Proceedings of Machine Learning Research vol 236:979–1007, 2024 3rd Conference on Causal Learning and Reasoning

Bootstrap aggregation and confidence measures to improve time series causal discovery

Kevin Debeire KEVIN.DEBEIRE@DLR.DE
Deutsches Zentrum für Luft- und Raumfahrt (DLR), Institut für Physik der Atmosphäre, Oberpfaffenhofen, Germany.
Deutsches Zentrum für Luft- und Raumfahrt (DLR), Institut für Datenwissenschaften, Jena, Germany.

Andreas Gerhardus* ANDREAS.GERHARDUS@DLR.DE
Deutsches Zentrum für Luft- und Raumfahrt (DLR), Institut für Datenwissenschaften, Jena, Germany.

Jakob Runge* JAKOB.RUNGE@DLR.DE
Deutsches Zentrum für Luft- und Raumfahrt (DLR), Institut für Datenwissenschaften, Jena, Germany.
Technische Universität Berlin, Faculty of Computer Science, Berlin, Germany.

Veronika Eyring VERONIKA.EYRING@DLR.DE
Deutsches Zentrum für Luft- und Raumfahrt (DLR), Institut für Physik der Atmosphäre, Oberpfaffenhofen, Germany.
University of Bremen, Institute of Environmental Physics (IUP), Bremen, Germany.

**These corresponding authors contributed equally.*

Editors: Francesco Locatello and Vanessa Didelez

Abstract

Learning causal graphs from multivariate time series is an ubiquitous challenge in all application domains dealing with time-dependent systems, such as in Earth sciences, biology, or engineering, to name a few. Recent developments for this causal discovery learning task have shown considerable skill, notably the specific time-series adaptations of the popular conditional independence-based learning framework. However, uncertainty estimation is challenging for conditional independence-based methods. Here, we introduce a novel bootstrap approach designed for time series causal discovery that preserves the temporal dependencies and lag-structure. It can be combined with a range of time series causal discovery methods and provides a measure of confidence for the links of the time series graphs. Furthermore, next to confidence estimation, an aggregation, also called bagging, of the bootstrapped graphs by majority voting results in bagged causal discovery methods. In this work, we combine this approach with the state-of-the-art conditional-independence-based algorithm PCMCI+. With extensive numerical experiments we empirically demonstrate that, in addition to providing confidence measures for links, Bagged-PCMCI+ improves in precision and recall as compared to its base algorithm PCMCI+, at the cost of higher computational demands. These statistical performance improvements are especially pronounced in the more challenging settings (short time sample size, large number of variables, high autocorrelation). Our bootstrap approach can also be combined with other time series causal discovery algorithms and can be of considerable use in many real-world applications.

Keywords: Causal discovery, Bootstrap aggregation, time series, confidence estimation

1. Introduction

Since a rigorous mathematical framework for causal inference has been established in the seminal works of Pearl, Spirtes, Glymour, Scheines, and Rubin (Imbens and Rubin, 2015; Pearl, 2009; Spirtes et al., 2000; Rubin, 1974), causal inference has undergone continuous developments to address the challenges of real-world problem settings. Learning causal graphs from data (termed causal discovery) is a main pillar of causal inference and of high interest in many fields where not even qualitative causal knowledge in the form of graphs is available, such as in biology (Friedman et al., 2000), neuroscience (Kaminski et al., 2001), or Earth system sciences (Ebert-Uphoff and Deng, 2012; Runge et al., 2019a; Kretschmer et al., 2016; Galytska et al., 2023; Karmouche et al., 2023). Furthermore, data in these fields typically comes in the form of time series, constituting a more general problem setting than the standard *i.i.d.*-case. In (Runge et al., 2019a), the authors give an overview of a few main categories of **time series causal discovery** methods: Granger causality and its extensions (Granger, 2001), nonlinear state-space methods (CCM (Sugihara et al., 2012)), causal network learning algorithms (for example, naive adaptations of the PC-algorithm (Spirtes and Glymour, 1991), PCMCI and its extensions PCMCI+ (Runge et al., 2019b; Runge, 2020; Gerhardus and Runge, 2020), FCI (Spirtes et al., 2000; Zhang, 2008)), Bayesian score-based approaches (Chickering, 1996), and the structural causal model framework (e.g., VarLiNGAM (Hyvärinen et al., 2010)). While the state-of-the-art time series causal discovery methods have seen considerable improvements over the years, this is still a very challenging task. A particular drawback in practical applications is that most methods output single graphs without the option to assess the uncertainty or confidence in the causal links. Approaches in this direction are based on taking the absolute p-value over all conditioning sets (Strobl et al., 2016), which presents a conservative option and does not permit to assess the uncertainty in orientations.

Independent of the development of causal discovery methods, Breiman (1996) introduced **bootstrap aggregation** (bagging) which has initially been used to improve the accuracy and stability of machine learning algorithms. In bagging, a random sample in the training set is selected with replacement —meaning that each data point can be drawn more than once. Several data samples are generated in this fashion to produce a set of replicates (also called resamples). The machine learning models are then trained independently on each replicate and, finally, the outputs are averaged for prediction tasks or aggregated for classification tasks (for example by majority voting). Since its introduction, bagging has been extensively used in combination with other machine learning algorithms (Dietterich, 2000; Galar et al., 2012; Ganaie et al., 2022).

Combining bagging and causal graphical model algorithms has been proposed to improve the stability of graphical model learning (Meinshausen and Buehlmann, 2008; Li et al., 2011), as the estimation of graphical models is relatively sensitive to small changes of the original data. For example, Wang and Peng (2014) introduce an aggregation approach of directed acyclic graphs (DAGs) by minimizing the overall distance of the aggregated graph (based on structural hamming distance) to the ensemble of DAGs. Bootstrapping a causal discovery algorithm and returning a summary graph constructed by a voting scheme is a feature of the TETRAD project (Ramsey et al., 2018). Guo et al. (2021) propose a two-phase causality ensemble framework to combine results from different data partitions (instead of bootstrap samples) and causal discovery algorithms into a single output graph using majority voting. In addition, the idea of measuring the uncertainty or confidence for an edge of an estimated graph from the edge frequency based on the graphs learned on bootstrap samples has been suggested in Friedman et al. (1999); Imoto et al. (2002); Mooij et al. (2016).

However, none of these approaches is directly transferable to time series causal discovery because their bootstrap sampling needs to be adapted to the lagged interdependencies of time series data.

Our **main contribution** is the introduction of a bootstrap method for time series causal discovery which preserves temporal dependencies. Our method allows (1.) to obtain **uncertainty estimates** for the links of the output graph, and (2.) **improves the stability and accuracy** by aggregating the ensemble of bootstrap graphs to one single output graph with majority voting at the level of each individual edge. In principle, our method can be paired with any time series causal discovery algorithm. In the main text, we investigate the combination with PCMCI+ (Runge, 2020) as a representative of a state-of-the-art constraint-based time series causal discovery method. Results for further methods are presented in the Appendices. **The paper is structured as follows**. In Section 2, we give an overview of time series causal discovery and the PCMCI+ method. In Section 3, we present our bagging and confidence measure technique which we combine with PCMCI+ (Bagged-PCMCI+). With a range of numerical experiments, we show in Section 4 that Bagged-PCMCI+ outperforms standard PCMCI+ and that our method to measure confidence for links is effective. Finally, we summarize the paper in Section 5. The paper is accompanied by an Appendix.

2. Time series causal discovery

2.1. Preliminaries

We consider discrete-time structural causal processes $\mathbf{X}_t = (X_t^1, ..., X_t^N)$ such that

$$X_t^j := f_j(\mathrm{pa}(X_t^j), \eta_t^j) \quad \forall j \in \{1, \ldots, N\} \quad \forall t. \tag{1}$$

Here, f_j are arbitrary measurable functions that depend non-trivially on all their arguments and η_t^j are mutually and temporally independent noises. In a time series graph $\mathcal{G}$, the nodes represent the variables X_t^j at different time lags. The *causal parents* $\mathrm{pa}(X_t^j)$ are the set of variables on which X_t^j depends, and a causal link from $X_{t-\tau}^i$ to X_t^j exists if $X_{t-\tau}^i \in \mathrm{pa}(X_t^j)$ for a time lag τ. A link $X_{t-\tau}^i \to X_t^j$ is a called *lagged* if $\tau > 0$, else it is called *contemporaneous*. In this work, we assume *stationarity* of the causal links: that is, if the causal link $X_{t-\tau}^i \to X_t^j$ exists for some time t, then $X_{t'-\tau}^i \to X_{t'}^j$ also exists for all times $t' \neq t$. We define the set $\mathcal{A}(X_t^j)$ of non-future adjacencies of variable X_t^j as the set of all variables $X_{t-\tau}^i$ for $\tau \geq 0$ that have a causal link with X_t^j.

2.2. PCMCI+

We focus on the combination of our bagging approach with PCMCI+ as it is a widely-used state-of-the-art algorithm for the setting it considers, that is, lag-resolved time series causal discovery with contemporaneous edges but without hidden confounders. PCMCI+ learns the causal time series graph including lagged and contemporaneous links (up to Markov equivalence) under the standard assumptions of Causal Sufficiency, Faithfulness, and the Causal Markov condition, as well as causal stationarity (Runge, 2020). To increase the detection power and maintain well-calibrated tests, PCMCI+ optimizes the choice of conditioning sets in the conditional independence (CI) tests. It is based on two central ideas: (1) separating the skeleton edge removal phase into a lagged and contemporaneous conditioning phase, and (2) constructing conditioning sets in the contemporaneous conditioning phase via the so-called momentary conditional independence (MCI) approach (Runge et al., 2019c), as explained below. Moreover, PCMCI+ is order-independent (Colombo and Maathuis, 2014), which implies that the output does not depend on the order of the variables X^j. More details and examples of PCMCI+ can be found in Runge (2020), here we briefly summarize it.

In its **first phase**, PCMCI+ starts with a fully connected graph and then removes adjacencies among variable pairs by conditional independence testing. To this end, the PC$_1$ algorithm tests all lagged pairs $(X^i_{t-\tau}, X^j_t)$ for $\tau > 0$ conditioning on subsets $\mathbf{S}_k \subseteq \mathcal{A}(X^j_t) \cap \mathbf{X}^-_t$ with the lagged variables $\mathbf{X}^-_t = (\mathbf{X}_{t-1}, \mathbf{X}_{t-2}, \ldots, \mathbf{X}_{t-\tau_{\max}})$ up to a maximum time lag $\tau_{\max}$. If (conditional) independence is detected, the adjacency is removed. The subsets $\mathbf{S}_k$ are chosen with increasing cardinality k: For $k = 0$ all $X^i_{t-\tau}$ with $X^i_{t-\tau} \perp\!\!\!\perp X^j_t$ are removed, for $k = 1$ those with $X^i_{t-\tau} \perp\!\!\!\perp X^j_t | \mathbf{S}_1 \backslash X^i_{t-\tau}$ where $\mathbf{S}_1$ is the adjacency with largest association (not counting $X^i_{t-\tau}$) with X^j_t from the previous step, for $k = 2$ those with $X^i_{t-\tau} \perp\!\!\!\perp X^j_t | \mathbf{S}_2 \backslash X^i_{t-\tau}$ where $\mathbf{S}_2$ are the two adjacencies with largest association (not counting $X^i_{t-\tau}$) with X^j_t from the previous step, and so on. Association strength is measured by the absolute test statistic value of the CI test. This procedure improves recall and speeds up the skeleton phase as compared to the standard PC algorithm skeleton phase. The resulting lagged adjacency sets for each X^j_t of this first phase are denoted $\mathcal{B}^-_t(X^j_t)$. In its **second phase**, the graph $\mathcal{G}$ is initialized with all contemporaneous adjacencies plus all lagged adjacencies from $\hat{\mathcal{B}}^-_t(X^j_t)$ for all X^j_t found in the PC$_1$ algorithm in the first phase. The second phase of PCMCI+ tests all, contemporaneous and lagged, adjacent pairs $(X^i_{t-\tau}, X^j_t)$ for $\tau \geq 0$, but iterates only through contemporaneous conditions $\mathbf{S} \subseteq \mathcal{A}(X^j_t) \cap \mathbf{X}_t$ with the MCI test $X^i_{t-\tau} \perp\!\!\!\perp X^j_t \mid \mathbf{S}, \hat{\mathcal{B}}^-_t(X^j_t) \backslash \{X^i_{t-\tau}\}, \hat{\mathcal{B}}^-_t(X^i_{t-\tau})$. The conditioning on $\hat{\mathcal{B}}^-_t(X^j_t)$ blocks paths through lagged parents while the conditioning on $\hat{\mathcal{B}}^-_t(X^i_{t-\tau})$ has been shown to lead to well-calibrated tests even for highly autocorrelated time series (Runge et al., 2019c; Runge, 2020). After these tests, the time-series-adapted collider orientation phase and rule orientation phase are applied: the former rule orients the collider motifs that contain contemporaneous links based on unshielded triples while the latter rule orients the remaining contemporaneous links based on the Meek rules (Meek, 1995).

