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

The paper introduces a novel conditional independence (CI) based method for linear and nonlinear, lagged and contemporaneous causal discovery from observational time series in the causally sufficient case. Existing CI-based methods such as the PC algorithm and also common methods from other frameworks suffer from low recall and partially inflated false positives for strong autocorrelation which is an ubiquitous challenge in time series. The novel method, PCMCI$^+$, extends PCMCI [Runge et al., 2019b] to include discovery of contemporaneous links. PCMCI$^+$ improves the reliability of CI tests by optimizing the choice of conditioning sets and even benefits from autocorrelation. The method is order-independent and consistent in the oracle case. A broad range of numerical experiments demonstrates that PCMCI$^+$ has higher adjacency detection power and especially more contemporaneous orientation recall compared to other methods while better controlling false positives. Optimized conditioning sets also lead to much shorter runtimes than the PC algorithm. PCMCI$^+$ can be of considerable use in many real world application scenarios where often time resolutions are too coarse to resolve time delays and strong autocorrelation is present.

1 INTRODUCTION

A number of frameworks address the problem of causal discovery from observational data utilizing different assumptions. Next to Bayesian score-based methods [Chickering, 2002], classical Granger causality (GC) [Granger, 1969], and the more recent restricted structural causal models (SCM) framework [Peters et al., 2017, Spirtes and Zhang, 2016], conditional independence (CI) based network learning algorithms [Spirtes et al., 2000] form a main pillar. A main representative of the CI framework in the causally sufficient case (no unobserved common drivers) is the PC algorithm [Spirtes and Glymour, 1991]. 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 score- and SCM-based methods are more difficult to adapt to nonlinear high-dimensional causal discovery.

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). This forms the basis of GC which, however, cannot deal with contemporaneous links and suffers from the curse of dimensionality [Runge et al., 2019b]. SCM-based methods such as LiNGAM [Hyvärinen et al., 2010] and also CI-based methods [Runge et al., 2019b, Entner and Hoyer, 2010, Malinsky and Spirtes, 2018] have been adapted to the time series case. In [Moneta et al., 2011] GC is augmented by the PC algorithm. However, properties such as non-stationarity and especially autocorrelation can make causal discovery much less reliable.

Here I show that autocorrelation, an ubiquitous property of time series (e.g., temperature data), is especially detrimental and propose a novel CI-based method, PCMCI$^+$, that extends the PCMCI method from [Runge et al., 2019b] to also include discovery of contemporaneous links, which requires substantial changes. PCMCI$^+$ is based on two central ideas

Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), PMLR volume 124, 2020.

that deviate from the PC algorithm and the time-series adaptations of FCI in [Entner and Hoyer, 2010, Malinsky and Spirtes, 2018]: First, an edge removal phase is conducted separately for lagged and contemporaneous conditioning sets and the lagged phase uses much fewer CI tests. Secondly, and more importantly, PCMCI$^+$ optimizes the choice of conditioning sets for the individual CI tests to make them better calibrated under autocorrelation and increase detection power by utilizing the momentary conditional independence idea [Runge et al., 2019b]. The paper is structured as follows. Section 2 briefly introduces the problem and Sect. 3 describes the method and states theoretical results. Numerical experiments in Sect. 4 show that PCMCI$^+$ benefits from strong autocorrelation and yields much more adjacency detection power and especially more orientation recall for contemporaneous links while better controlling false positives at much shorter runtimes than the PC algorithm. A Supplementary Material (SM) contains proofs and further numerical experiments.

2 TIME SERIES CAUSAL DISCOVERY

2.1 PRELIMINARIES

We are interested in discovering *time series graphs* (e.g., [Runge, 2018a]) that can represent the temporal dependency structure underlying complex dynamical systems. Consider an underlying discrete-time structural causal process $\mathbf{X}_t = (X_t^1, \ldots, X_t^N)$ with

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

where f_j are arbitrary measurable functions with non-trivial dependencies on their 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 $\mathcal{G}$ (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 $\mathcal{P}(X_t^j) \subset \mathbf{X}_{t+1}^- = (\mathbf{X}_t, \mathbf{X}_{t-1}, \ldots)\backslash\{X_t^j\}$. We denote *lagged parents* by $\mathcal{P}_t^-(X_t^j) = \mathcal{P}(X_t^j) \cap \mathbf{X}_t^-$. A lagged ($\tau > 0$) or contemporaneous ($\tau = 0$) causal link $X_{t-\tau}^i \to X_t^j$ exists if $X_{t-\tau}^i \in \mathcal{P}(X_t^j)$. Throughout this work the graph $\mathcal{G}$ is assumed *acyclic* and the causal links *stationary* meaning that if $X_{t-\tau}^i \to X_t^j$ for some t, then $X_{t'-\tau}^i \to X_{t'}^j$ for all $t' \neq t$. Then we can always fix one variable at t and take $\tau \geq 0$. Note that the stationarity assumption may be relaxed. The graph is actually infinite in time, but in practice only considered up to some maximum time lag $\tau_{\max}$. 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 $k \in \{1, \ldots, m\}$ the link $X_t^{i+k-1} \to X_t^{i+k}$ occurs. We call X_t^i a *contemporaneous 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 CI-based causal discovery algorithm for the causally sufficient case and utilizes the Markov and Faithfulness assumptions as formally defined in Sect. S1. Adapted to time series (analogously to the methods for the latent case in [Entner and Hoyer, 2010, Malinsky and Spirtes, 2018]), 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 $\alpha_{\rm PC}$ (Alg. 2 with the PC option). For lagged links, time-order automatically provides orientations, while for contemporaneous links a collider phase (Alg. S2) and rule phase (Alg. S3) determine the orientation of links. CI-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 modification of PC that removes an undesired dependence on the order of variables, called PC-stable [Colombo and Maathuis, 2014]. These modifications also include either the *majority* or *conservative* [Ramsey et al., 2006] rule for handling ambiguous triples where separating sets are inconsistent, and conflicting links where different triples in the collider or orientation phase lead to conflicting link orientations. With the *conservative* rule the PC algorithm is consistent already under the weaker Adjacency Faithfulness condition [Ramsey et al., 2006]. Another approach for the time series case (considered in the numerical experiments) 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 AUTOCORRELATION

To illustrate the challenge of autocorrelation, in Fig. 1 we consider a linear example with lagged and contemporaneous ground truth links shown for the PCMCI$^+$ case (right panel). The PC algorithm (Alg. 2 with ParCorr CI test) starts by testing all unconditional independencies ($p = 0$). Here the coupled pairs (X^5, X^6) as well

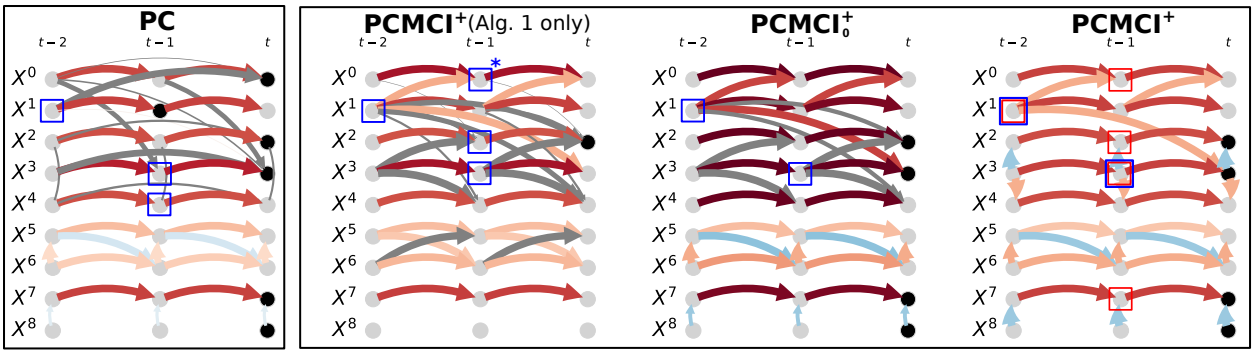

Figure 1: The curse and blessing of autocorrelation. Linear example of model (3) with ground truth links shown for the PCMCI$^+$ case (right panel). All autodependency coefficients are 0.95 (except 0.475 for $X^{5,6}$) and all cross-coupling coefficients are 0.4 ($\pm$ indicated by red/blue links). The graphs show true and false link detection rates as the link width (if > 0.06) for true (color indicating ParCorr) and incorrect links (grey) for the PC algorithm, Alg. 1, and the variants PCMCI$^+$ and PCMCI$^+_0$ as explained in the text (detection rates based on 500 realizations run at $\alpha_{\rm PC} = 0.01$ for $T = 500$).

as (X^7, X^8) are independent of the 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 GC framework. Due to the strong autocorrelation the remaining variables, on the other hand, are almost all adjacent to each other at multiple time lags in this iteration. In the next iteration, CI for all remaining links is tested conditional on all one-dimensional ($p = 1$) conditioning sets. Here the PC algorithm removes the true lagged link $X^1_{t-1} \to X^0_t$ (black dots) due to the incorrect CI result $X^1_{t-1} \perp\!\!\!\perp X^0_t | X^1_{t-2}$ (condition marked by blue box). Later this then leads to the false positive $X^1_{t-2} \to X^0_t$ (grey link) since X^1_{t-1} is not conditioned on. In a similar way the true link $X^1_{t-2} \to X^3_t$ is missed leading to the false positive $X^0_{t-1} \to X^3_t$. Further, the true contemporaneous link $X^2_t \circ\!\!-\!\!\circ X^3_t$ (similarly $X^3_t \circ\!\!-\!\!\circ X^4_t$) is removed when conditioning on $\mathcal{S} = (X^4_{t-1}, X^3_{t-1})$ (blue boxes), which leads to the false positive autodependencies at lag 2 for X^2_t, X^4_t, while the false autodependency $X^3_{t-2} \to X^3_t$ is due to missing $X^1_{t-2} \to X^3_t$. This illustrates the pattern of a cascade of false negative errors (missing links) leading to false positives in later stages of the PC algorithm.

