Furthermore, with the conservative rule those triples where only Adjacency Faithfulness, but not standard Faithfulness holds will be correctly marked as ambiguous triples.

Under standard Faithfulness the completeness requires to prove that PCMCi⁺ without the conservative orientation rule yields correct subset information. By Lemma 2 also here we need to consider only separating sets among subsets of contemporaneous neighbors of X_t^j . Algorithm 2 tests $X_{t-\tau}^i \perp\!\!\!\perp X_t^j \mid \mathcal{S}, \widehat{\mathcal{B}}_t^-(X_t^j) \setminus \{X_{t-\tau}^i\}, \widehat{\mathcal{B}}_{t-\tau}^-(X_{t-\tau}^i)$

for all $S \subseteq \widehat{\mathcal{A}}_t(X_t^j) \setminus \{X_{t-\tau}^i\}$. And again, since PCMC⁺ is sound, all adjacency information is correct and since all CI tests are assumed correct, all information on contemporaneous separating sets is valid, from which the completeness for this case follows. \square

S2.5 Proof of Theorem 3

Proof. Order-independence follows straightforwardly from sticking to PC algorithm version in [Colombo and Maathuis, 2014]. In particular, Algorithm 1 and Algorithm 2 are order-independent since they are based on PC stable where adjacencies are removed only after each loop over conditions of cardinality p . Furthermore, the collider phase (Algorithm S3) and rule orientation phase (Algorithm S4) are order-independent by marking triples with inconsistent separating sets as ambiguous and consistently marking conflicting link orientations by the introduction of bi-directed links \leftrightarrow . \square

S2.6 Proof of Theorem 4

Proof. The proof states that under Assumptions 1 the effect size for the PCMC⁺ CI tests in Algorithm 2 for $p = 0$ for contemporaneous true links $X_t^i \rightarrow X_t^j \in \mathcal{G}$ of PCMC⁺ is greater than that of PCMC₀⁺: $I(X_t^i; X_t^j | \widehat{\mathcal{B}}_t^-(X_t^j), \widehat{\mathcal{B}}_t^-(X_t^i)) > \min(I(X_t^i; X_t^j | \widehat{\mathcal{B}}_t^-(X_t^j)), I(X_t^i; X_t^j | \widehat{\mathcal{B}}_t^-(X_t^i)))$. We will use an information-theoretic framework here and consider the conditional mutual information.

To prove this statement, we denote by $\mathcal{B}_i = \widehat{\mathcal{B}}_t^-(X_t^i) \setminus \widehat{\mathcal{B}}_t^-(X_t^j)$ the lagged conditions of X_t^i that are not already contained in those of X_t^j and, correspondingly, $\mathcal{B}_j = \widehat{\mathcal{B}}_t^-(X_t^j) \setminus \widehat{\mathcal{B}}_t^-(X_t^i)$. Further, we denote the common lagged conditions as $\mathcal{B}_{ij} = \widehat{\mathcal{B}}_t^-(X_t^j) \cap \widehat{\mathcal{B}}_t^-(X_t^i)$ and make use of the following conditional independencies, which hold by the Markov assumption: (1) $\mathcal{B}_i \perp\!\!\!\perp X_t^j | \mathcal{B}_j, \mathcal{B}_{ij}, X_t^i$ and (2) $\mathcal{B}_j \perp\!\!\!\perp X_t^i | \mathcal{B}_i, \mathcal{B}_{ij}$. We first prove that, given a contemporaneous true link $X_t^i \rightarrow X_t^j \in \mathcal{G}$, $I(X_t^i; X_t^j | \mathcal{B}_{ij}, \mathcal{B}_j) > I(X_t^i; X_t^j | \mathcal{B}_{ij}, \mathcal{B}_i)$ by using the following two ways to apply the chain rule of conditional mutual information:

$$\begin{aligned} I(X_t^i; \mathcal{B}_i; X_t^j, \mathcal{B}_j | \mathcal{B}_{ij}) &= \\ &= I(X_t^i; \mathcal{B}_i; \mathcal{B}_j | \mathcal{B}_{ij}) + I(X_t^i; \mathcal{B}_i; X_t^j | \mathcal{B}_{ij}, \mathcal{B}_j) \\ &= I(\mathcal{B}_i; \mathcal{B}_j | \mathcal{B}_{ij}) + \underbrace{I(X_t^i; \mathcal{B}_j | \mathcal{B}_{ij}, \mathcal{B}_i)}_{=0 \text{ (Markov)}} \\ &\quad + I(X_t^i; X_t^j | \mathcal{B}_{ij}, \mathcal{B}_j) + \underbrace{I(\mathcal{B}_i; X_t^j | \mathcal{B}_{ij}, \mathcal{B}_j, X_t^i)}_{=0 \text{ (Markov)}} \end{aligned} \quad (\text{S1})$$

and

$$\begin{aligned} I(X_t^i; \mathcal{B}_i; X_t^j, \mathcal{B}_j | \mathcal{B}_{ij}) &= \\ &= I(\mathcal{B}_i; X_t^j, \mathcal{B}_j | \mathcal{B}_{ij}) + I(X_t^i; X_t^j, \mathcal{B}_j | \mathcal{B}_{ij}, \mathcal{B}_i) \\ &= I(\mathcal{B}_i; \mathcal{B}_j | \mathcal{B}_{ij}) + \underbrace{I(\mathcal{B}_i; X_t^j | \mathcal{B}_{ij}, \mathcal{B}_j)}_{>0 \text{ if } X_t^i \rightarrow X_t^j} \\ &\quad + I(X_t^i; X_t^j | \mathcal{B}_{ij}, \mathcal{B}_i) + \underbrace{I(X_t^i; \mathcal{B}_j | \mathcal{B}_{ij}, \mathcal{B}_i, X_t^j)}_{>0 \text{ if } X_t^i \rightarrow X_t^j} \end{aligned} \quad (\text{S2})$$

where (S1) and (S2) denote two different applications of the chain rule. From this it follows that $I(X_t^i; X_t^j | \mathcal{B}_{ij}, \mathcal{B}_j) > I(X_t^i; X_t^j | \mathcal{B}_{ij}, \mathcal{B}_i)$.

Hence, it remains to prove that $I(X_t^i; X_t^j | \mathcal{B}_{ij}, \mathcal{B}_j, \mathcal{B}_i) > I(X_t^i; X_t^j | \mathcal{B}_{ij}, \mathcal{B}_i)$, which we also do by the chain rule:

$$\begin{aligned} I(X_t^i; X_t^j, \mathcal{B}_j | \mathcal{B}_{ij}, \mathcal{B}_i) &= \\ &= I(X_t^i; X_t^j | \mathcal{B}_{ij}, \mathcal{B}_i) + \underbrace{I(X_t^i; \mathcal{B}_j | \mathcal{B}_{ij}, \mathcal{B}_i, X_t^j)}_{>0 \text{ if } X_t^i \rightarrow X_t^j} \quad (\text{S3}) \\ &= \underbrace{I(X_t^i; \mathcal{B}_j | \mathcal{B}_{ij}, \mathcal{B}_i)}_{=0 \text{ (Markov)}} + I(X_t^i; X_t^j | \mathcal{B}_{ij}, \mathcal{B}_i, \mathcal{B}_j) \quad (\text{S4}) \end{aligned}$$

\square

S3 Further pseudo code

Algorithms S3 and S4 details the pseudo code for the PCMC⁺ / PCMC₀⁺ / PC collider phase with different collider rules and the orientation phase.

S4 Further numerical experiments

In Fig. S1 we show a version of numerical experiments with no cross-links and just autocorrelation to illustrate that PCMC⁺ has ill-calibrated CI tests for large autocorrelation.

The remaining appended pages contain results of further numerical experiments that evaluate different $N, T, \alpha_{\text{PC}}, \tau_{\text{max}}$ for the ParCorr and GPDC CI tests as indicated in the descriptors of the subpanels.