In the final graph, the following link types can connect a pair $(X^i_{t-\tau}, X^j_t)$ of vertices: no link (i.e., pair is non-adjacent), direct link $X^i_{t-\tau} {\rightarrow} X^j_t$, opposite direct link $X^i_t {\leftarrow} X^j_t$ (only for $\tau = 0$), unoriented link $X^i_t {\circ\!\!-\!\!\circ} X^j_t$ (only for $\tau = 0$), conflict-indicating link $X^i_t {\times\!\!-\!\!\times}\ X^j_t$ (due to finite sample effects or violations of assumptions, only for $\tau = 0$).

Like other CI-based methods, PCMCI+ has the free parameters α_{PC} (significance level of CI tests), $\tau_{\max}$ (maximal considered time lag), and the choice of the CI test. α_{PC} in PCMCI+ turned out empirically to be an upper bound on the false positives. As opposed to such a statistically-motivated choice, it can also be chosen based on cross-validation or an information criterion (Runge, 2020). $\tau_{\max}$ should be larger or equal to the maximum assumed true time lag of any parent and can in practice also be chosen based on model selection. However, the numerical experiments indicate that a too large $\tau_{\max}$ does not degrade performance much (Runge, 2020). PCMCI+ can flexibly be combined with different CI tests for nonlinear causal discovery, and for different variable types (discrete or continuous, univariate or multivariate).

3. Bootstrapping time series causal discovery

In this section, we motivate and explain our technical contributions to bagging and bootstrap uncertainty quantification of time series causal discovery methods.

3.1. Motivational example

Figure 1 illustrates the benefits and overall structure of our bootstrapping approach by an example. To begin, we generate a multivariate time series according to an instance of eq. (1) and plot its $N = 4$ components in **Figure 1A**. We then estimate the causal dependencies from these data by applying a time series causal discovery algorithm, here PCMCI+, which outputs the graph in **Figure 1E**. This estimated graph (1) deviates from the ground truth causal graph in **Figure 1D** due to finite-sample effects and (2) does not entail any information about the associated uncertainty respectively confidence of output links. Bootstrapping helps to address both of these problems. To this end, we randomly generate B bootstrap datasets from the original data, here $B = 100$, using the sampling approach explained in Sec. 3.2, where B is the number of chosen bootstrap realizations. Next, we individually apply the to-be-bootstrapped algorithm (here PCMCI+) to all B bootstrap datasets, thus giving rise to B estimated graphs, see **Figure 1B**. We now use this ensemble of estimated graphs in two ways.

First, for bagging, we aggregate the entire ensemble of graphs into a single graph by majority voting (see aggregation prescription in Sec. 3.3), which here outputs the graph in **Figure 1C**. The intuition is the following: PCMCI+ is asymptotically consistent but can make errors on finite samples (e.g., false positives and false negatives regarding adjacencies and orientations). Random finite-sample effects tend to cancel out in the aggregated graph, thus making it on average a more accurate estimate than the single estimate on the original data. Specifically, we expect that false links tend to appear in the minority of bootstrap graphs, such that there are fewer false positives in the aggregated graph.

Second, for uncertainty quantification, we count how often each individual edge type (no link, $\rightarrow$, $\leftarrow$, $\circ\!\!-\!\!\circ$, or $\times\!\!-\!\!\times$) of a specific link between each pair $(X^i_{t-\tau}, X^j_t)$ of the aggregated graph appears in the ensemble of graphs and employ the according frequencies as proxies for the confidence in the respective edges. For example, if an edge type for a specific link occurs in 80% of the ensemble of graphs, we have stronger confidence in that link than in a different link where the majority edge type occurs only among 45% of the ensemble of graphs. In **Figure 1C**, we visualize the frequencies by the thickness of edges, and we further motivate and discuss this approach to uncertainty quantification in Sec. 3.4.

3.2. Temporal-dependencies preserving sampling approach

Our sampling approach is suitable for all algorithms that, as PCMCI+, internally use a moving-window scheme to generate the sets of samples upon which they further operate.

Specifically, PCMCI+ tests marginal and conditional independencies $X \perp\!\!\!\perp Y \mid Z$ where X, Y, and the potentially empty Z can contain variables at different time steps; for example, $X = \{X^1_{t-1}\}$, $Y = \{X^2_t\}$ and $Z = \{X^3_{t-2}, X^2_{t-1}\}$. To create samples for these tests, the algorithm makes a stationarity assumption and employs the following moving window approach: Let $\mathbf{D} = \{w_s \mid s \in \mathcal{I}_\mathbf{D}\}$ with $\mathcal{I}_\mathbf{D} = \{0, 1, \ldots, T-2, T-1\}$ be the time series dataset where $w_s = (x^1_s, x^2_s, \ldots, x^N_s)$ are the measured values at time s. Then, for testing $X \perp\!\!\!\perp Y \mid Z$ with $X = \{X^1_{t-1}\}$, $Y = \{X^2_t\}$ and $Z = \{X^3_{t-2}, X^2_{t-1}\}$ the algorithm uses the set of samples $\mathbf{S}^{(X,Y,Z)} = \{v^{(X,Y,Z)}_s \mid s \in \mathcal{I}_\mathbf{S}\}$ with index set $\mathcal{I}_\mathbf{S} = \{2 \cdot \tau_{\max}, 2 \cdot \tau_{\max} + 1, \ldots, T-2, T-1\}$ and samples $v^{(X,Y,Z)}_s = (x^1_{s-1}, x^2_s, x^3_{s-2}, x^2_{s-1})$. The choice that $\mathcal{I}_\mathbf{S}$ starts with $2 \cdot \tau_{\max}$ instead of 0 is specific to PCMCI+ and ensures that all independence tests employ the same number of samples, but this choice is irrelevant to our sampling approach described here.

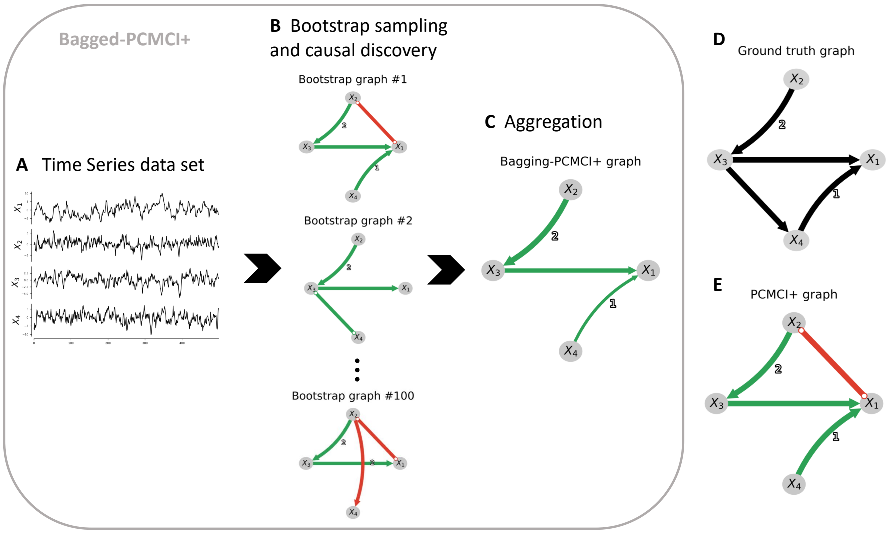

Figure 1: Motivational example and schematic of our approach. **(A)** Time series for a model as in eq. (1) with linear functions f_j. **(B)** Ensemble of $B = 100$ estimated causal graphs obtained by applying PCMCI+ to each bootstrap dataset randomly drawn as described in Sec. 3.2. Green links indicate true positives, red links represent false positives, and the numbers next to the edges indicate non-zero time lags. **(C)** Graph obtained by aggregating the ensemble of graphs as explained in Sec. 3.3, with confidence scores obtained as explained in Sec. 3.4 and visualized as the thickness of edges. **(D)** Ground truth causal graph of the model. **(E)** Estimated causal graph obtained by applying PCMCI+ to the data in part **A**.

To bootstrap in a non-temporal setting, one would draw bootstrap datasets $\mathbf{D}^*$ from the original dataset $\mathbf{D}$ by letting $\mathbf{D}^* = \{w_s \mid s \in \mathcal{I}^*_{\mathbf{D}}\}$ where $\mathcal{I}^*_{\mathbf{D}}$ is sampled with replacement from $\mathcal{I}_{\mathbf{D}}$. Using this procedure in the temporal setting and then applying the moving window approach to $\mathbf{D}^*$ would, however, combine values from possibly far away time steps into single samples. Such combinations of values do not reflect the actual temporal dependencies, destroy lag relationships, and, hence, spoil the downstream use of the set of samples (for PCMCI+, spoil the independence testing). Thus, we instead sample with replacement the index set $\mathcal{I}^*_{\mathbf{S}}$ from $\mathcal{I}_{\mathbf{S}}$ and let $\mathbf{S}^{*,(X,Y,Z)} = \{v_s^{(X,Y,Z)} \mid s \in \mathcal{I}^*_{\mathbf{S}}\}$. Here, $v_s^{(X,Y,Z)}$ is exactly as in the previous paragraph and combines values from different time steps in the correct way. For each bootstrap realization, we run the to-be-bootstrapped causal discovery algorithm in combination with the temporal-dependencies preserving sampling approach. That is, the causal discovery algorithm internally operates with sets of samples which are resampled with the aforementioned approach. We stress that, as in the standard bootstrap, resampling happens only once per bootstrap realization. For each of the b boostrap realizations (where $1 \leq b \leq B$), the set $\mathcal{I}^{*,(b)}_{\mathbf{S}}$ is drawn once from $\mathcal{I}_{\mathbf{S}}$, and then the sets $\mathbf{S}^{*,(X,Y,Z)} = \{v_s^{(X,Y,Z)} \mid s \in \mathcal{I}^{*,(b)}_{\mathbf{S}}\}$ are used *for all* (conditional) independence tests that are called in the b-th boostrap realization. The to-be-bootstrapped algorithm outputs a causal graph for each bootstrap realization, thus generating an ensemble $\mathcal{C} = \{\mathcal{G}_1, \ldots, \mathcal{G}_B\}$ of B causal graphs as an intermediate result.