What determines the removal of a true link in the finite sample case? Detection power depends on sample size, the significance level $\alpha_{\rm PC}$, the CI test dimension $(p+2)$, and effect size, e.g., the absolute ParCorr (population) value, here denoted $I(X^i_{t-\tau}; X^j_t | \mathcal{S})$ for some conditioning set $\mathcal{S}$. Within each p-iteration the sample size, $\alpha_{\rm PC}$, and the dimension are the same and a link will be removed if $I(X^i_{t-\tau}; X^j_t | \mathcal{S})$ falls below the $\alpha_{\rm PC}$-threshold for *any* considered $\mathcal{S}$. Hence, the overall minimum effect size $\min_{\mathcal{S}}[I(X^i_{t-\tau}; X^j_t | \mathcal{S})]$ determines whether a link is removed. The PC algorithm will iterate through *all* subsets of adjacencies such that this minimum can become very small. Low effect size can be understood as a low (causal) signal-to-noise ratio: Here $I(X^1_{t-1}; X^0_t | X^1_{t-2})$ is small since the signal X^1_{t-1} is reduced by conditioning on its autodependency X^1_{t-2} and the 'noise' in X^0_t is large due to its 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. On the other hand, the detection rate here is rather weak since, as above, the signal (link from X^8_t) is small compared to the noise (autocorrelation in X^7).

This illustrates the curse and blessing of autocorrelation. In summary, the PC algorithm often results in false negatives (low recall) and these then lead to false positives. Another reason for false positives are ill-calibrated tests: To correctly model the null distribution, each individual CI test would need to account for autocorrelation, which is difficult in a complex multivariate and potentially nonlinear setting [Runge, 2018a]. In the experiments we will see that the PC algorithm features 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 time-delayed feedbacks while an aggregated *summary graph* would contain a cyclic dependency and summary graph-based methods assuming acyclic graphs would not work. The orientation of the contemporaneous link $X^6_t \to X^5_t$ is achieved via rule R1 in the orientation phase of PC (Alg. S3).

3 PCMCI$^+$

3.1 ALGORITHM

The goal of PCMCI$^+$ is to optimize the choice of conditioning sets in CI tests in order 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 edge removal phase into a lagged and contemporaneous conditioning phase with much fewer CI tests and (2) utilizing the momentary conditional independence (MCI) test [Runge et al., 2019b] idea in the contemporaneous conditioning phase. Below, I explain the reasoning behind.

First, the goal of PC's skeleton phase is to remove all those adjacencies that are due to indirect paths and common causes by conditioning on subsets $\mathcal{S}$ of the variables' neighboring adjacencies in each iteration. Consider a variable X_t^j. If we test lagged adjacencies from nodes $X_{t-\tau}^i \in \mathbf{X}_t^-$ conditional on the whole past, i.e., $\mathcal{S} = \mathbf{X}_t^- \setminus \{X_{t-\tau}^i\}$, the only indirect adjacencies remaining are due to paths through contemporaneous parents of X_t^j. This is in contrast to conditioning sets on contemporaneous adjacencies which can also open up paths $X_t^j \to X_t^k \leftarrow X_{t-\tau}^i$ if X_t^k is conditioned on. One reason why the PC algorithm tests *all* combinations of subsets $\mathcal{S}$ is to avoid opening up such collider paths. Therefore, one approach would be to start by $\mathcal{S} = \mathbf{X}_t^- \setminus \{X_{t-\tau}^i\}$ and then iterate through contemporaneous conditions. A similar idea lies behind the combination of GC and the PC algorithm in [Moneta et al., 2011]. However, conditioning on large-dimensional conditioning sets strongly affects detection power [Runge et al., 2019b]. To avoid this, the lagged conditioning phase of PCMCI$^+$ (Alg. 1) tests all pairs $(X_{t-\tau}^i, X_t^j)$ for $\tau > 0$ conditional on only the *strongest* p adjacencies of X_t^j in each p-iteration without going through all p-dimensional subsets of adjacencies. This choice (i) improves the causal signal-to-noise ratio and recall since for a given test $X_{t-\tau}^i \perp\!\!\!\perp X_t^j \mid \mathcal{S}$ the 'noise' in X_t^j due to other lagged adjacencies is conditioned out, (ii) leads to fewer CI tests further improving recall, and (iii) speeds up the skeleton phase. We denote the lagged adjacency set resulting from Alg. 1 as $\widehat{\mathcal{B}}_t^-(X_t^j)$. Lemma 1 in Sect. 3.2 states that the only remaining indirect adjacencies in $\widehat{\mathcal{B}}_t^-(X_t^j)$ 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 $\widehat{\mathcal{B}}_t^-(X_t^j)$ for all X_t^j. Algorithm 2 tests all (unordered lagged and ordered contemporaneous) adjacent pairs $(X_{t-\tau}^i, X_t^j)$ and iterates through contemporaneous conditions $\mathcal{S} \subseteq \mathcal{A}_t(X_t^j)$ with the MCI test

$$X_{t-\tau}^i \perp\!\!\!\perp X_t^j \mid \mathcal{S}, \widehat{\mathcal{B}}_t^-(X_t^j) \backslash \{X_{t-\tau}^i\}, \widehat{\mathcal{B}}_{t-\tau}^-(X_{t-\tau}^i). \quad (2)$$

The condition on $\widehat{\mathcal{B}}_t^-(X_t^j)$ blocks paths through lagged parents and the advantage of the additional conditioning on $\widehat{\mathcal{B}}_{t-\tau}^-(X_{t-\tau}^i)$ is discussed in the following. We denote the variant without the condition on $\widehat{\mathcal{B}}_{t-\tau}^-(X_{t-\tau}^i)$ as PCMCI$_0^+$. Both versions are followed by the collider orientation phase (Alg. S2) and rule orientation phase (Alg. S3) which are deferred to the SM since they are equivalent to the PC algorithm with the modification that the additional CI tests in the collider phase for the conservative or majority rule are also based on the test (2).

We now discuss PCMCI$_0^+$ and PCMCI$^+$ on the example in Fig. 1. Algorithm 1 tests $X_{t-1}^1 \to X_t^0$ 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 (causal signal-to-noise ratio) is much larger than for the condition on $\mathcal{S} = \{X_{t-2}^1\}$ which lead to the removal of $X_{t-1}^1 \to X_t^0$ in the PC algorithm. In Sect. 3.2 we elaborate more rigorously on effect size. In the example $\widehat{\mathcal{B}}_t^-(X_t^2)$ 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, Alg. 2 with the PCMCI$_0^+$ option then recovers all lagged links (3rd panel), but it still the misses contemporaneous adjacencies $X_t^2 \circ\!\!-\!\!\circ X_t^3$ and $X_t^3 \circ\!\!-\!\!\circ X_t^4$ 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 for PCMCI$_0^+$ in the $p = 0$ loop, like the original PC algorithm, will test *ordered* contemporaneous pairs. Hence, first $X_t^2 \circ\!\!-\!\!\circ X_t^3$ conditional on $\widehat{\mathcal{B}}_t^-(X_t^3)$ and, if the link is not removed, $X_t^3 \circ\!\!-\!\!\circ X_t^2$ conditional on $\widehat{\mathcal{B}}_t^-(X_t^2)$. Here $X_t^2 \circ\!\!-\!\!\circ X_t^3$ is removed conditional on $\widehat{\mathcal{B}}_t^-(X_t^3)$ (indicated by blue boxes in the panel) because $I(X_t^2; X_t^3 | \widehat{\mathcal{B}}_t^-(X_t^3))$ falls below the significance threshold.

The second central idea of PCMCI$^+$ is to improve the effect size of CI tests for contemporaneous links by conditioning on *both* lagged adjacencies $\widehat{\mathcal{B}}_t^-$ in the CI test (2) (see blue and red boxes in Fig. 1 right panel). At least for the initial phase $p = 0$ one can prove that for non-empty $\widehat{\mathcal{B}}_t^-$ the effect size of the PCMCI$^+$ CI test is always strictly larger than that of the PCMCI$_0^+$ test (Thm. 4). I conjecture that this similarly holds for PCMCI$^+$ vs. the PC algorithm. Higher effect size leads to higher recall

and PCMCI$^+$ now recovers all lagged as well as contemporaneous links and also correctly removes the lagged false positives that PCMCI$^+_0$ obtains. Also the contemporaneous coupled pair (X^7, X^8) is now much better detected since the MCI effect size $I(X^7_t; X^8_t|X^7_{t-1})$ is larger than $I(X^7_t; X^8_t)$, one of the two PCMCI$^+_0$ and PC algorithm effect sizes tested here.

Another advantage, discussed in [Runge et al., 2019b] is that PCMCI$^+$ CI tests are better calibrated, in contrast to PCMCI$^+_0$ and PC algorithm tests, since the condition on both parents removes 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 $\widehat{\mathcal{B}}^-_{t-\tau}(X^i_{t-\tau})$ can only reduce effect size (see [Runge et al., 2012]). This is the cost of avoiding inflated false positives.

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. S2, S3. The other advantage of PCMCI$^+$ compared to the PC algorithm is a much faster and, as numerical examples show, also much less variable runtime.