Algorithm S3 (Detailed PCMCI⁺ / PCMCI₀⁺ / PC collider phase with different collider rules)

Require: \mathcal{G} and sepset from Algorithm 2, rule = {'none', 'conservative', 'majority'}, time series dataset $\mathbf{X} = (X^1, \dots, X^N)$, significance threshold α_{PC} , $CI(X, Y, \mathbf{Z})$, PCMCI⁺ / PCMCI₀⁺: $\hat{\mathcal{B}}_t^-(X_t^j)$ for all X_t^j in \mathbf{X}_t

- 1: **for all** unshielded triples $X_{t-\tau}^i \rightarrow X_t^k \circ \circ X_t^j$ ($\tau > 0$) or $X_{t-\tau}^i \circ \circ X_t^k \circ \circ X_t^j$ ($\tau = 0$) in \mathcal{G} where $(X_{t-\tau}^i, X_t^j)$ are not adjacent **do**
- 2: **if** rule = 'none' **then**
- 3: **if** X_t^k is not in sepset($X_{t-\tau}^i, X_t^j$) **then**
- 4: Orient $X_{t-\tau}^i \rightarrow X_t^k \circ \circ X_t^j$ ($\tau > 0$) or $X_{t-\tau}^i \circ \circ X_t^k \circ \circ X_t^j$ ($\tau = 0$) as $X_{t-\tau}^i \rightarrow X_t^k \leftarrow X_t^j$
- 5: **else**
- 6: PCMCI⁺ / PCMCI₀⁺: Define contemporaneous adjacencies $\hat{\mathcal{A}}(X_t^j) = \hat{\mathcal{A}}_t(X_t^j) = \{X_t^i \neq X_t^j \in \mathbf{X}_t : X_t^i \circ \circ X_t^j \text{ in } \mathcal{G}\}$
- 7: PC: Define full adjacencies $\hat{\mathcal{A}}(X_t^j)$ for all (lagged and contemporaneous) links in \mathcal{G}
- 8: **for all** for all $\mathcal{S} \subseteq \hat{\mathcal{A}}(X_t^j) \setminus \{X_{t-\tau}^i\}$ and for all $\mathcal{S} \subseteq \hat{\mathcal{A}}(X_t^i) \setminus \{X_t^j\}$ (if $\tau = 0$) **do**
- 9: Evaluate $CI(X_{t-\tau}^i, X_t^j, \mathbf{Z})$ with
- 9: PCMCI⁺: $\mathbf{Z} = (\mathcal{S}, \hat{\mathcal{B}}_t^-(X_t^j) \setminus \{X_{t-\tau}^i\}, \hat{\mathcal{B}}_{t-\tau}^-(X_{t-\tau}^i))$
- 9: PCMCI₀⁺: $\mathbf{Z} = (\mathcal{S}, \hat{\mathcal{B}}_t^-(X_t^j) \setminus \{X_{t-\tau}^i\})$
- 9: PC: $\mathbf{Z} = \mathcal{S}$
- 10: Store all subsets \mathcal{S} with p -value $> \alpha_{PC}$ as separating subsets
- 11: **if** no separating subsets are found **then**
- 12: Mark triple as ambiguous
- 13: **else**
- 14: Compute fraction n_k of separating subsets that contain X_t^k
- 15: **if** rule = 'conservative' **then**
- 16: Orient triple as collider if $n_k=0$, leave unoriented if $n_k=1$, and mark as ambiguous if $0 < n_k < 1$
- 17: **else if** rule = 'majority' **then**
- 18: Orient triple as collider if $n_k < 0.5$, leave unoriented if $n_k > 0.5$, and mark as ambiguous if $n_k = 0.5$
- 19: Mark links in \mathcal{G} with conflicting orientations as bi-directed \leftrightarrow
- 20: **return** \mathcal{G} , sepset, ambiguous triples, conflicting links

Algorithm S4 (Detailed PCMCI⁺ / PCMCI₀⁺ / PC rule orientation phase)