3.3. Aggregation by edge-wise majority vote

The prescription for aggregating the bootstrap ensemble $\mathcal{C} = \{\mathcal{G}_1, \dots, \mathcal{G}_B\}$ of estimated graphs to a single final estimated graph $\mathcal{G}_{bagged}$ is independent of the sampling approach. While a variety of choices seems possible, in this work we choose to employ the following strategy: *majority voting*.

In Appendix B, **Alg. 1** summarizes the aggregation of the B causal graphs $\mathcal{C} = \{\mathcal{G}_1, \dots, \mathcal{G}_B\}$ to a single final output graph $\mathcal{G}_{bagged}$. For each ordered pair of distinct vertices $(X^i_{t-\tau}, X^j_t)$, we first run through the ensemble of $\mathcal{C}$ and record the relative frequency of each of the possible edge types. For PCMCI+, the possible edge types are *no edge* and $\rightarrow$ for lagged pairs ($\tau > 0$), and for contemporaneous pairs ($\tau = 0$) there additionally are the edge types $\leftarrow$, $\circ\!\!-\!\!\circ$, and $\times\!\!-\!\!\times$. Here, the edges $X^i_t {\circ\!\!-\!\!\circ} X^j_t$ and $X^i_t {\times\!\!-\!\!\times} X^j_t$ indicate inconclusiveness respectively conflicting information about the direction of the link.[1] In $\mathcal{G}_{bagged}$, we then connect $X^i_{t-\tau}$ and X^j_t by an edge of the type with the highest frequency (where connection by *no edge* means that the vertices are not, actually, connected by an edge). To resolve potential ties, we employ the preference order *no edge*, $\times\!\!-\!\!\times$, $\circ\!\!-\!\!\circ$, and, at the same preference level, $\rightarrow$ and $\leftarrow$ (from highest to lowest). In case of a tie between only $\rightarrow$ and $\leftarrow$, we resolve to a conflict-indicating link $\times\!\!-\!\!\times$. This choice of preference order is conservative in the sense that, among edge types other than *no edge*, it prefers edge types that convey less conclusive claims.

The adaptation of this aggregation strategy to other sets of possible edge types and other tie-resolving strategies is straightforward. For example, we have also explored an *alternative* aggregation strategy. In the first step of this alternative approach, the orientation of edges is ignored, and the focus is only on determining the adjacency of each pair of vertices. This is done through majority voting between *no edge* and all other edge types. In the second step, the adjacencies identified in the first step are oriented based on majority voting. This alternative approach ensures that *no edge* can only be voted on if it appears in more than half of the bootstrap ensemble of graphs.

The key advantage of aggregating at the level of individual edges is the simplicity of implementation and interpretation. However, this way of aggregating does not in general preserve acyclicity and, more generally speaking, graphical properties that the to-be-bootstrapped algorithms might presuppose. One might view this non-preservation of graphical properties as a disadvantage. Alternatively, one might view this property as a useful feature. The presence of cycles points the method users to a large uncertainty in the respective parts of the graph and to a potential violation of the respective graphical assumption, thus advising them to interpret the results with great care. It is not hard to come up with alternative aggregation methods that preserve acyclicity. For example, one could first aggregate the graphs as we currently do and then, for each cycle in the aggregated graph, remove the edge that appears with the lowest frequency in the ensemble $\mathcal{C}$. We leave the exploration of this and other alternative aggregation methods, for example using current techniques that minimize a modified structural Hamming distance of the aggregated graph to the entire set of graphs (Wang and Peng, 2014), to future research.

3.4. Edge frequencies as confidence scores

Assuming stationarity, the original time series dataset $\mathbf{D}$ is a (non-*iid*) sample from the stationary distribution of the stochastic process defined by the respective instance eq. (1). The sample set $\mathbf{S}^{(X,Y,Z)}$ defined in Sec. 3.2 is, thus, a (non-*iid*) sample from a t-independent distribution $F^{(X,Y,Z)}$.

[1] Conflicting information can arise due to a violation of assumptions or incorrect results of the independence tests.

This sample set defines an empirical distribution $\hat{F}_T^{(X,Y,Z)}$ that approximates $F^{(X,Y,Z)}$, and by the design of our sampling approach the bootstrap sample set $\mathbf{S}^{*,(X,Y,Z)}$ is a sample from $\hat{F}_T^{(X,Y,Z)}$.

Apart from autodependencies, we are hence in the standard bootstrap setting and can use the bootstrap to approximate uncertainties with respect to $F^{(X,Y,Z)}$. Specifically, we can quantify the uncertainty respectively confidence in a given edge e in the aggregated graph $\mathcal{G}_{bagged}$ by the frequency r_e^{boot} with which this edge appears in the bootstrap ensemble $\mathcal{C}$ of estimated graphs. The precise meaning of this measure of uncertainty is as follows: As $T \to \infty$ and $B \to \infty$, we expect that r_e^{boot} converges to the frequency r_e with which the same edge e appears in the hypothetical ensemble of graphs obtained by repeatedly applying the to-be-boostraped method on time series datasets $\mathbf{D}_1, \mathbf{D}_2, \ldots$ that are generated *from the model* (as opposed to sampling from $\mathbf{D}$). Thus, a higher r_e^{boot} means a higher confidence in the edge e. In our visualizations of aggregated graphs, we represent r_e^{boot} by the edge's thickness, where edges with higher r_e^{boot} are thicker.

The advantages of this edge-wise uncertainty measure are (1) its simplicity and (2) that it conforms well with the edge-wise aggregation of the bootstrap ensemble $\mathcal{C}$ of graphs to $\mathcal{G}_{bagged}$. However, in the bootstrap ensemble, the presence, absence and orientation of an edge might be correlated with the presence, absence and orientation of other edges. Since our aggregation scheme operates *edge-wise*, that is, on the level of individual edges, the aggregation does not take into these potential dependencies between edges. As a result, information about such dependencies is lost during the edge-wise aggregation process.

4. Numerical experiments

4.1. Evaluation of Bagged-PCMCI+ performance on synthetic data

The numerical experiments model several typical challenges in time series causal discovery (Runge et al., 2019a): contemporaneous and time-lagged causal dependencies, nonlinearity, non-Gaussian noise distribution, strong autocorrelation, large numbers of variables and considered time lags. For a better comparison to PCMCI+, we use a similar setup to the numerical experiments presented in Runge (2020). The synthetic data is generated according to the following additive model:

$$X_t^j = a_j X_{t-1}^j + \sum_i c_i f_i(X_{t-\tau_i}^i) + \eta_t^j \quad \forall j \in \{1, \ldots, N\} \tag{2}$$

Autocorrelation coefficients a_j are uniformly drawn from $[\max(0, a - 0.3), a]$. The values of the autocorrelation a and of other parameters of the data-generating model are indicated in the header of Figures 2, 3, 4, and Figures in the Appendix. The noise terms η^j are independent and identically distributed zero-mean Gaussians $\mathcal{N}(0, \sigma^2)$ with standard deviation σ drawn from $[0.5, 2]$ or Weibull (scale parameter 2) depending on the setup. In addition to autodependency links, for each model $L = \lfloor 1.5 \cdot N \rfloor$ (except when $N = 2$ where we set $L = 1$) cross-links are chosen with linear functional dependencies $f_i(x) = x$ or with nonlinear functional dependencies $f_i(x) = x + 5x^2 e^{-x^2/20}$ depending on the setup. Coefficients c_i are drawn uniformly from $\pm[0.1, 0.5]$. 30% of the links are contemporaneous ($\tau_i = 0$) and the remaining 70% are lagged links with τ_i drawn from $\{1, \ldots, 5\}$. Only stationary models are considered. We have an average cross-in-degree of $d = 1.5$ for all network sizes (plus an auto-dependency) implying that models become sparser for larger N. We consider several model setups: linear with Gaussian noise, linear with mixed noise (50% Gaussian and 50% Weibull), and nonlinear (50% linear and 50% nonlinear dependencies) with Gaussian noise. We consider the PCMCI+ algorithm and Bagged(B)-PCMCI+ where the number of bootstrap realizations B varies. If no aggregation strategy is mentioned, the summary graph of Bagged(B)-PCMCI+ is obtained by simple majority voting. To test conditional independence, we use the partial correlation test (ParCorr) for linear experiments or the GPDC test (Runge et al., 2019c) for nonlinear experiments implemented in the tigramite Python package (see Appendix A). To show that our bagging approach is general, we also include numerical experiments that, instead of PCMCI+, use a modified PC algorithm adapted to time series and the LPCMCI algorithm (Gerhardus and Runge, 2020) as the base algorithm. We chose the adapted PC algorithm because it can be combined with our resampling approach easily, and decided to also experiment with LPCMCI as a state-of-the-art algorithm for time series causal discovery in the presence of contemporaneous edges and hidden confounders. For the experiment with LPCMCI, we randomly choose 70% from the N variables of each model as observed ($\tilde{N} = \lceil 0.7N \rceil$). The rest of the variables are unobserved.

Performance is evaluated with recall (equivalent to True Positive Rate, TPR), precision, and the area under the precision-recall curve (PR-AUC). For adjacencies, precision and recall are distinguished between lagged cross-links ($i \neq j$), contemporaneous ($\tau = 0$), and autodependency links ($i = j$). Due to time order, lagged links (and autodependencies) are automatically oriented. Performance of contemporaneous orientation is evaluated with contemporaneous orientation precision, which is measured as the fraction of correctly oriented links ($\circ\!\!-\!\!\circ$,$\rightarrow$,$\leftarrow$) among all estimated adjacencies, and with recall as the fraction of correct orientations among all true contemporaneous links. False positive rates (FPR) are also shown to evaluate whether the methods control false positives at the chosen significance level α_{PC}. Furthermore, the fraction of conflicting links among all detected contemporaneous adjacencies is calculated. Finally, we give the average runtimes that were evaluated on an AMD EPYC 7763 processor. All metrics (and their standard errors) are computed across all estimated graphs from 500 different additive models (and associated realizations) described in Equation 2 with time series length T.

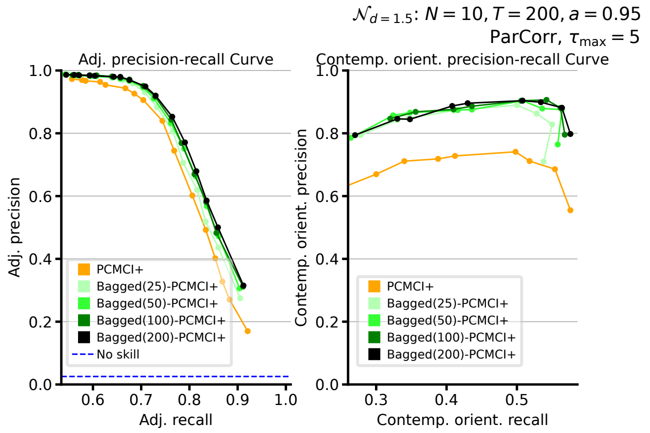

Figure 2: Precision-recall curves for adjacencies (left) and contemporaneous orientations (right) obtained by varying the significance level α_{PC} in PCMCI+ and Bagged-PCMCI+ for the model setup as shown in the top right. Results are shown for PCMCI+ (orange line) and Bagged-PCMCI+ with different numbers of bootstrap replicas B (lines with different shades of green).