The full algorithm is detailed in pseudo-code Algorithms 1,2,S2,S3 with differences to PC and PCMCI$^+_0$ indicated. Note that pairs $(X^i_{t-\tau}, X^j_t)$ in lines 5 and 6 of Alg. 2 are ordered for $\tau = 0$ and unordered for $\tau > 0$. One can construct (rather conservative) p-values for the skeleton adjacencies $(X^i_{t-\tau}, X^j_t)$ by taking the maximum p-value over all CI tests conducted in Alg. 2. A link strength can be defined corresponding to the test statistic value of the maximum p-value. Based on the PC stable variant, PCMCI$^+$ is fully order-independent. Here shown is the majority-rule implementation of the collider phase, the version without handling ambiguous triples and for the conservative rule are detailed in Alg. S2. Note that the tests in the collider phase also use the CI tests (2).

Like other CI-based methods, PCMCI$^+$ has the free parameters $\alpha_{\rm PC}$, $\tau_{\max}$, and the choice of the CI test. $\alpha_{\rm PC}$ can be chosen based on cross-validation or an information criterion (implemented in `tigramite`). $\tau_{\max}$ should be larger or equal to the maximum true time lag of any parent and can in practice also be chosen based on model selection. However, the numerical experiments indicate that, in contrast to GC, a too large $\tau_{\max}$ does not degrade performance much and $\tau_{\max}$ can also be chosen based on the lagged dependence functions, see [Runge et al., 2019b]. PCMCI$^+$ can flexibly be combined with different CI tests for nonlinear causal discovery, and for different variable types (discrete or continuous, univariate or multivariate).

The computational complexity of PCMCI$^+$ strongly depends on the network structure. 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 Alg. 1 only has polynomial complexity [Runge et al., 2019b] and Alg. 2 only iterates through contemporaneous conditioning sets, hence the worst-case exponential complexity only applies to N and not to $N\tau_{\max}$.

3.2 THEORETICAL RESULTS

This section states asymptotic consistency, finite sample order-independence, and further results regarding effect size and false positive control. 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 CI tests (oracle). In the present time series context we also assume stationarity and time-order and that the maximum time lag $\tau_{\max} \geq \tau^{\mathcal{P}}_{\max}$, where $\tau^{\mathcal{P}}_{\max}$ is the maximum time lag of any parent in the SCM* (1)*. Furthermore, we rule out* selection variables *and* measurement error.

Definitions of these assumptions, adapted from [Spirtes et al., 2000] to the time series context, are in Sect. S1 and all proofs are in Sect. S2. We start with the following lemma.

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

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

This establishes that the conditions $\widehat{\mathcal{B}}^-_t(X^j_t)$ estimated in the first phase of PCMCI$^+$ will suffice to block all lagged confounding paths that do not go through contemporaneous links. This enables to prove the soundness of Alg. 2, even though Alg. 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 $\mathcal{G}^*$ denotes the skeleton of the time series graph.*

Algorithm 1 (PCMCI$^+$ / PCMCI$_0^+$ lagged skeleton phase)

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

for all X_t^j in $\mathbf{X}_t$ **do**
  Initialize $\widehat{\mathcal{B}}_t^-(X_t^j)$=$\mathbf{X}_t^-$=$(\mathbf{X}_{t-1}, \ldots, \mathbf{X}_{t-\tau_{\max}})$ and $I^{\min}(X_{t-\tau}^i, X_t^j) = \infty \ \forall\, X_{t-\tau}^i \in \widehat{\mathcal{B}}_t^-(X_t^j)$
  Let $p = 0$
  while any $X_{t-\tau}^i \in \widehat{\mathcal{B}}_t^-(X_t^j)$ satisfies $|\widehat{\mathcal{B}}_t^-(X_t^j)\backslash\{X_{t-\tau}^i\}| \geq p$ **do**
    for all $X_{t-\tau}^i$ in $\widehat{\mathcal{B}}_t^-(X_t^j)$ satisfying $|\widehat{\mathcal{B}}_t^-(X_t^j)\backslash\{X_{t-\tau}^i\}| \geq p$ **do**
      $\mathcal{S}$ = first p variables in $\widehat{\mathcal{B}}_t^-(X_t^j) \setminus \{X_{t-\tau}^i\}$
      (p-value, I) $\leftarrow$ CI($X_{t-\tau}^i$, X_t^j, $\mathcal{S}$)
      $I^{\min}(X_{t-\tau}^i, X_t^j) = \min(|I|, I^{\min}(X_{t-\tau}^i, X_t^j))$
      if p-value $> \alpha_{\rm PC}$ **then** mark $X_{t-\tau}^i$ for removal
    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
    Let $p = p + 1$
return $\widehat{\mathcal{B}}_t^-(X_t^j)$ for all X_t^j in $\mathbf{X}_t$

Algorithm 2 (PCMCI$^+$ / PCMCI$_0^+$ contemporaneous skeleton phase / PC full skeleton phase)

Require: Time series dataset $\mathbf{X} = (X^1, \ldots, X^N)$, max. time lag $\tau_{\max}$, significance threshold $\alpha_{\rm PC}$, $\mathrm{CI}(X, Y, \mathbf{Z})$, PCMCI$^+$ / PCMCI$_0^+$: $\widehat{\mathcal{B}}_t^-(X_t^j)$ for all X_t^j in $\mathbf{X}_t$

PCMCI$^+$ / PCMCI$_0^+$: 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 $X_t^i \neq X_t^j \in \mathbf{X}_t$
PC: Form fully connected time series graph $\mathcal{G}$ with lagged and contemporaneous links
PCMCI$^+$ / PCMCI$_0^+$: Initialize contemporaneous adjacencies $\widehat{\mathcal{A}}(X_t^j) := \widehat{\mathcal{A}}_t(X_t^j) = \{X_t^i \neq X_t^j \in \mathbf{X}_t : X_t^i \circ\!\!-\!\!\circ X_t^j \text{ in } \mathcal{G}\}$
PC: Initialize full adjacencies $\widehat{\mathcal{A}}(X_t^j)$ for all (lagged and contemporaneous) links in $\mathcal{G}$
Initialize $I^{\min}(X_{t-\tau}^i, X_t^j) = \infty$ for all links in $\mathcal{G}$
Let $p = 0$
while any adjacent pairs $(X_{t-\tau}^i, X_t^j)$ for $\tau \geq 0$ in $\mathcal{G}$ satisfy $|\widehat{\mathcal{A}}(X_t^j)\backslash\{X_{t-\tau}^i\}| \geq p$ **do**
  Select new adjacent pair $(X_{t-\tau}^i, X_t^j)$ for $\tau \geq 0$ satisfying $|\widehat{\mathcal{A}}(X_t^j)\backslash\{X_{t-\tau}^i\}| \geq p$
  while $(X_{t-\tau}^i, X_t^j)$ are adjacent in $\mathcal{G}$ and not all $\mathcal{S} \subseteq \widehat{\mathcal{A}}(X_t^j)\backslash\{X_{t-\tau}^i\}$ with $|\mathcal{S}| = p$ have been considered **do**
    Choose new $\mathcal{S}\subseteq\widehat{\mathcal{A}}(X_t^j)\backslash\{X_{t-\tau}^i\}$ with $|\mathcal{S}|$=p
    PCMCI$^+$: Set $\mathbf{Z}$=$(\mathcal{S}, \widehat{\mathcal{B}}_t^-(X_t^j)\backslash\{X_{t-\tau}^i\}, \widehat{\mathcal{B}}_{t-\tau}^-(X_{t-\tau}^i))$
    PCMCI$_0^+$: Set $\mathbf{Z}$=$(\mathcal{S}, \widehat{\mathcal{B}}_t^-(X_t^j)\backslash\{X_{t-\tau}^i\})$
    PC: Set $\mathbf{Z}$=$\mathcal{S}$
    (p-value, I) $\leftarrow$ CI($X_{t-\tau}^i, X_t^j, \mathbf{Z}$)
    $I^{\min}(X_{t-\tau}^i, X_t^j) = \min(|I|, I^{\min}(X_{t-\tau}^i, X_t^j))$
    if p-value $> \alpha_{\rm PC}$ **then**
      Delete link $X_{t-\tau}^i \to X_t^j$ for $\tau > 0$ (or $X_t^i \circ\!\!-\!\!\circ X_t^j$ for $\tau = 0$) from $\mathcal{G}$
      Store (unordered) $\mathrm{sepset}(X_{t-\tau}^i, X_t^j) = \mathcal{S}$
  Let $p = p + 1$
  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
return $\mathcal{G}$, sepset

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 (Alg. S2) and rule orientation phase (Alg. S3) can be restricted as follows: In the collider orientation phase only unshielded triples $X_{t-\tau}^i \to X_t^k \circ\!\!-\!\!\circ X_t^j$ (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 $X_{t-\tau}^i \to X_t^k \circ\!\!-\!\!\circ X_t^j$ where $(X_{t-\tau}^i, X_t^j)$ are not adjacent,*

for orientation rule R2 triples $X_t^i \to X_t^k \to X_t^j$ with $X_t^i \circ\!\!-\!\!\circ X_t^j$, and for orientation rule R3 pairs of triples $X_t^i \circ\!\!-\!\!\circ X_t^k \to X_t^j$ and $X_t^i \circ\!\!-\!\!\circ X_t^l \to X_t^j$ 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 parts of separating sets are relevant for the collider phase.

Theorem 2 (PCMCI$^+$ is complete)**.** *PCMCI$^+$ (Algorithms 1,2,S2,S3) when used with the conservative rule for orienting colliders in Alg. S2 returns the correct CPDAG under Assumptions 1. Under standard Faithfulness also PCMCI$^+$ when used with the majority rule or the standard orientation rule is complete.*

Also the proof of order-independence follows straightforwardly from the proof in [Colombo and Maathuis, 2014]. Of course, order independence does not apply to time-order.