Require: \mathcal{G} , ambiguous triples, conflicting links

- 1: **while** any unambiguous triples suitable for rules R1-R3 are remaining **do**
- 2: Apply rule R1 (orient unshielded triples that are not colliders):
- 3: **for all** unambiguous triples $X_{t-\tau}^i \rightarrow X_t^k \circ \circ X_t^j$ where $(X_{t-\tau}^i, X_t^j)$ are not adjacent **do**
- 4: Orient as $X_{t-\tau}^i \rightarrow X_t^k \rightarrow X_t^j$
- 5: Mark links with conflicting orientations as bi-directed \leftrightarrow
- 6: Apply rule R2 (avoid cycles):
- 7: **for all** unambiguous triples $X_t^i \rightarrow X_t^k \rightarrow X_t^j$ with $X_t^i \circ \circ X_t^j$ **do**
- 8: Orient as $X_t^i \rightarrow X_t^j$
- 9: Mark links with conflicting orientations as bi-directed \leftrightarrow
- 10: Apply rule R3 (orient unshielded triples that are not colliders and avoid cycles):
- 11: **for all** pairs of unambiguous triples $X_t^i \circ \circ X_t^k \rightarrow X_t^j$ and $X_t^i \circ \circ X_t^l \rightarrow X_t^j$ where (X_t^k, X_t^l) are not adjacent and $X_t^i \circ \circ X_t^j$ **do**
- 12: Orient as $X_t^i \rightarrow X_t^j$
- 13: Mark links with conflicting orientations as bi-directed \leftrightarrow
- 14: **return** \mathcal{G} , conflicting links

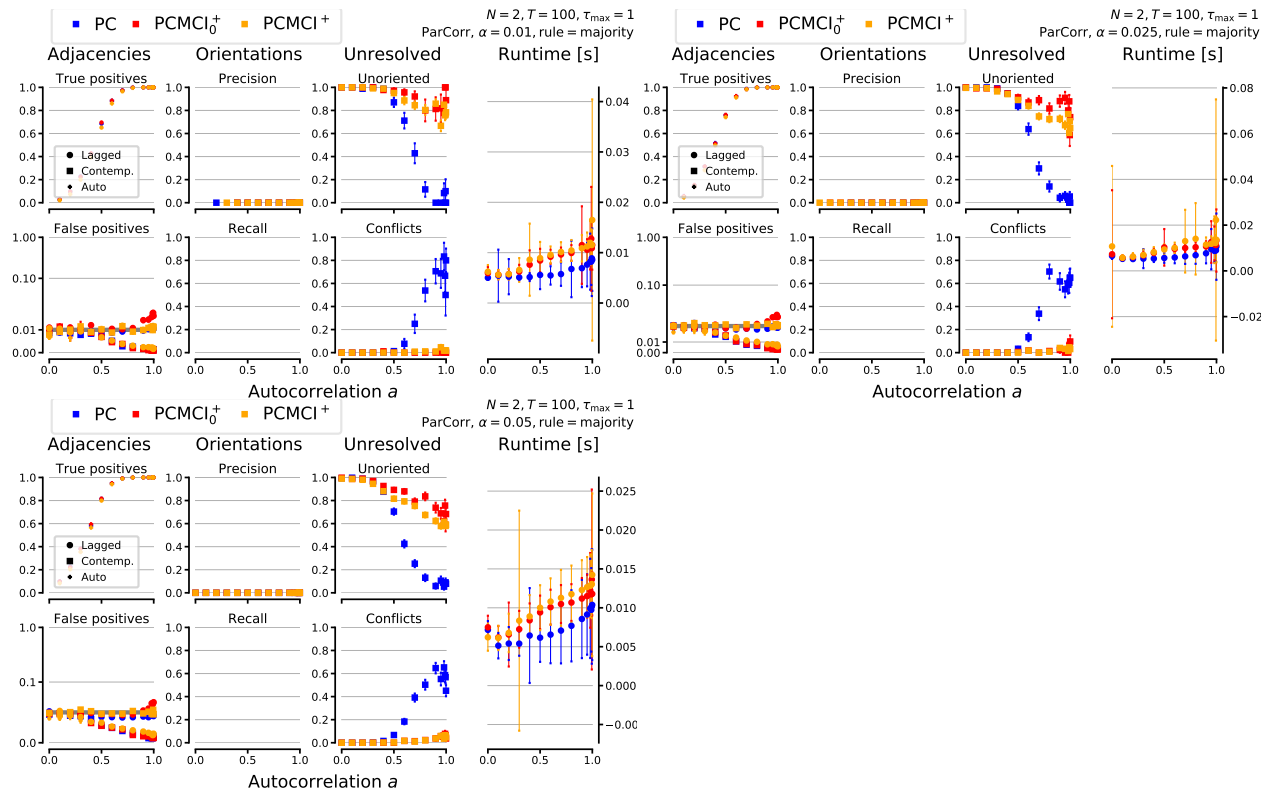


Figure S1: Results of numerical experiments for independent variables with autocorrelation.