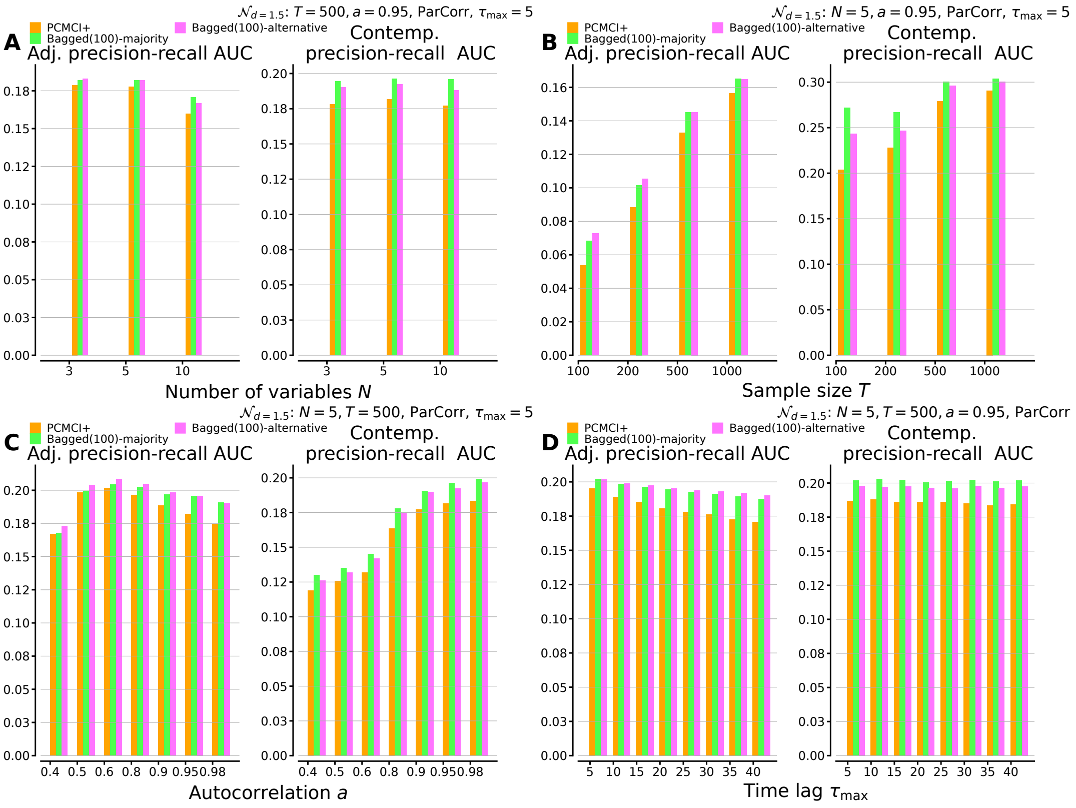

Figure 3: PR-AUC of adjacencies and contemporaneous orientations for the linear and Gaussian noise setup for a varying: (**A**) number of variables N, (**B**) sample size T, (**C**) autocorrelation a, (**D**) PCMCI+ maximum time lag $\tau_{\max}$. The model setup is shown at the top right. The Precision-recall curves are obtained by varying the significance level α_{PC} of PCMCI+. PR-AUC is the area under these curves. The results for PCMCI+ (orange bars), Bagged-PCMCI+ with $B = 100$ bootstrap replicas for the majority voting aggregation strategy (green bars) and the alternative aggregation strategy (pink bars) are shown.

Figure 2 depicts **precision-recall curves** for adjacencies as well as contemporaneous orientations obtained by varying the hyperparameter α_{PC} for a model setup that stands exemplary for others (see Appendix C). The precision-recall values of Bagged-PCMCI+ are systematically higher than those of PCMCI+, regarding adjacencies and even more pronounced for contemporaneous orientations. Moreover, larger numbers of bootstrap replicates B result in enhanced performance, but we observe no strong differences between $B = 50$ and $B = 200$. More precision-recall curves are shown in Appendix C for different model parameters (autocorrelation, number of variables, and time sample size). In **Figure 3**, we plot the area under the precision-recall curves for varying model parameters. Bagged-PCMCI+ has systematically **higher PR-AUC** than PCMCI+, indicating that Bagged-PCMCI+ outperforms PCMCI in terms of precision and recall. Noticeably, we found that Bagged-PCMCI+ outperforms PCMCI+ more clearly in statistically more challenging regimes with higher autocorrelation a, larger number of variables N, and smaller time sample sizes T. We also compare the PR-AUC of Bagged(100)-PCMCI+ with the simple majority voting strategy to the alternative aggregation strategy introduced in Section 3.3. The alternative aggregation strategy generally leads to improvement in terms of the PR-AUC for the adjacencies, whereas the simple majority voting strategy is superior regarding contemporaneous orientations.

In the Appendix C, **Figure 5** gives further details for the linear Gaussian setup for varying significance level α_{PC}. The right column of Figure 5 clearly shows that both methods (PCMCI+ and Bagged-PCMCI+) control the FPR below the significance level α_{PC} (grey line), but Bagged-PCMCI+ consistently exhibits **lower FPR** compared to PCMCI+ for all types of links (lagged, contemporaneous, auto links) and across significance levels α_{PC} ranging from 0.001 to 0.1. Thus, treating the significance level as a hyperparameter, one can use a higher level α_{PC} for Bagged-PCMCI+ than for PCMCI+ while still controlling the FPR below α_{PC}. There appear to be slightly fewer conflicts for Bagged-PCMCI+. Runtimes are, as expected, higher for Bagged-PCMCI+, but these can be reduced by embarrassingly parallelization. In the Appendix C, we provide additional figures for varying one of autocorrelation a, number of variables N, time sample size T, and maximum time lag $\tau_{\max}$ with $\alpha_{PC} = 0.01$. We found that Bagged-PCMCI+ seems robust to large maximum time lags $\tau_{\max}$ (even when $\tau_{\max}$ is much larger than the true maximum time lag of 5) for the studied sample size $T = 500$. We have observed similar improvements of Bagged-PCMCI+ in terms of precision and recall compared to PCMCI+ in the nonlinear and mixed noise experiments (see **Appendix C.2**).

Finally, we have combined our bagging approach with a modified PC algorithm adapted to time series and LPCMCI. We provide results for these experiments in the **Appendix D** and **Appendix E** respectively. Similar to PCMCI+, we have found that Bagged-PC enhances its base algorithm PC in terms of precision and recall. For Bagged-LPCMCI, we have identified improvements in terms of contemporaneous orientation precision and recall relative to LPCMCI.

4.2. Evaluation of Bagged-PCMCI+ confidence measure

We conduct numerical experiments to assess the ability of Bagged-PCMCI+ to determine a confidence degree for links in the output graph. Ideally, we would like the Bagged-PCMCI+ link frequency obtained on a single data sample to approximate closely the frequency of links along graphs obtained independently by PCMCI+ on an infinite number of data samples, which we call reference link frequencies. In practice, it is only possible to approximate the reference link frequencies by using a large but limited number of data samples. Here, we design two experiments to evaluate the ability of the proposed confidence measures to approximate the reference link frequencies.

In the **first experiment**, we generate 250 different additive models (see Equation 2). For each of these models, we generate $D = 100$ independent data samples with the same additive model. That is, only the noise terms change across the samples. For each of these 100 samples, causal graphs are estimated independently using PCMCI+ and Bagged-PCMCI+. For each edge, we estimate its *reference link frequency* by calculating the frequency of the most frequent link types across the PCMCI+ ensemble of 100 graphs. We use the mean of our proposed confidence measure (i.e. the mean of the Bagged-PCMCI+ link frequencies along the 100 Bagged-PCMCI+ graphs) to reduce the amount of noise in the estimation. We also calculate the standard deviation of the Bagged-PCMCI+ link frequencies along the 100 Bagged-PCMCI+ graphs to estimate its uncertainty. If effective, we expect the Bagged-PCMCI+ confidence measure to approximately follow the estimated reference link frequency. In **Figure 4**, we show the results of this first experiment for model parameters and method parameters indicated at the top right and for $B = 1000$. We plot averaged confidence measures (y-axis) against the reference link frequencies (x-axis) for different causal dependencies (lagged, contemporaneous, or all) and for existing/absent links of the ground truth graphs. Figure 4 shows that the Bagged-PCMCI+ link frequency (confidence measure) overall tends to follow the reference link frequency. We can notice a bias for low reference link frequencies as Bagged-PCMCI+ tends to overestimate the reference link frequencies between 40% to 60%. This bias seems consistent across different types of causal dependencies. The one standard deviation error bars give indications of the uncertainties in the confidence measure and the estimated reference link frequency. When taking both uncertainties into account, the error bars cross the expected $x = y$ diagonal line more than 99% of the estimated link frequencies. This high percentage is a good indication that the Bagged-PCMCI+ confidence measure approximates the reference link frequency of PCMCI+ reasonably well.

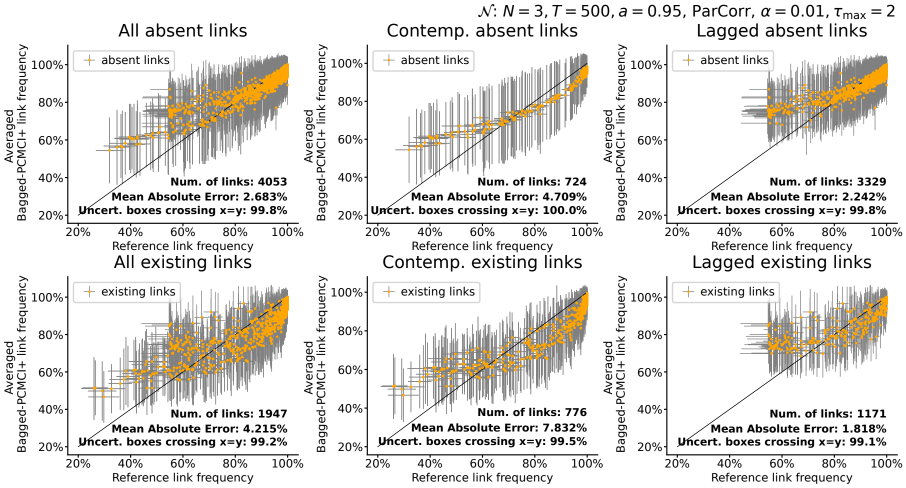

Figure 4: Evaluation of the proposed confidence measures for a linear Gaussian setup with parameters indicated at the top right. y-values (averaged confidence measure) and x-values (reference) are as explained in the text. Grey bars indicate the one standard deviation error around the estimated value.

In the **second experiment**, we quantify the average absolute error of the Bagged-PCMCI+ link frequency relative to the reference link frequency for different causal dependencies (lagged, contemporaneous, or all) and for existing links and absent links of the ground truth model. For each additive model, the reference link frequency is calculated as the frequency of the most recurrent link type of PCMCI+ graphs over $D = 5000$ independent samples. Here, the Bagged-PCMCI+ link frequency is a confidence measure for the first data sample of each additive model (the confidence is, as compared to the first experiment, not averaged). We calculate the mean absolute error between the confidence measures and the estimated reference link frequencies. We generate a total of 500 different additive models and average the mean absolute link frequency errors. To study the effect of the number of bootstrap realizations B on the error, we vary B from 25 to 2000. We present the results of the mean absolute link frequency errors in **Fig 13**. For lagged links and all links, the mean absolute errors are slightly below 3% which seems relatively low. For the contemporaneous links, the 7% errors demonstrate that the estimation of the contemporaneous links is a more difficult task for the current method. It is also not clear if the mean absolute error converges to zero if B goes to infinity. We observe that a larger B leads to a lower mean absolute error of Bagged-PCMCI+. This confirms that increasing B enhances the performance of the Bagged-PCMCI+. The mean absolute frequency error reduces by about 10% when increasing B from 25 to 500. Unfortunately, without parallelization, the increase in performance comes at the cost of a 20-fold increase in runtime. This is why we recommend using a number of bootstrap realizations which is adapted for the application and available computing resources, but preferably larger or equal to 100.