Theorem 3 (Order independence)**.** *Under Assumptions 1 PCMCI$^+$ with the conservative or majority rule in Alg. S2 is independent of the order of variables $(X^1, \ldots, X^N)$.*

Next, we consider effect size. 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, I will base my argument in an information-theoretic framework and consider the conditional mutual information as a general test statistic, denoted I. In Alg. 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$^+$ oracle case CI tests in Alg. 2 for $p = 0$ for contemporaneous true links $X_t^i \to X_t^j \in \mathcal{G}$ have an effect size that is always greater than that of the PCMCI$_0^+$ CI tests, i.e., $I(X_t^i; X_t^j | \widehat{\mathcal{B}}_t^-(X_t^j), \widehat{\mathcal{B}}_t^-(X_t^i)) > \min(I(X_t^i; X_t^j | \widehat{\mathcal{B}}_t^-(X_t^j)), I(X_t^i; X_t^j | \widehat{\mathcal{B}}_t^-(X_t^i)))$ if both X_t^i and X_t^j have parents that are not shared with the other.*

I conjecture that this result holds similarly for $p > 0$ and also that PCMCI$^+$ has greater effect sizes than the PC algorithm since the latter iterates over *all* subsets of adjacencies and, hence, the minimum is taken generally over an even larger set leading to even smaller effect sizes. For lagged links the effect size of the PCMCI$^+$ tests is always smaller (or equal) than that of the PCMCI$_0^+$ tests (see [Runge et al., 2012]).

Last, we discuss false positive control. While the effect size result regards detection power, in the following I give a mathematical intuition why the MCI tests are better calibrated than the PC algorithm CI tests and control false positives below the expected significance level. Lemma 1 implies that even though Alg. 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, I conjecture, is key to achieve well-calibrated CI tests in Alg. 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 numerical experiments below we can see that the PC algorithm has inflated false positive for high autocorrelation, while PCMCI$^+$ well controls false positives, but a formal proof of correct false positive control for this challenging nonlinear, high-dimensional setting is beyond the scope of this paper.

4 NUMERICAL EXPERIMENTS

We consider a number of typical challenges [Runge et al., 2019a], contemporaneous and time lagged causal dependencies, strong autocorrelation, large number of variables and considered time lags, different noise distributions and nonlinearity, in the following additive variant of model (1):

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

for $j \in \{1, \ldots, N\}$. Autocorrelations a_j are uniformly drawn from $[\max(0, a - 0.3), a]$ for a as indicated in Fig. 2 and η^j is *i.i.d.* and follows a zero-mean Gaussian $\mathcal{N}$ or Weibull $\mathcal{W}$ (scale parameter 2) distribution (depending on setup) with standard deviation drawn from $[0.5, 2]$. In addition to autodependency links, for each model $L = \lfloor 1.5 \cdot N \rfloor$ (except for $N = 2$ with $L = 1$) cross-links are chosen whose functional dependencies are linear or $f_i(x) = f^{(2)}(x) = (1 + 5xe^{-x^2/20})x$ (depending on setup), with $f^{(2)}$ designed to yield more stationary dynamics. 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 $[1, 5]$. 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 Gaussian, linear mixed noise (among the N

variables: 50% Gaussian, 50% Weibull), and nonlinear mixed noise (50% linear, 50% $f^{(2)}(x)$; 66% Gaussian, 34% Weibull).

For the linear model setups we consider the PC algorithm and PCMCI$^+$ in the majority-rule variant with ParCorr and compare these with GCresPC [Moneta et al., 2011], a combination of GC with PC applied to residuals, and a autoregressive model version of LiNGAM [Hyvärinen et al., 2010], a representative of the SCM framework (implementation details in Sect. S4). For the LiNGAM implementation I could not find a way to set a significance level and used the LASSO option which prunes 'non-active' links to zero. Both GCresPC and LiNGAM assume linear dependencies and LiNGAM also non-Gaussianity. For the nonlinear setup the PC algorithm and PCMCI$^+$ are implemented with 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.

Performance is evaluated as follows: True (TPR) and false positive rates (FPR, shown to evaluate false positive control, not applicable to LiNGAM) for adjacencies are distinguished between lagged cross-links ($i \neq j$), contemporaneous, and autodependency links. Due to time order, lagged links (and autodependencies) are automatically oriented. Contemporaneous orientation precision is measured as the fraction of correctly oriented links ($\circ\!\!-\!\!\circ$ or $\rightarrow$) among all estimated adjacencies, and recall as the fraction of correct orientations among all true contemporaneous links. Further shown is the fraction of conflicting links among all detected contemporaneous adjacencies (not applicable to LiNGAM). All metrics (and their std. errors) are computed across all estimated graphs from 500 realizations of model (3) at time series length T. The average runtimes were evaluated on Intel Xeon Platinum 8260. In Fig. 2 results for the linear Gaussian setup with default model parameters $N = 5$, $T = 500$, $a = 0.95$ and method parameters $\tau_{\max} = 5$ and $\alpha = 0.01$ (not applicable to LiNGAM) are shown. Each of the four panels shows results for varying one of a, N, T, $\tau_{\max}$. The insets show ANOVA statistics $r \pm \bar{\Delta} r$ [per unit], where r is the performance metric at the leftmost parameter on the x-axis (a, N, T, $\tau_{\max}$, respectively) and $\bar{\Delta} r$ denotes the average change per parameter unit. In the adjacency subplots the statistics refer to lagged links.

Figure 2A demonstrates that the TPR of PCMCI$^+$ and GCresPC for contemporaneous links is stable even under high autocorrelation while PC and LiNGAM show strong declines. Since LiNGAM has no α_{PC} for FPR-control we focus on its relative changes rather than absolute performance. Lagged TPR decreases strongly for PC while the other methods are more robust. FPR is well-controlled for PCMCI$^+$ while PC and slightly also GCresPC show inflated lagged FPR for high autocorrelation. LiNGAM features a strong increase of lagged FPR. These adjacency results translate into higher contemporaneous orientation recall for PCMCI$^+$ which increases with autocorrelation, while it decreases for all other methods. GCresPC has steady low recall since it does not use lagged links in the orientation phase. Except for GCresPC, all methods have increasing precision with PCMCI$^+$ and PC outperforming LiNGAM. PCMCI$^+$ shows almost no conflicts while PC's conflicts increase with autocorrelation until low power reduces them again. Finally, runtimes are almost constant for GCresPC and LiNGAM, while they increase for PCMCI$^+$ and much stronger for PC.

Figure 2B shows that PCMCI$^+$ and GCresPC have the highest TPR for increasing number of variables N, especially for contemporaneous links. FPR is well controlled only for PCMCI$^+$ while PC has 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 while inflated FPR appears for GCresPC. PCMCI$^+$ has more than twice as much contemporaneous recall compared to the other methods and is almost not affected by higher N. Orientation precision is decreasing for all methods (except PC) with a higher decrease for PCMCI$^+$. Runtime is increasing at a much smaller rate for PCMCI$^+$ compared to PC, which also has a very high runtime variability across the different model realizations. LiNGAM and especially GCresPC are fastest.

PCMCI$^+$, GCresPC, and LiNGAM benefit similarly and PC less so for increasing sample size regarding TPR (Fig. 2C). FPR is still not controlled for PC for large sample sizes, lagged FPR increases for GCresPC. PCMCI$^+$ shows the highest increases in contemporaneous recall and precision. Runtime increases are moderate compared to PC, conflicts decrease.

Last, Fig. 2D shows that all methods are relatively robust to large maximum time lags $\tau_{\max}$ (beyond the true max. time lag 5) for the considered sample size $T = 500$. Contemporaneous FPR and runtime increase for PC.

In the SM further results are shown. For too large $N\tau_{\max}$ (relative to T) GCresPC and LiNGAM (despite LASSO-regularization) sharply drop in performance. For the linear mixed noise setup (Fig. S2) results are almost unchanged for all methods except for LiNGAM for which recall and precision rise, as expected. Recall is then higher than PCMCI$^+$ for low autocorrelation, but still much lower for high autocorrelation and large N or $\tau_{\max}$, at similar precision.

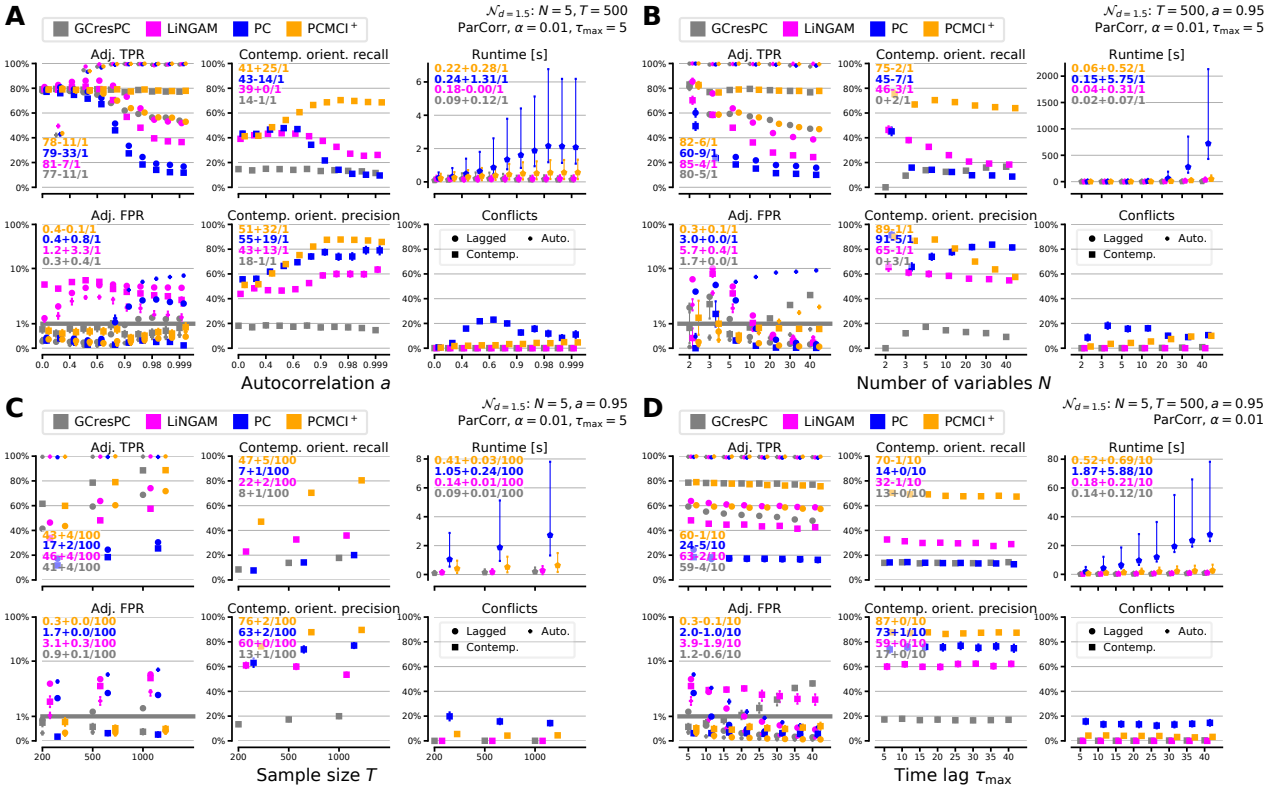

Figure 2: Numerical experiments with linear Gaussian setup for varying (**A**) autocorrelation strength a (**B**) number of variables N (**C**) sample size T and (**D**) maximum time lag $\tau_{\max}$. All remaining setup parameters indicated in the top right. Errorbars show std. errors or the 90% range (for runtime). The insets show ANOVA statistics.

In the nonlinear mixed noise setup (Fig. S3), the difference between PC and PCMCI$^+$ is similar. We observe slight FPR inflation for high autocorrelation. GPDC seems to not work well in high-dimensional, highly autocorrelated settings. Runtime for GPDC compared to ParCorr is orders of magnitude longer, especially for PC. Further figures in the SM show many combinations of a, N, T, $\tau_{\max}$ and $\alpha_{\rm PC}$ for the model setups and demonstrate that the above findings are robust.

5 CONCLUSIONS

PCMCI$^+$ improves the reliability of CI tests by optimizing the choice of conditioning sets and yields much higher recall, well-controlled false positives, and faster runtime than the original PC algorithm for highly autocorrelated time series, while maintaining similar performance for low autocorrelation. The algorithm well exploits sparsity in high-dimensional settings and can flexibly be combined with different CI tests for nonlinear causal discovery, and for different variable types (discrete or continuous, univariate or multivariate). Autocorrelation is actually key to increase contemporaneous orientation recall since it creates triples $X^i_{t-1} \to X^i_t \circ\!\!-\!\!\circ X^j_t$ that can often be oriented while an isolated link $X^i_t \circ\!\!-\!\!\circ X^j_t$ stays undirected in the Markov equivalence class, a drawback of CI-based methods. If the data is at least non-Gaussian, a SCM method like LiNGAM can exploit this property and recover directionality in such cases. Still, we saw that LiNGAM suffers from large autocorrelation. PCMCI$^+$ is available as part of the *tigramite* Python package at `https://github.com/jakobrunge/tigramite`. A next step will be to extend the present ideas to an algorithm accounting for latent confounders and to explore combinations between SCM-based methods and PCMCI$^+$. The numerical results will be contributed to the causality benchmark platform `www.causeme.net` [Runge et al., 2019a] to facilitate a further expanded method evaluation.

Acknowledgments

DKRZ provided computational resources (grant no. 1083). 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.

[Hyvärinen et al., 2010] Hyvärinen, A., Zhang, K., Shimizu, S., and Hoyer, P. O. (2010). Estimation of a structural vector autoregression model using non-gaussianity. *J. Mach. Learn. Res.*, 11:1709–1731.

[Malinsky and Spirtes, 2018] Malinsky, D. and Spirtes, P. (2018). Causal structure learning from multivariate time series in settings with unmeasured confounding. In *Proc. of 2018 ACM SIGKDD Work. on Causal Discovery*, pages 23–47.

[Moneta et al., 2011] Moneta, A., Chlaß, N., Entner, D., and Hoyer, P. (2011). Causal search in structural vector autoregressive models. In *NIPS Mini-Symp. 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 *Proc. 22nd Conf. on Uncertainty in Art. Int.*, 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 Comm.*, 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. *Phys. Rev. E*, 86(6):061121.

[Runge et al., 2019b] Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., and Sejdinovic, D. (2019b). Detecting and quantifying 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, MA.

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

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

S1 Definitions

The following definitions are adaptations of the standard assumptions of causal discovery to the time series case. Here we consider the causally sufficient case and assume that all variables $\mathbf{X} = (X^1, \dots, X^N)$ of the underlying SCM (1) are observed. Additionally, we assume that the maximum PCMCI$^+$ time lag $\tau_{\max} \geq \tau^{\mathcal{P}}_{\max}$, where $\tau^{\mathcal{P}}_{\max}$ is the maximum time lag of any parent in the SCM (1).

Definition S1 (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 $X^j_t \in \mathbf{X}_t$ every non-descendent of X^j_t in $\mathcal{G}$ is independent of X^j_t given the parents $\mathcal{P}(X^j_t)$. In particular, $\mathbf{X}^-_t \backslash \mathcal{P}(X^j_t) \perp\!\!\!\perp X^j_t \mid \mathcal{P}(X^j_t)$ since all variables in $\mathbf{X}^-_t$ are non-descendants of X^j_t by time-order.*

Note that for the SCM (1) with independent noise terms the Causal Markov Condition is automatically fulfilled.

Definition S2 (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 $X^i_{t-\tau}, X^j_t, \mathcal{S} \in \mathbf{X}^-_{t+1}$ with $\tau > 0$*

$$\begin{aligned} &X^i_{t-\tau} \perp\!\!\!\perp X^j_t \mid \mathcal{S} \;\Rightarrow\; X^i_{t-\tau} \to X^j_t \notin \mathcal{G} \\ &X^i_{t-\tau} \to X^j_t \in \mathcal{G} \;\Rightarrow\; X^i_{t-\tau} \not\perp\!\!\!\perp X^j_t \mid \mathcal{S} \;\textit{(contrapositive)} \end{aligned}$$

and with $\tau = 0$

$$\begin{aligned} &X^i_t \perp\!\!\!\perp X^j_t \mid \mathcal{S} \;\Rightarrow\; X^i_t \circ\!\!-\!\!\circ X^j_t \notin \mathcal{G} \\ &X^i_t \circ\!\!-\!\!\circ X^j_t \in \mathcal{G} \;\Rightarrow\; X^i_t \not\perp\!\!\!\perp X^j_t \mid \mathcal{S} \;\textit{(contrapositive)}\,. \end{aligned}$$

Furthermore, the variables fulfill the (standard) Faithfulness Condition iff for $\tau \geq 0$

$$\begin{aligned} &X^i_{t-\tau} \perp\!\!\!\perp X^j_t \mid \mathcal{S} \;\Rightarrow\; X^i_{t-\tau} \bowtie X^j_t \mid \mathcal{S} \\ &X^i_{t-\tau} \not\bowtie X^j_t \mid \mathcal{S} \;\Rightarrow\; X^i_{t-\tau} \not\perp\!\!\!\perp X^j_t \mid \mathcal{S} \;\textit{(contrapositive)}\,. \end{aligned}$$

S2 Proofs

S2.1 Proof of Lemma 1

We first consider the following Lemma:

Lemma S1. *Algorithm 1 returns a superset of lagged parents under Assumptions 1, i.e., $\mathcal{P}^-_t(X^j_t) \subseteq \widehat{\mathcal{B}}^-_t(X^j_t)$ for all X^j_t in $\mathbf{X}_t$.*

Proof. We need to show that for arbitrary $(X^i_{t-\tau}, X^j_t)$ with $\tau > 0$ we have $X^i_{t-\tau} \notin \widehat{\mathcal{B}}^-_t(X^j_t) \;\Rightarrow\; X^i_{t-\tau} \notin \mathcal{P}^-_t(X^j_t)$. Algorithm 1 removes $X^i_{t-\tau}$ from $\widehat{\mathcal{B}}^-_t(X^j_t)$ iff $X^i_{t-\tau} \perp\!\!\!\perp X^j_t \mid \mathcal{S}$ for some $\mathcal{S} \subseteq \widehat{\mathcal{B}}^-_t(X^j_t)\backslash\{X^i_{t-\tau}\}$ in the iterative CI tests. Then Adjacency Faithfulness directly implies that $X^i_{t-\tau}$ is not adjacent to X^j_t and in particular $X^i_{t-\tau} \notin \mathcal{P}^-_t(X^j_t)$. □

With this step we can prove Lemma 1.