5. Conclusion

In this paper, we have made **two major contributions** to time series causal discovery. First, we propose a **bootstrap aggregation** by majority voting that can be combined with any time series causal discovery algorithm. Here, we combine our bootstrap aggregation approach with the state-of-the-art time series causal discovery PCMCI+ algorithm (referred to as Bagged-PCMCI+). Through extensive numerical experiments, we show that Bagged-PCMCI+ greatly reduces the number of false positives compared to the base PCMCI+ algorithm. In addition, Bagged-PCMCI+ has a higher precision-recall regarding adjacencies and orientations of contemporaneous edges compared to PCMCI+. The same outcomes hold true when combining our bootstrap aggregation approach with a modified version of the PC algorithm or with the LPCMCI algorithm, suggesting that our numerical results are consistent and generalizable. Our second contribution is a **confidence measure for links** in a time series graph, which is calculated as the link frequencies along the graphs learned on bootstrap replicates. Numerical experiments show that the proposed method gives a pertinent confidence measure for links of the output graph. The **main strengths** of our method are that it can be coupled with a wide range of time series causal discovery algorithms, and it can be of substantial use in many real-world applications, especially for orienting contemporaneous causal links or when confidence measures for links are desired. The **main weaknesses** of our method so far are its higher computational cost and longer runtime. One solution to decrease runtime is to embarrassingly parallelize the bootstrap process. In addition, the current method of aggregating through majority voting has a limitation. It can lead to cyclic graphs that are not always desirable. Therefore, exploring alternative methods of aggregation will constitute a crucial step for future research.

Acknowledgments

DKRZ provided computational resources (grant no. 1083). K.D. and V.E have received funding provided by the European Research Council (ERC) Synergy Grant “Understanding and Modelling the Earth System with Machine Learning (USMILE)” under the Horizon 2020 research and innovation programme (Grant agreement No. 855187). J.R. has received funding from the European Research Council (ERC) Starting Grant CausalEarth under the European Union’s Horizon 2020 research and innovation program (Grant Agreement No. 948112). We thank Tom Hochsprung for his helpful comments.

References

R. Ayesha Ali, Thomas S. Richardson, and Peter Spirtes. Markov equivalence for ancestral graphs. *The Annals of Statistics*, 37:2808–2837, 2009. ISSN 0090-5364, 2168-8966. doi: 10.1214/08-AOS626.

L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

David Maxwell Chickering. Learning equivalence classes of bayesian-network structures. *J. Mach. Learn. Res.*, 2:445–498, 1996.

Diego Colombo and Marloes H. Maathuis. Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.*, 15(1):3741–3782, jan 2014. ISSN 1532-4435.

Thomas G. Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems*, pages 1–15, Berlin, Heidelberg, 2000. Springer Berlin Heidelberg. ISBN 978-3-540-45014-6.

Imme Ebert-Uphoff and Yi Deng. Causal discovery for climate research using graphical models. *Journal of Climate*, 25(17):5648–5665, 2012.

Nir Friedman, Moisés Goldszmidt, and Abraham J. Wyner. Data analysis with bayesian networks: A bootstrap approach. *ArXiv*, abs/1301.6695, 1999.

Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe'er. Using bayesian networks to analyze expression data. In *Annual International Conference on Research in Computational Molecular Biology*, 2000.

Mikel Galar, Alberto Fernandez, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera. A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4):463–484, 2012. doi: 10.1109/TSMCC.2011.2161285.

Evgenia Galytska, Katja Weigel, Dörthe Handorf, Ralf Jaiser, Raphael Köhler, Jakob Runge, and Veronika Eyring. Evaluating causal arctic-midlatitude teleconnections in cmip6. *Journal of Geophysical Research: Atmospheres*, 128(17):e2022JD037978, 2023. doi: https://doi.org/10.1029/2022JD037978.

M.A. Ganaie, Minghui Hu, A.K. Malik, M. Tanveer, and P.N. Suganthan. Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, 115:105151, 2022. ISSN 0952-1976. doi: https://doi.org/10.1016/j.engappai.2022.105151.

Andreas Gerhardus and Jakob Runge. High-recall causal discovery for autocorrelated time series with latent confounders. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12615–12625. Curran Associates, Inc., 2020.

Clive W. J. Granger. *Investigating Causal Relations by Econometric Models and Cross-Spectral Methods*, volume 2 of *Econometric Society Monographs*, page 31–47. Cambridge University Press, 2001. doi: 10.1017/CBO9780511753978.002.

Pei Guo, Yiyi Huang, and Jianwu Wang. Scalable and flexible two-phase ensemble algorithms for causality discovery. *Big Data Research*, 26:100252, 2021. ISSN 2214-5796. doi: https://doi.org/10.1016/j.bdr.2021.100252.

Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O. Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11 (56):1709–1731, 2010.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015. doi: 10.1017/CBO9781139025751.

Seiya Imoto, Sun Yong Kim, Hidetoshi Shimodaira, Sachiyo Aburatani, Kousuke Tashiro, Satoru Kuhara, and Satoru Miyano. Bootstrap analysis of gene networks based on bayesian networks and nonparametric regression. *Genome Informatics*, 13:369–370, 2002.

Maciej Kaminski, Mingzhou Ding, Wilson A. Truccolo, and Steven L. Bressler. Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance. *Biological Cybernetics*, 85:145–157, 2001.

Soufiane Karmouche, Evgenia Galytska, Jakob Runge, Gerald A. Meehl, Adam S. Phillips, Katja Weigel, and Veronika Eyring. Regime-oriented causal model evaluation of atlantic–pacific teleconnections in cmip6. *Earth System Dynamics*, 2023.

Marlene Kretschmer, Dim Coumou, Jonathan F Donges, and Jakob Runge. Using causal effect networks to analyze different arctic drivers of midlatitude winter circulation. *Journal of climate*, 29(11):4069–4081, 2016.

Shuang Li, Li Hsu, Jie Peng, and Pei Wang. Bootstrap inference for network construction with an application to a breast cancer microarray study. *The annals of applied statistics*, 7 1:391–417, 2011.

Christopher Meek. Causal inference and causal explanation with background knowledge. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, page 403–410, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. ISBN 1558603859.

Nicolai Meinshausen and Peter H. Buehlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72, 2008.

Joris M. Mooij, Sara Magliacane, and Tom Claassen. Joint causal inference from multiple contexts. *J. Mach. Learn. Res.*, 21:99:1–99:108, 2016.

Judea Pearl. *Causality*. Cambridge University Press, 2 edition, 2009. doi: 10.1017/CBO9780511803161.

Joseph Ramsey, Kun Zhang, Madelyn Glymour, Ruben Sanchez Romero, Biwei Huang, Immé, Ebert-Uphoff, Savini M. Samarasinghe, Elizabeth A. Barnes, and Clark Glymour. Tetrad - a toolbox for causal discovery. In *8th International Workshop on Climate Informatics*, 2018.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.

Jakob Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Conference on Uncertainty in Artificial Intelligence*, 2020.

Jakob Runge, Sebastian Bathiany, Erik M. Bollt, Gustau Camps-Valls, Dim Coumou, Ethan R. Deyle, Clark Glymour, Marlene Kretschmer, Miguel D. Mahecha, Jordi Muñoz-Marí, Egbert H. van Nes, J. Peters, Rick Quax, Markus Reichstein, Marten Scheffer, Bernhard Schölkopf, Peter L. Spirtes, George Sugihara, Jie Sun, Kun Zhang, and Jakob Zscheischler. Inferring causation from time series in earth system sciences. *Nature Communications*, 10, 2019a.

Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, 5 (11):eaau4996, 2019b. doi: 10.1126/sciadv.aau4996.

Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science advances*, 5 (11):eaau4996, 2019c.

Peter Spirtes, Clark Glymour, Scheines N., and Richard. *Causation, Prediction, and Search*. Mit Press: Cambridge, 2000.

Peter L. Spirtes and Clark Glymour. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9:62 – 72, 1991.

Eric V. Strobl, Peter L. Spirtes, and Shyam Visweswaran. Estimating and controlling the false discovery rate of the pc algorithm using edge-specific p-values. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10:1 – 37, 2016.

George Sugihara, Robert M. May, Hao Ye, Chih-hao Hsieh, Ethan R. Deyle, Michael Fogarty, and Stephan B. Munch. Detecting causality in complex ecosystems. *Science*, 338:496 – 500, 2012.

Ru Wang and Jie Peng. Learning directed acyclic graphs via bootstrap aggregating. *arXiv: Machine Learning*, 2014.

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

Appendix A. Code availability

Code for the bootstrap aggregation of PCMCI+, LPCMCI, and other variants as well as code for the conditional independence tests ParCorr and GPDC are provided as part of the tigramite Python package at `https://github.com/jakobrunge/tigramite`. Numerical experiments can be reproduced with the code available at `https://github.com/EyringMLClimateGroup/debeire24clear_Bagged_TimeSeries_Causality`.

Appendix B. Description of the aggregation by edge-wise majority (pseudo-code)

Algorithm 1 (Aggregation of bootstrap causal graphs by majority voting of edges)

```
input  : Ensemble of bootstrap causal graphs $\mathcal{C} = \{\mathcal{G}_1, \dots, \mathcal{G}_B\}$
output: Aggregated graph $\mathcal{G}_{bagged}$, the relative frequency $F$ of link types for each pair of vertices
        $(X^i_{t-\tau}, X^j_t)$ and link type $e$ across $\mathcal{C}$
Initialize an empty graph $\mathcal{G}_{bagged}$ with the same nodes as in the bootstrap causal graphs
Initialize a two-level nested dictionary $F$ for recording relative frequencies of link types
forall pairs $(X^i_{t-\tau}, X^j_t)$ with $0 \leq \tau \leq \tau_{\max}$, $1 \leq i, j \leq N$, and $i < j$ if $\tau = 0$ do
    forall possible link types $e$ do
        $n(e) \leftarrow$ number of graphs in $\mathcal{C}$ in which $(X^i_{t-\tau}, X^j_t)$is connected by $e$
        $F[(X^i_{t-\tau}, X^j_t)][e] \leftarrow \frac{n(e)}{B}$
    end
    Link $(X^i_{t-\tau}, X^j_t)$ in $\mathcal{G}_{bagged} \leftarrow \underset{e}{\text{argmax}}\, F[(X^i_{t-\tau}, X^j_t)][e]$ [a]
end
return $\mathcal{G}_{bagged}, F$
```

[a] Ties are resolved according to the preferred order given in section 3.3.