Proof. The lemma states that under Assumptions 1 with Adjacency Faithfulness replaced by standard Faithfulness Alg. 1 for all $X^j_t \in \mathbf{X}_t$ returns $\widehat{\mathcal{B}}^-_t(X^j_t) = \bigcup_{X^i_t \in \{X^j_t\} \cup \mathcal{C}_t(X^j_t)} \mathcal{P}^-_t(X^i_t)$ where $\mathcal{C}_t(X^j_t)$ denotes the contemporaneous ancestors of X^j_t. We need to show that for arbitrary $X^i_{t-\tau}, X^j_t \in \mathbf{X}^-_{t+1}$ with $\tau > 0$: (1) $X^i_{t-\tau} \notin \widehat{\mathcal{B}}^-_t(X^j_t) \;\Rightarrow\; X^i_{t-\tau} \notin \bigcup_{X^i_t \in \{X^j_t\} \cup \mathcal{C}_t(X^j_t)} \mathcal{P}^-_t(X^i_t)$ and (2) $X^i_{t-\tau} \in \widehat{\mathcal{B}}^-_t(X^j_t) \;\Rightarrow\; X^i_{t-\tau} \in \bigcup_{X^i_t \in \{X^j_t\} \cup \mathcal{C}_t(X^j_t)} \mathcal{P}^-_t(X^i_t)$.

Ad 1) Algorithm 1 removes $X^i_{t-\tau}$ from $\widehat{\mathcal{B}}^-_t(X^j_t)$ iff $X^i_{t-\tau} \perp\!\!\!\perp X^j_t \mid \mathcal{S}$ for some $\mathcal{S} \subseteq \widehat{\mathcal{B}}^-_t(X^j_t)\backslash\{X^i_{t-\tau}\}$ in the iterative CI tests. Then standard Faithfulness implies that $X^i_{t-\tau} \bowtie X^j_t \mid \mathcal{S}$ and in particular $X^i_{t-\tau} \notin \mathcal{P}^-_t(X^j_t)$, as proven already in Lemma S1 under the weaker Adjacency Faithfulness Condition. To show that $X^i_{t-\tau} \notin \bigcup_{X^i_t \in \{X^j_t\} \cup \mathcal{C}_t(X^j_t)} \mathcal{P}^-_t(X^i_t)$ we note that $\mathcal{S} \subseteq \widehat{\mathcal{B}}^-_t(X^j_t)\backslash\{X^i_{t-\tau}\}$ does not include any contemporaneous conditions and, hence, all contemporaneous directed paths from contemporaneous ancestors of X^j_t are open and also paths from parents of those ancestors are open. If $X^i_{t-\tau} \in \bigcup_{X^i_t \in \mathcal{C}_t(X^j_t)} \mathcal{P}^-_t(X^i_t)$, by the contraposition of standard Faithfulness we should observe $X^i_{t-\tau} \not\perp\!\!\!\perp X^j_t \mid \mathcal{S}$. Then the fact that on the contrary we observe $X^i_{t-\tau} \perp\!\!\!\perp X^j_t \mid \mathcal{S}$ implies that $X^i_{t-\tau} \notin \bigcup_{X^i_t \in \mathcal{C}_t(X^j_t)} \mathcal{P}^-_t(X^i_t)$.

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

S2.2 Proof of Theorem 1

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

Ad (1): Algorithm 2 deletes a link $X^i_{t-\tau} \star\!\!-\!\!\star X^j_t$ from $\widehat{\mathcal{G}^*}$ iff $X^i_{t-\tau} \perp\!\!\!\perp X^j_t \mid \mathcal{S}, \widehat{\mathcal{B}}^-_t(X^j_t)\backslash\{X^i_{t-\tau}\}, \widehat{\mathcal{B}}^-_{t-\tau}(X^i_{t-\tau})$ for some $\mathcal{S} \subseteq \widehat{\mathcal{A}}_t(X^j_t)$ in the iterative CI tests with $\widehat{\mathcal{B}}^-_t(X^j_t)$ estimated in Alg. 1. $\widehat{\mathcal{A}}_t(X^j_t)$ denotes the contemporaneous adjacencies. Then Adjacency Faithfulness directly implies that $X^i_{t-\tau}$ is not adjacent to X^j_t: $X^i_{t-\tau} \star\!\!-\!\!\star X^j_t \notin \mathcal{G}^*$.

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

$\mathcal{S} = \mathcal{P}_t(X_t^j)$ and Alg. 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 $\widehat{\mathcal{G}^*}$. □

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 $t = t_j$. We only need to consider cases with $t_i, t_k \leq t$. If $t_k > t_j$, the triple is oriented already by time order and the case $t_i > t_j$ is symmetric.

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 $X_{t_k}^k \to X_t^j$ 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 $\tau = t - t_i$ ($\tau_{\max} \geq \tau \geq 0$), leaving the two cases of unshielded triples $X_{t-\tau}^i \to X_t^k \circ\!\!-\!\!\circ X_t^j$ (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. Since X_t^k is contemporaneous to X_t^j, this restriction implies that only contemporaneous parts of 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 $X_{t-\tau}^i \to X_t^k$ that were not oriented by the collider phase (or by time order). This leaves $X_{t-\tau}^i \to X_t^k \circ\!\!-\!\!\circ X_t^j$ where $\tau_{\max} \geq \tau \geq 0$.

For rule R2 the original PC algorithm considers $X_{t_i}^i \to X_{t_k}^k \to X_t^j$ 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 $X_t^i \to X_t^k \to X_t^j$ with $X_t^i \circ\!\!-\!\!\circ X_t^j$.

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

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 Alg. S2. We need to show that all separating sets estimated in Alg. S2 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 Alg. S2 for unshielded triples $X_{t-\tau}^i \to X_t^k \circ\!\!-\!\!\circ X_t^j$ (for $\tau > 0$) or $X_t^i \circ\!\!-\!\!\circ X_t^k \circ\!\!-\!\!\circ X_t^j$ (for $\tau = 0$) the separating sets among subsets of contemporaneous neighbors of X_t^j and, if $\tau = 0$, of X_t^i, are correct. Algorithm S2 tests $X_{t-\tau}^i \perp\!\!\!\perp X_t^j \mid \mathcal{S}, \widehat{\mathcal{B}}_t^-(X_t^j) \backslash \{X_{t-\tau}^i\}, \widehat{\mathcal{B}}_{t-\tau}^-(X_{t-\tau}^i)$ for all $\mathcal{S} \subseteq \widehat{\mathcal{A}}_t(X_t^j) \backslash \{X_{t-\tau}^i\}$ and for all $\mathcal{S} \subseteq \widehat{\mathcal{A}}_t(X_t^i) \backslash \{X_t^j\}$ (if $\tau = 0$). Since PCMCI$^+$ is sound, all adjacency information is correct and since all CI tests are assumed correct, all information on separating sets is correct. 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 separating set 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 $X_{t-\tau}^i \perp\!\!\!\perp X_t^j \mid \mathcal{S}, \widehat{\mathcal{B}}_t^-(X_t^j) \backslash \{X_{t-\tau}^i\}, \widehat{\mathcal{B}}_{t-\tau}^-(X_{t-\tau}^i)$ for all $\mathcal{S} \subseteq \widehat{\mathcal{A}}_t(X_t^j) \backslash \{X_{t-\tau}^i\}$. And again, since PCMCI$^+$ is sound, all adjacency information is correct and since all CI tests are assumed correct, all information on separating sets is correct, from which the completeness for this case follows. □

S2.5 Proof of Theorem 3

Proof. Order-independence follows straightforwardly from sticking to the PC algorithm version in [Colombo and Maathuis, 2014]. In particular, Alg. 1 and Alg. 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 (Alg. S2) and rule orientation phase (Alg. S3) are order-independent by marking triples with inconsistent separating sets as ambiguous and consistently marking conflicting link orientations by $\times\!-\!\times$. □

S2.6 Proof of Theorem 4

Proof. The theorem states that under Assumptions 1 the effect size for the PCMCI$^+$ oracle case CI tests in Alg. 2 for $p = 0$ for contemporaneous true links $X^i_t \to X^j_t \in \mathcal{G}$ is greater than that of PCMCI$^+_0$: $I(X^i_t; X^j_t|\widehat{\mathcal{B}}^-_t(X^j_t), \widehat{\mathcal{B}}^-_t(X^i_t)) > \min(I(X^i_t; X^j_t|\widehat{\mathcal{B}}^-_t(X^j_t)),\, I(X^i_t; X^j_t|\widehat{\mathcal{B}}^-_t(X^i_t)))$ if both X^i_t and X^j_t have parents that are not shared with the other. We will use an information-theoretic framework here and consider the conditional mutual information.