Appendix C. Additional Bagged-PCMCI+ experiments

C.1. Further numerical experiments for the linear and Gaussian noise setup

We investigate the impact of model parameters on Bagged-PCMCI+ in the linear and Gaussian noise experiments compared to the model setup of **Figure 2** in the main text. Precision-recall curves for additional model setups show the impact of a smaller number of variables of $N = 5$ instead of $N = 10$ in the main text (**Figure 6**), increased sample size T from 200 to 500 for $N = 5$ (**Figure 8**), and a decreased autocorrelation coefficient a from 0.95 to 0.6 for $T = 500$ and $N = 5$ (**Figure 10**). For all these model setups we also provide the individual precisions, recalls, F1-scores plots for adjacencies and contemporaneous orientations, as well as the runtimes and number of conflicts for varying α_{PC} (see **Figures 5, 7, 9, and 11**). F_1 scores are calculated for adjacencies and contemporaneous orientations as the harmonic mean of precision and recall: $F_1 = 2\frac{precision \cdot recall}{precision + recall}$.

Across all these model setups, for a given α_{PC} Bagged-PCMCI+ has similar recall and higher precision as compared to PCMCI+, particularly in orienting contemporaneous links. Moreover, these improvements are stronger in the more challenging settings (high autocorrelation a, short time sample size T, and high number of variables N).

While, for a given α_{PC}, PCMCI+ can have higher adjacency recall, the fair comparison here is the area under the precision-recall curve, which is higher for Bagged-PCMCI+. This implies that one can always choose a higher α_{PC} to obtain a better recall with Bagged-PCMCI+, while still retaining the same or better precision.

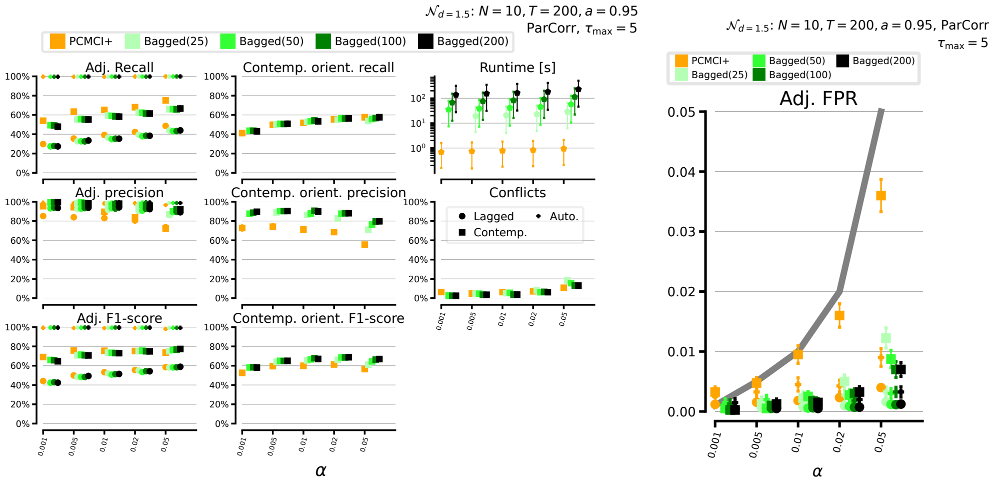

Figure 5: Numerical experiments with linear Gaussian setup for varying significance level α_{PC} of PCMCI+.

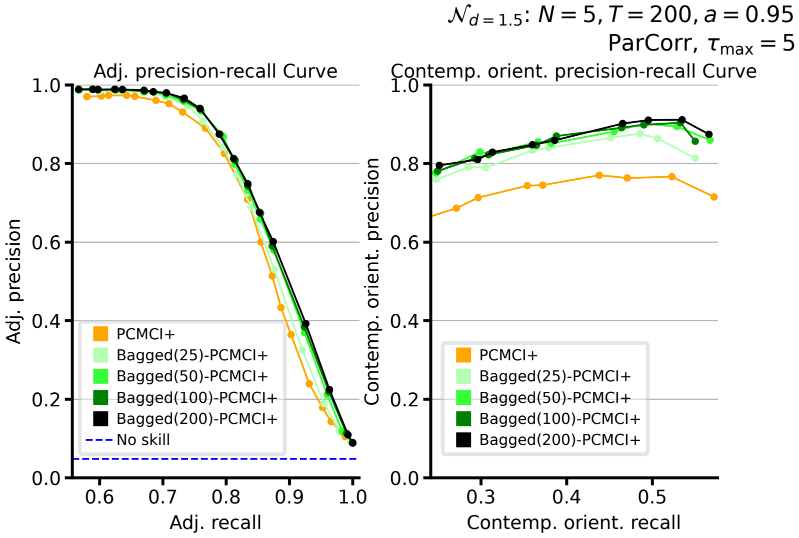

Figure 6: Precision-recall curves for adjacencies (left) and contemporaneous orientations (right) obtained by varying the significance level α_{PC} in PCMCI+ and Bagged-PCMCI+ for the model setup as shown in the header. Results are for PCMCI+ (orange line) and Bagged-PCMCI+ with different numbers of bootstrap replicas B (lines with different shades of green). Here $N=5$, $T=200$, and $a=0.95$.

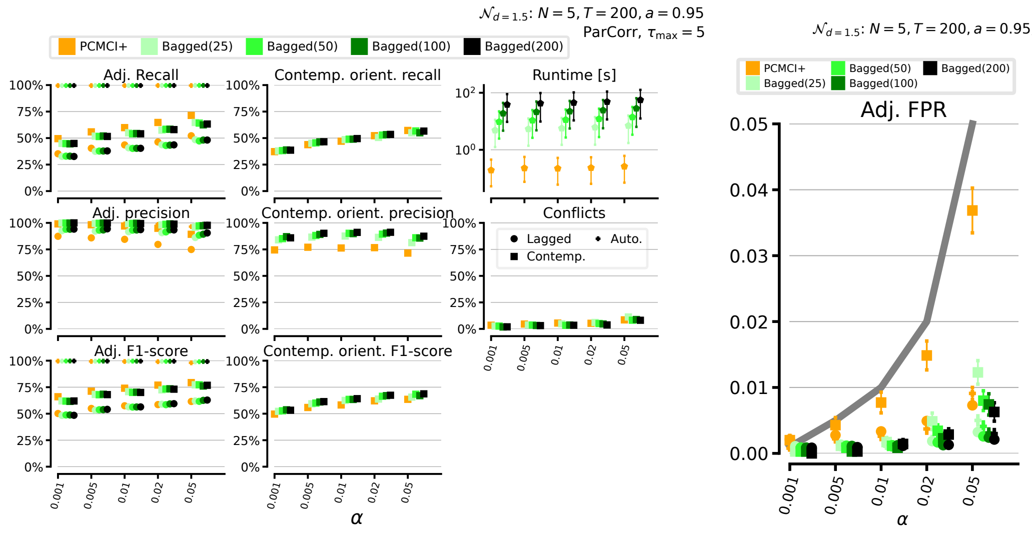

Figure 7: Numerical experiments with linear Gaussian setup for varying α_{PC} of PCMCI+. Here $N=5$, $T=200$, and $a=0.95$.

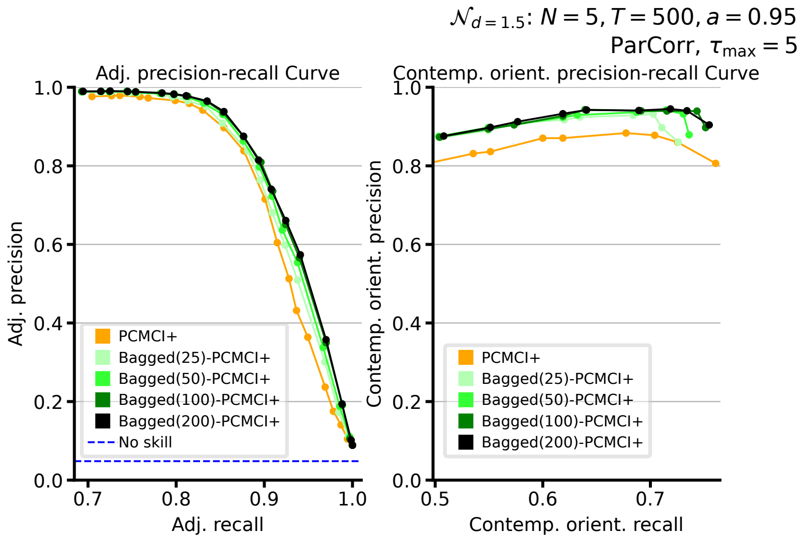

Figure 8: Precision-recall curves for adjacencies (left) and contemporaneous orientations (right) obtained by varying the significance level α_{PC} in PCMCI+ and Bagged-PCMCI+ for the model setup as shown in the header. Results are for PCMCI+ (orange line) and Bagged-PCMCI+ with different numbers of bootstrap replicas B (lines with different shades of green). Here $N=5$, $T=500$, and $a=0.95$.

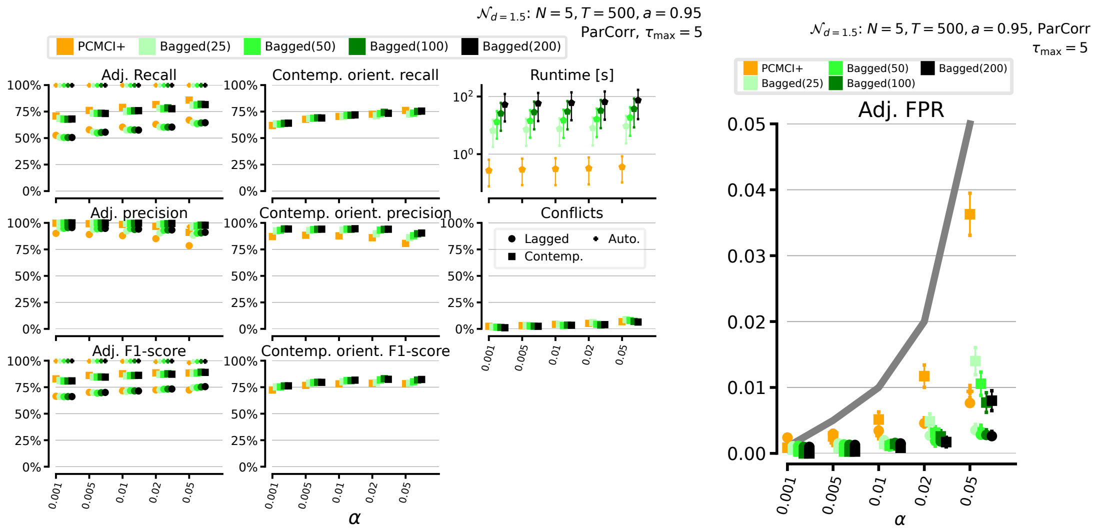

Figure 9: Numerical experiments with linear Gaussian setup for varying α_{PC} of PCMCI+. Here $N = 5$, $T = 500$, and $a = 0.95$.

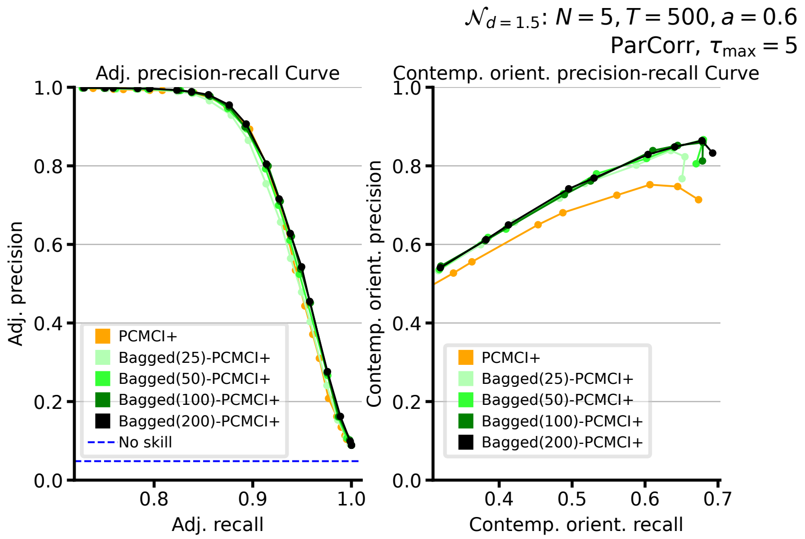

Figure 10: Precision-recall curves for adjacencies (left) and contemporaneous orientations (right) obtained by varying the significance level α_{PC} in PCMCI+ and Bagged-PCMCI+ for the model setup as shown in the header. Results are for PCMCI+ (orange line) and Bagged-PCMCI+ with different numbers of bootstrap replicas B (lines with different shades of green). Here $N = 5$, $T = 500$, and $a = 0.6$.