To prove this statement, we denote by $\mathcal{B}_i = \widehat{\mathcal{B}}^-_t(X^i_t) \setminus \widehat{\mathcal{B}}^-_t(X^j_t)$ the lagged conditions of X^i_t that are not already contained in those of X^j_t and, correspondingly, $\mathcal{B}_j = \widehat{\mathcal{B}}^-_t(X^j_t) \setminus \widehat{\mathcal{B}}^-_t(X^i_t)$. Since both X^i_t and X^j_t have parents that are not shared with the other and we assume the oracle case, both these sets are non-empty. Further, we denote the common lagged conditions as $\mathcal{B}_{ij} = \widehat{\mathcal{B}}^-_t(X^j_t) \cap \widehat{\mathcal{B}}^-_t(X^i_t)$ and make use of the following conditional independencies, which hold by the Markov assumption: (1) $\mathcal{B}_i \perp\!\!\!\perp X^j_t|\mathcal{B}_j, \mathcal{B}_{ij}, X^i_t$ and (2) $\mathcal{B}_j \perp\!\!\!\perp X^i_t|\mathcal{B}_i, \mathcal{B}_{ij}$. We first prove that, given a contemporaneous true link $X^i_t \to X^j_t \in \mathcal{G}$, $I(X^i_t; X^j_t|\mathcal{B}_{ij}, \mathcal{B}_j) > I(X^i_t; X^j_t|\mathcal{B}_{ij}, \mathcal{B}_i)$ by using the following two ways to apply the chain rule of conditional mutual information:

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

and

$$
\begin{aligned}
&I(X^i_t, \mathcal{B}_i; X^j_t, \mathcal{B}_j|\mathcal{B}_{ij}) = \\
&= I(\mathcal{B}_i; X^j_t, \mathcal{B}_j|\mathcal{B}_{ij}) + I(X^i_t; X^j_t, \mathcal{B}_j|\mathcal{B}_{ij}\mathcal{B}_i) \\
&= I(\mathcal{B}_i; \mathcal{B}_j|\mathcal{B}_{ij}) + \underbrace{I(\mathcal{B}_i; X^j_t|\mathcal{B}_{ij}, \mathcal{B}_j)}_{>0 \;\text{ since } X^i_t \to X^j_t} \\
&\quad + I(X^i_t; X^j_t|\mathcal{B}_{ij}, \mathcal{B}_i) + \underbrace{I(X^i_t; \mathcal{B}_j|\mathcal{B}_{ij}, \mathcal{B}_i, X^j_t)}_{>0 \;\text{ since } X^i_t \to X^j_t}
\end{aligned} \tag{S2}
$$

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

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

$$
\begin{aligned}
&I(X^i_t; X^j_t, \mathcal{B}_j|\mathcal{B}_{ij}, \mathcal{B}_i) = \\
&= I(X^i_t; X^j_t|\mathcal{B}_{ij}, \mathcal{B}_i) + \underbrace{I(X^i_t; \mathcal{B}_j|\mathcal{B}_{ij}, \mathcal{B}_i, X^j_t)}_{>0 \;\text{ since } X^i_t \to X^j_t} \qquad \text{(S3)} \\
&= \underbrace{I(X^i_t; \mathcal{B}_j|\mathcal{B}_{ij}, \mathcal{B}_i)}_{=0 \;\text{(Markov)}} + I(X^i_t; X^j_t|\mathcal{B}_{ij}, \mathcal{B}_i, \mathcal{B}_j) \qquad \text{(S4)}
\end{aligned}
$$

□

S3 Further pseudo code

Algorithms S2 and S3 detail the pseudo-code for the PCMCI$^+$ / PCMCI$^+_0$ / PC collider phase with different collider rules and the orientation phase.

Algorithm S2 (Detailed PCMCI$^+$ / PCMCI$^+_0$ / PC collider phase with different collider rules)

Require: $\mathcal{G}$ and sepset from Alg. 2, rule = {'none', 'conservative', 'majority'}, time series dataset $\mathbf{X} = (X^1, \ldots, X^N)$, significance threshold $\alpha_{\rm PC}$, CI(X, Y, $\mathbf{Z}$), PCMCI$^+$ / PCMCI$^+_0$: $\widehat{\mathcal{B}}^-_t(X^j_t)$ for all X^j_t in $\mathbf{X}_t$

for all unshielded triples $X^i_{t-\tau} \to X^k_t \circ\!\!-\!\!\circ X^j_t$ ($\tau > 0$) or $X^i_t \circ\!\!-\!\!\circ X^k_t \circ\!\!-\!\!\circ X^j_t$ ($\tau = 0$) in $\mathcal{G}$ where $(X^i_{t-\tau}, X^j_t)$ are not adjacent **do**
  if rule = 'none' **then**
    if X^k_t is not in sepset$(X^i_{t-\tau}, X^j_t)$ **then**
      Orient $X^i_{t-\tau} \to X^k_t \circ\!\!-\!\!\circ X^j_t$ ($\tau > 0$) or $X^i_{t-\tau} \circ\!\!-\!\!\circ X^k_t \circ\!\!-\!\!\circ X^j_t$ ($\tau = 0$) as $X^i_{t-\tau} \to X^k_t \leftarrow X^j_t$
  else
    PCMCI$^+$ / PCMCI$^+_0$: Define contemporaneous adjacencies $\widehat{\mathcal{A}}(X^j_t) = \widehat{\mathcal{A}}_t(X^j_t) = \{X^i_t \neq X^j_t \in \mathbf{X}_t : X^i_t \circ\!\!-\!\!\circ X^j_t \text{ in } \mathcal{G}\}$
    PC: Define full adjacencies $\widehat{\mathcal{A}}(X^j_t)$ for all (lagged and contemporaneous) links in $\mathcal{G}$
    for all for all $\mathcal{S} \subseteq \widehat{\mathcal{A}}(X^j_t) \backslash \{X^i_{t-\tau}\}$ and for all $\mathcal{S} \subseteq \widehat{\mathcal{A}}(X^i_t) \backslash \{X^j_t\}$ (if $\tau = 0$) **do**
      Evaluate CI$(X^i_{t-\tau}, X^j_t, \mathbf{Z})$ with
      PCMCI$^+$: $\mathbf{Z}=(\mathcal{S}, \widehat{\mathcal{B}}^-_t(X^j_t) \backslash \{X^i_{t-\tau}\}, \widehat{\mathcal{B}}^-_{t-\tau}(X^i_{t-\tau}))$
      PCMCI$^+_0$: $\mathbf{Z}=(\mathcal{S}, \widehat{\mathcal{B}}^-_t(X^j_t) \backslash \{X^i_{t-\tau}\})$
      PC: $\mathbf{Z}=\mathcal{S}$
      Store all subsets $\mathcal{S}$ with p-value $> \alpha_{\rm PC}$ as separating subsets
    if no separating subsets are found **then**
      Mark triple as ambiguous
    else
      Compute fraction n_k of separating subsets that contain X^k_t
      if rule = 'conservative' **then**
        Orient triple as collider if n_k=0, leave unoriented if n_k=1, and mark as ambiguous if $0<n_k<1$
      else if rule = 'majority' **then**
        Orient triple as collider if $n_k<0.5$, leave unoriented if $n_k>0.5$, and mark as ambiguous if $n_k=0.5$
      Mark links in $\mathcal{G}$ with conflicting orientations as $\times\!\!-\!\!\times$
return $\mathcal{G}$, sepset, ambiguous triples, conflicting links

Algorithm S3 (Detailed PCMCI$^+$ / PCMCI$^+_0$ / PC rule orientation phase)

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

while any unambiguous triples suitable for rules R1-R3 are remaining **do**
    Apply rule R1 (orient unshielded triples that are not colliders):
    for all unambiguous triples $X^i_{t-\tau} \to X^k_t \circ\!\!-\!\!\circ X^j_t$ where $(X^i_{t-\tau}, X^j_t)$ are not adjacent **do**
        Orient as $X^i_{t-\tau} \to X^k_t \to X^j_t$
        Mark links with conflicting orientations as $\times\!-\!\times$
    Apply rule R2 (avoid cycles):
    for all unambiguous triples $X^i_t \to X^k_t \to X^j_t$ with $X^i_t \circ\!\!-\!\!\circ X^j_t$ **do**
        Orient as $X^i_t \to X^j_t$
        Mark links with conflicting orientations as $\times\!-\!\times$
    Apply rule R3 (orient unshielded triples that are not colliders and avoid cycles):
    for all pairs of unambiguous triples $X^i_t \circ\!\!-\!\!\circ X^k_t \to X^j_t$ and $X^i_t \circ\!\!-\!\!\circ X^l_t \to X^j_t$ where (X^k_t, X^l_t) are not adjacent and $X^i_t \circ\!\!-\!\!\circ X^j_t$ **do**
        Orient as $X^i_t \to X^j_t$
        Mark links with conflicting orientations as $\times\!-\!\times$
return $\mathcal{G}$, conflicting links

S4 Implementation details

In the linear and nonlinear numerical experiments PCMCI$^+$ is compared with the PC algorithm, both implemented with the appropriate CI test (ParCorr for the linear case, GPDC for the nonlinear case). For the linear numerical experiments we additionally consider representatives from two further frameworks: GCresPC, a combination of GC with PC applied to residuals, and an autoregressive model version of LiNGAM [Hyvärinen et al., 2010], a representative of the SCM framework. Their implementations are as follows.

S4.1 LiNGAM

For LiNGAM the code was taken from `https://github.com/cdt15/lingam` which provides a class `VARLiNGAM`. The method was called follows:

```
Input: data, tau_max

model = lingam.VARLiNGAM(lags=tau_max, criterion=None, prune=True)
model.fit(data)
val_matrix = model.adjacency_matrices_.transpose(2,1,0)
graph = (val_matrix != 0.).astype('int')

Output: graph
```

The causal graph `graph` encodes the causal relations in an array of shape `(N, N, tau_max + 1)`. The option `criterion=None` just ignores the optional automatic selection of `lags`, which is here set to the same `tau_max` for all methods. I could not find a way to obtain p-values in the `VARLiNGAM` implementation, but with the parameter setting `prune=True` the resulting adjacency matrices are regularized with an adaptive LASSO approach using the BIC criterion to find the optimal regularization hyper-parameter (`sklearn.LassoLarsIC(criterion='bic')`). Non-zero adjacencies were then evaluated as causal links. Note that all other methods can be intercompared at different $\alpha_{\rm PC}$ levels while for comparison against LiNGAM we focus on its relative changes rather than absolute performance.