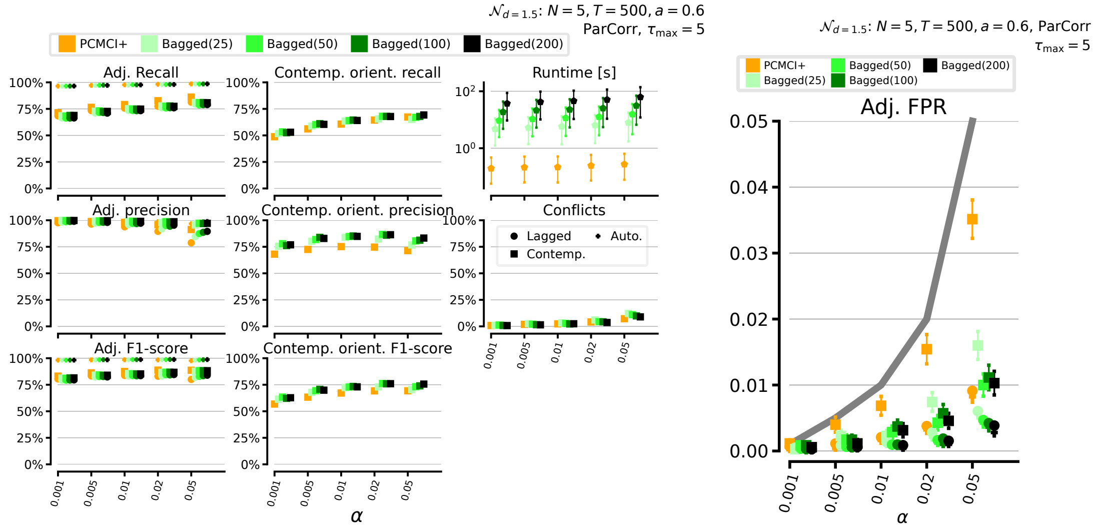

Figure 11: Numerical experiments with linear Gaussian setup for varying α_{PC} of PCMCI+. Here $N = 5$, $T = 500$, and $a = 0.6$.

C.2. Nonlinear and mixed noise experiments with Bagged-PCMCI+

In this subsection, we consider two model setups: linear with mixed noise (50% Gaussian and 50% Weibull), and nonlinear (50% linear and 50% nonlinear dependencies) with Gaussian noise. Results are shown in **Figure 12A** and **Figure 12B** respectively. Similar to the results of the linear Gaussian setup, the PR curves of Bagged-PCMCI+ systematically dominate the PR curves of the base PCMCI+ algorithm. This confirms the improvement of Bagged-PCMCI+ compared to PCMCI+ in terms of precision and recall for the nonlinear and mixed noise model setups.

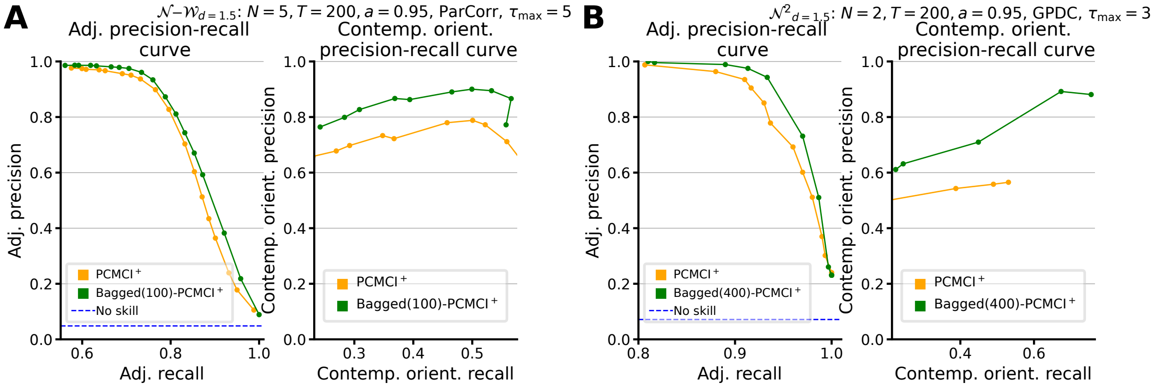

Figure 12: (**A**) Precision-recall curves for adjacencies (left) and contemporaneous orientations (right) obtained by varying the significance level α_{PC} in PCMCI+ (orange line) and Bagged-PCMCI+ with 100 bootstrap replicas (green line) for the model setup as shown in the top right. In particular, the model is **linear with mixed noise**: 50% Gaussian and 50% Weibull. (**B**) Precision-recall curves for adjacencies (left) and contemporaneous orientations (right) obtained by varying the significance level α_{PC} in PCMCI+ (orange line) and Bagged-PCMCI+ with 400 bootstrap replicas B (green line) for a **nonlinear** model ($f_i(x) = x + 5x^2 e^{-x^2/20}$, similar to Runge et al., (2020)) with Gaussian noise.

C.3. Additional Bagged-PCMCI+ confidence measure evaluation

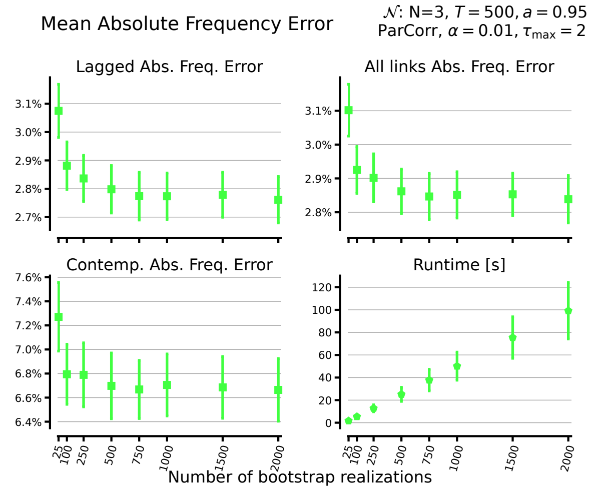

Figure 13: Mean absolute frequency error between estimated reference link frequencies and Bagged-PCMCI+ link frequencies for varying B and a linear Gaussian setup with parameters indicated at the top right.

Below, we evaluate our proposed confidence measures for varying significance level α_{PC} to study whether the bootstrapped confidence estimates approximate the estimated reference link frequencies for $\alpha_{PC} \to 0$ (**Figure 14**). To reduce computational time, the setup here was slightly modified compared to the main text. While we used $B = 1000$ and $L = 3$ (number of cross-links) in the main body of the paper, here we set $B = 250$ and $L = 5$. We vary α_{PC} from 0.01 to 10^{-5} to study the mean absolute error between the bootstrapped confidence estimates and the estimated reference link frequencies.

Table 1: Mean absolute error (in %) between the bootstrapped confidence estimates and the estimated reference link frequencies (extracted from Figure 14).

	$\alpha_{\mathrm{PC}} = 10^{-2}$	$\alpha_{\mathrm{PC}} = 10^{-4}$	$\alpha_{\mathrm{PC}} = 10^{-5}$
All absent links	2.7	1.0	1.3
All existing links	4.2	3.2	2.8
Contemp. absent links	4.7	2.3	2.5
Contemp. existing links	7.8	6.5	5.1
Lagged absent links	2.2	0.7	0.9
Lagged existing links	1.8	1.3	1.5

We summarize the results regarding mean absolute error in Tab. 1. There does seem to be a decrease in error from $\alpha_{\mathrm{PC}} = 10^{-2}$ to $\alpha_{\mathrm{PC}} = 10^{-4}$ across all types of link frequencies, while there are mixed results for $\alpha_{\mathrm{PC}} = 10^{-5}$. There is a visible recurrent positive bias for low values of the true frequencies (approximately 40-60%): The bootstrapped confidence measures tend to consistently overestimate the reference link frequencies for this range. More research is needed to clarify whether the bootstrap confidence estimates do approximate the reference link frequencies, or whether there are persistent biases. In this case, a question would be what this bias depends on (number of variables, graph structure, SCM properties, sample size, etc).

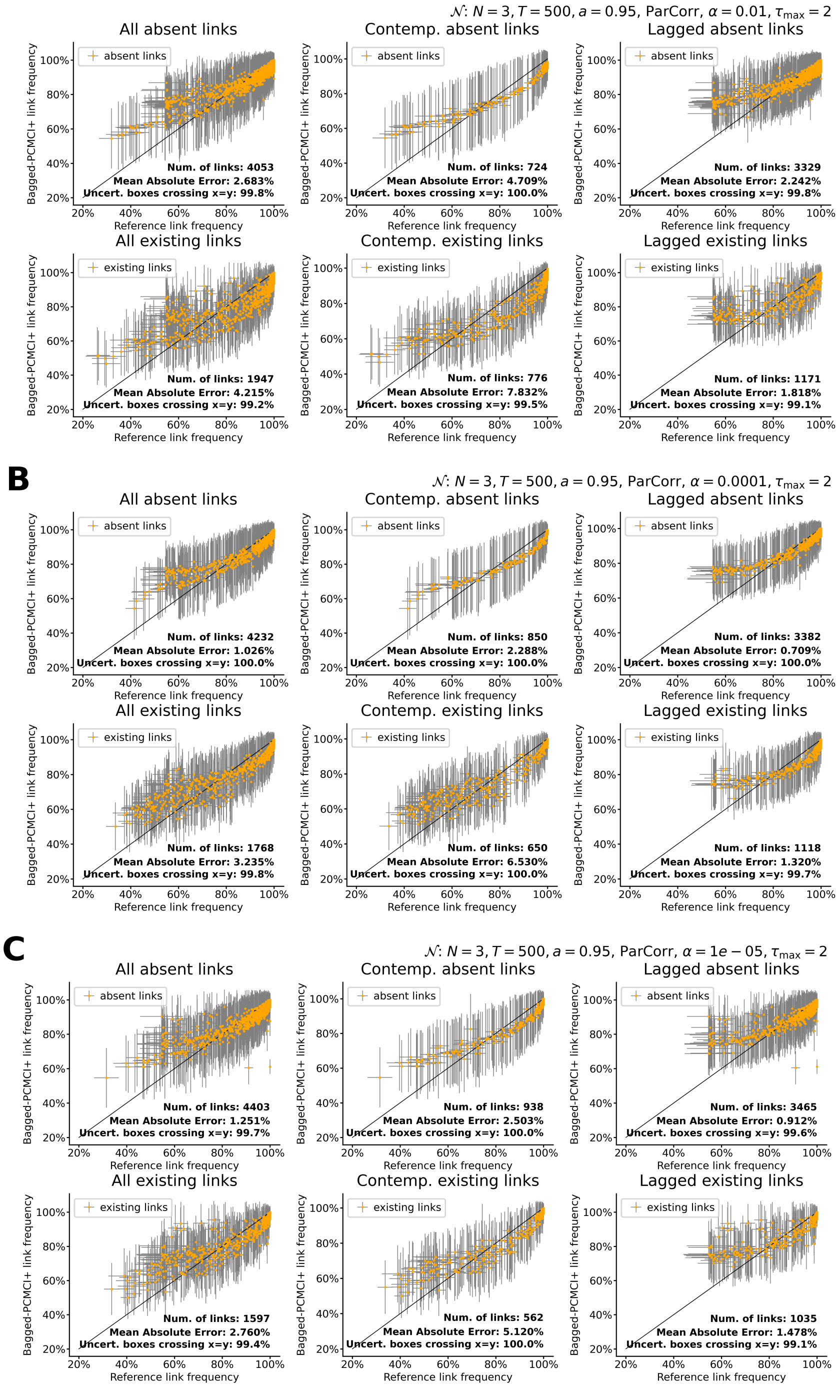

Figure 14: Estimated reference link frequencies against mean Bagged-PCMCI+ link frequencies ($B = 250$) for a linear Gaussian setup with parameters indicated at the top right. Grey bars indicate the one standard deviation error bars around the estimated value. The same model parameters are used in all three subfigures. Only the significance level α_{PC} changes: (**A**) $\alpha_{PC} = 0.01$, (**B**) $\alpha_{PC} = 10^{-4}$, (**C**) $\alpha_{PC} = 10^{-5}$.

Appendix D. Experiments for Bagged-PC

The numerical results with Bagged-PCMCI+ have shown that the bagging approach leads to enhanced precision-recall when paired with PCMCI+. To demonstrate that this conclusion not only applies to PCMCI+ but also to other causal discovery methods, we carried out further experiments, here with the PC algorithm. We combined our bagging approach with the PC algorithm (referred to as Bagged-PC) and compared its performance against the base PC algorithm. Both the base PC algorithm and Bagged-PC are adapted to time series as given in Runge (2020).

The results demonstrate that the gain using a bagging approach is similar here: Bagged-PC shows lower FPR and higher precision-recall compared to PC, especially for contemporaneous orientations (see **Figure 15** and **Figure 16**). Hence, our results show that combining a causal discovery method with our bagging approach can considerably improve the performance compared to the base causal discovery method, albeit at the expense of increased computational runtime (if not parallelized).

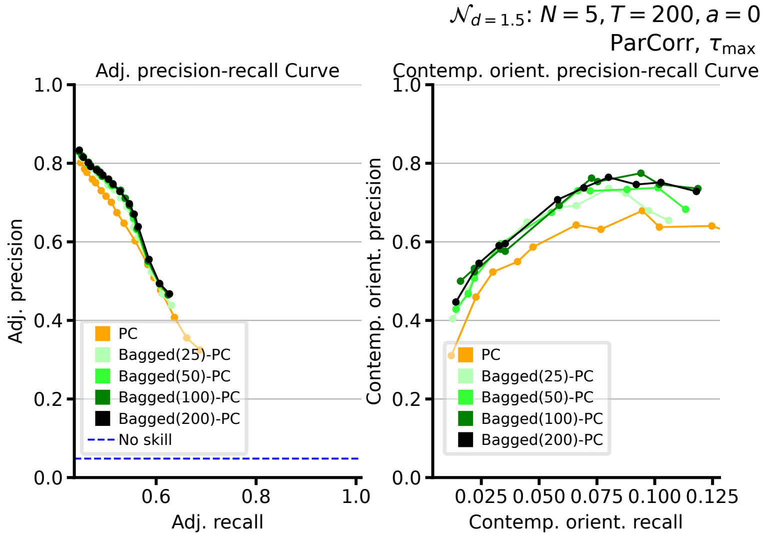

Figure 15: Precision-recall curves for adjacencies (left) and contemporaneous orientations (right) obtained by varying the significance level α_{PC} in **PC** and **Bagged-PC** for the model setup as shown in the header. Results are shown for PC (orange line) and Bagged-PC with different numbers of bootstrap replicas B (lines with different shades of green). Here the PC algorithm is adapted for time series data.

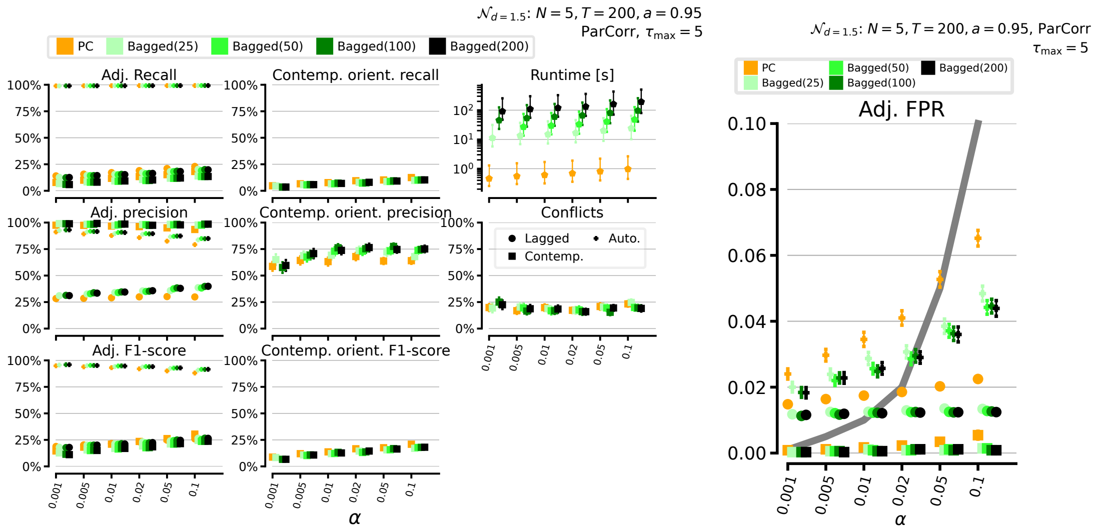

Figure 16: Numerical experiments with linear Gaussian setup for a varying α_{PC} of PC. Here $N = 5$, $T = 200$, and $a = 0.95$.

Appendix E. Experiments for Bagged-LPCMCI

We present here the experiment comparing Bagged-LPCMCI with the base LPCMCI algorithm (Gerhardus and Runge, 2020). LPCMCI is a further development of PCMCI+ that allows for unobserved (hidden/latent) variables. In terms of output, LPCMCI yields a partial ancestral graph (PAG, see Ali et al. (2009) and Zhang (2008)) while PCMCI+ outputs a completed partially directed acyclic graph (CPDAG, see Spirtes et al. (2000)). A PAG is a causal graph adapted to the presence of latent variables. Relative to a CPDAG, a PAG introduces the additional edge types $\leftarrow\!\circ$,$\circ\!\rightarrow$ and $\leftrightarrow$. Here, $\leftrightarrow$ indicates that there is a latent confounder causing both variables. The circle $\circ$ in $\leftarrow\!\circ$ and $\circ\!\rightarrow$ indicate the uncertainty about the correct causal direction. For example, $\circ\!\rightarrow$ could be $\rightarrow$ or $\leftrightarrow$ in the true graph. The design of LPCMCI is based on information-theoretical arguments showing that the effect sizes of (conditional) independence tests can often be increased if causal parents of the respective variables are included in the conditioning sets. To utilize this finding to improve recall, the algorithm intertwines the edge-removal phase (finding the graph's skeleton) with the edge-orientation phase and utilizes specific orientation rules to find causal parents and ancestors already before the final skeleton has been found.

The model setup considered here is linear with Gaussian noise, and results are shown in **Figure 17**. The results demonstrate the gain using a bagging approach: Bagged-LPCMCI shows lower FPR and higher precision-recall regarding contemporaneous orientations compared to the base LPCMCI. For this experiment, we have not found an improvement in the precision-recall of Bagged-LPCMCI regarding adjacencies. Hence, also for LPCMCI, our results show that combining a causal discovery method with our bagging approach can considerably improve the performance compared to the base causal discovery method, albeit at the expense of increased computational runtime (if not parallelized).

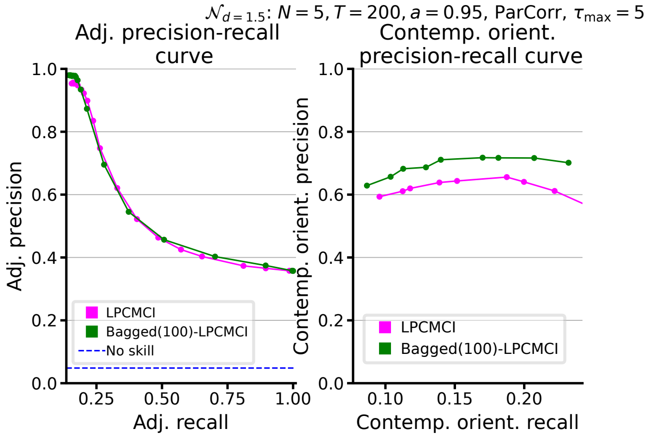

Figure 17: Precision-recall curves for adjacencies (left) and contemporaneous orientations (right) obtained by varying the significance level α_{PC} in **LPCMCI** (purple line) and **Bagged-LPCMCI** with 100 bootstrap replicas B (green line) for the model setup as shown in the top right. The hyperparameter k of LPCMCI is set to 4.

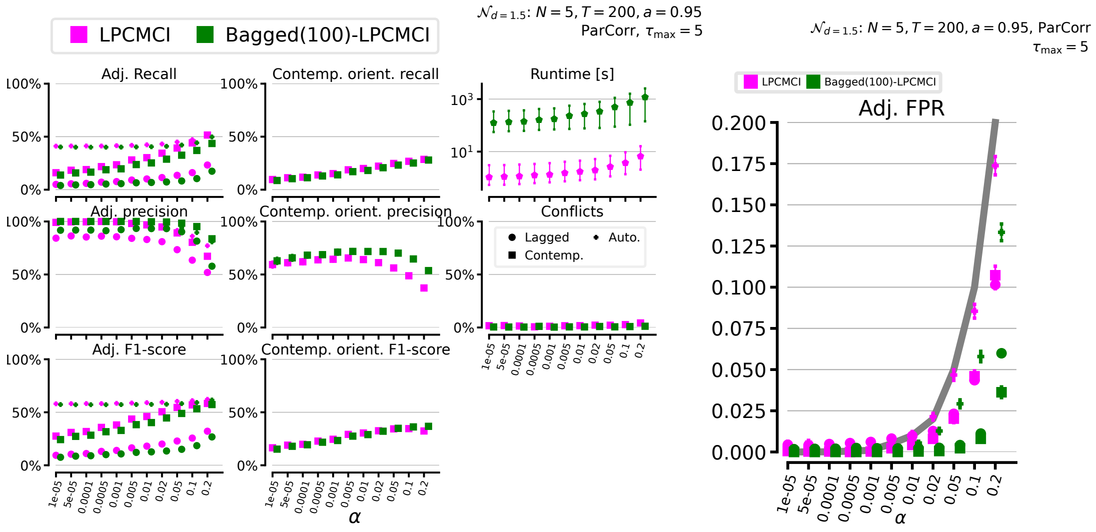

Figure 18: Numerical experiments with linear Gaussian setup for a varying α_{PC} of LPCMCI. Here $N = 5$, $T = 200$, and $a = 0.95$. The hyperparameter k of LPCMCI is set to 4.