S4.2 GCresPC

There was no code available for the method proposed in [Moneta et al., 2011]. The present implementation first fits a VAR model up to $\tau_{\max}$ and applies the PC algorithm on the residuals. To remove spurious lagged links (due to contemporaneous paths), the PC algorithm was additionally run on significant lagged and contemporaneous links, but the orientation phase was restricted to contemporaneous links, as proposed in [Moneta et al., 2011]. The following Python pseudo-code utilizes functionality from the `tigramite` package, `numpy`, and `statsmodels`:

```
Input: data, tau_max, alpha
import functions/classes ParCorr, PCMCI, DataFrame from tigramite

graph = np.zeros((N, N, tau_max + 1))

# 1. Estimate lagged adjacencies (to be updated in step 3.)
tsamodel = tsa.var.var_model.VAR(data)
results = tsamodel.fit(maxlags=tau_max, trend='nc')
pvalues = results.pvalues
values = results.coefs
residuals = results.resid

lagged_parents = significant lagged links at alpha

# 2. Run PC algorithm on residuals (with tau_max=0)
pcmci = PCMCI(dataframe=DataFrame(residuals), cond_ind_test=ParCorr())
pcmcires = pcmci.run_pcalg(pc_alpha=alpha,
```

```
   tau_min=0,
   tau_max=0)

# Update contemporaneous graph
graph[:,:,0] = pcmcires['graph'][:,:,0]

# 3. Run PC algorithm on significant lagged and contemporaneous adjacencies
# to remove spurious lagged links due to contemporaneous parents

selected_links = lagged_parents + significant contemporaneous adjacencies
pcmci = PCMCI(dataframe=DataFrame(data), cond_ind_test=ParCorr())
pcmcires = pcmci.run_pcalg(selected_links=selected_links,
   pc_alpha=alpha,
   tau_min=0,
   tau_max=tau_max)

# Update lagged part of graph
graph[:,:,1:] = pcmcires['graph'][:,:,1:]

Output: graph
```

Note that the contemporaneous graph structure in `graph` comes only from applying the PC algorithm to the residuals and, hence, does not utilize triples containing lagged adjacencies. Step 3 is necessary to remove spurious lagged links due to contemporaneous parents. The output of GCresPC depends on $\alpha_{\rm PC}$ as for PCMCI$^+$ and the PC algorithm.

S5 Further numerical experiments

Next to repeating the overview figure for the linear Gaussian model setup from the main text in Fig. S1, in Fig. S2 we show the linear mixed noise setup, and in Fig. S3 the nonlinear mixed noise setup. The remaining pages contain results of further numerical experiments that evaluate different a, N, T, $\tau_{\max}$ and $\alpha_{\rm PC}$ for the linear model setups. All results and more will be contributed to the causality benchmark platform `www.causeme.net` [Runge et al., 2019a] to facilitate a further expanded method evaluation.

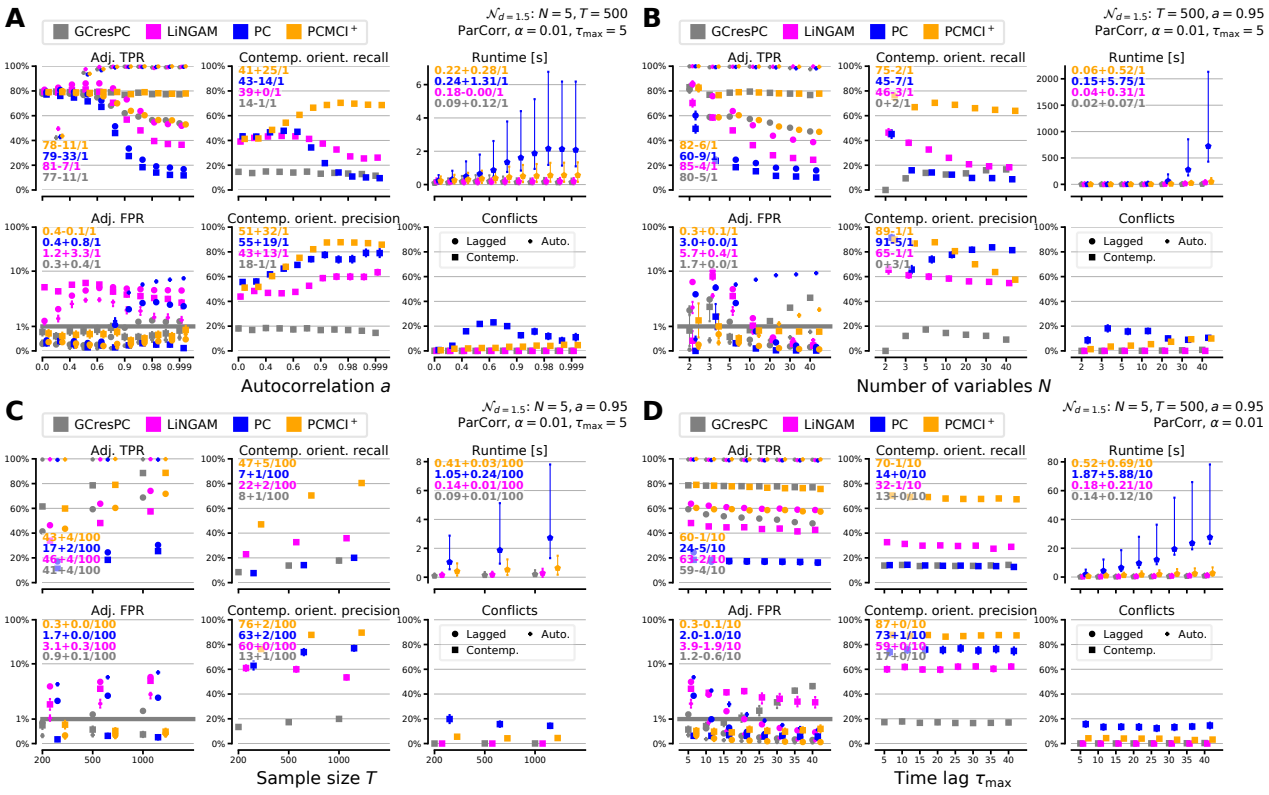

Figure S1: Numerical experiments with linear Gaussian setup for varying (**A**) autocorrelation strength a (**B**) number of variables N (**C**) sample size T and (**D**) maximum time lag $\tau_{\max}$. All remaining setup parameters indicated in the top right. Errorbars show std. errors or the 90% range (for runtime). The insets show ANOVA statistics.

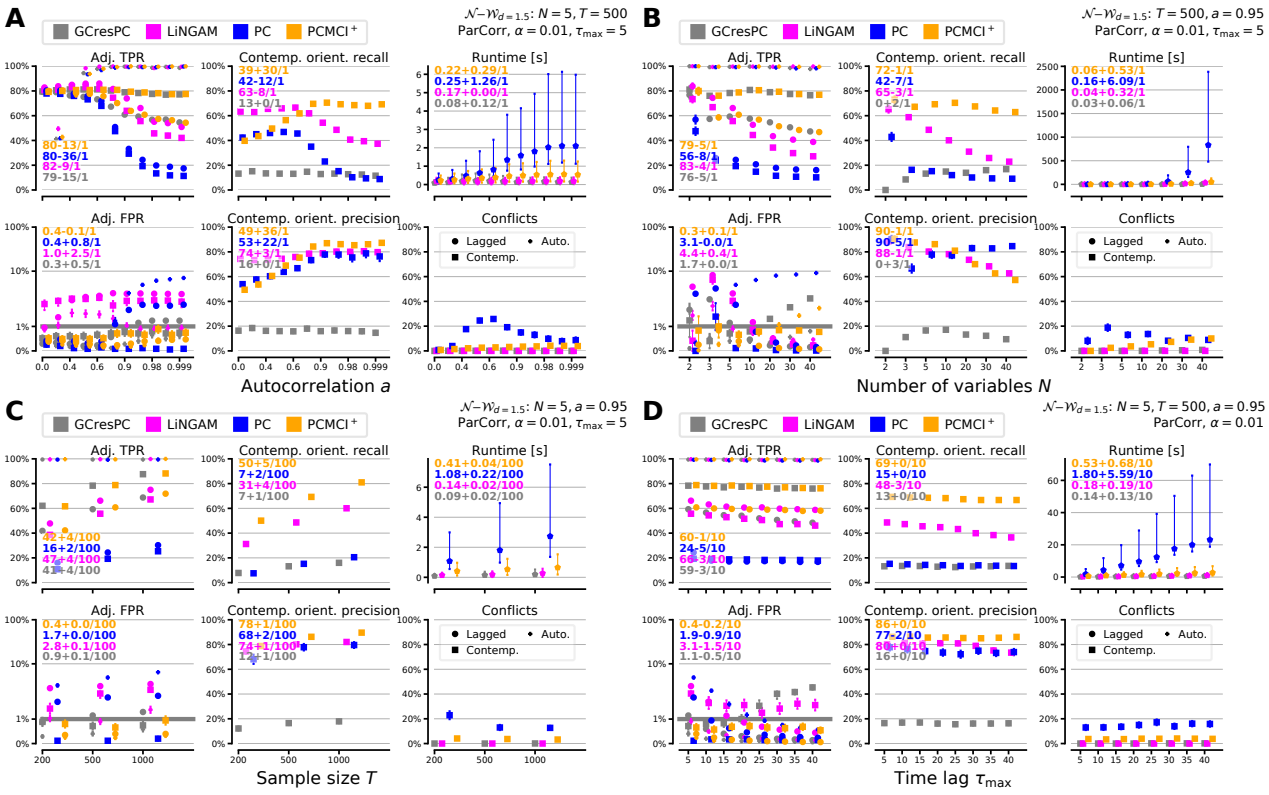

Figure S2: Numerical experiments with linear mixed noise setup for varying (**A**) autocorrelation strength a (**B**) number of variables N (**C**) sample size T and (**D**) maximum time lag $\tau_{\max}$. All remaining setup parameters indicated in the top right. Errorbars show std. errors or the 90% range (for runtime). The insets show ANOVA statistics.

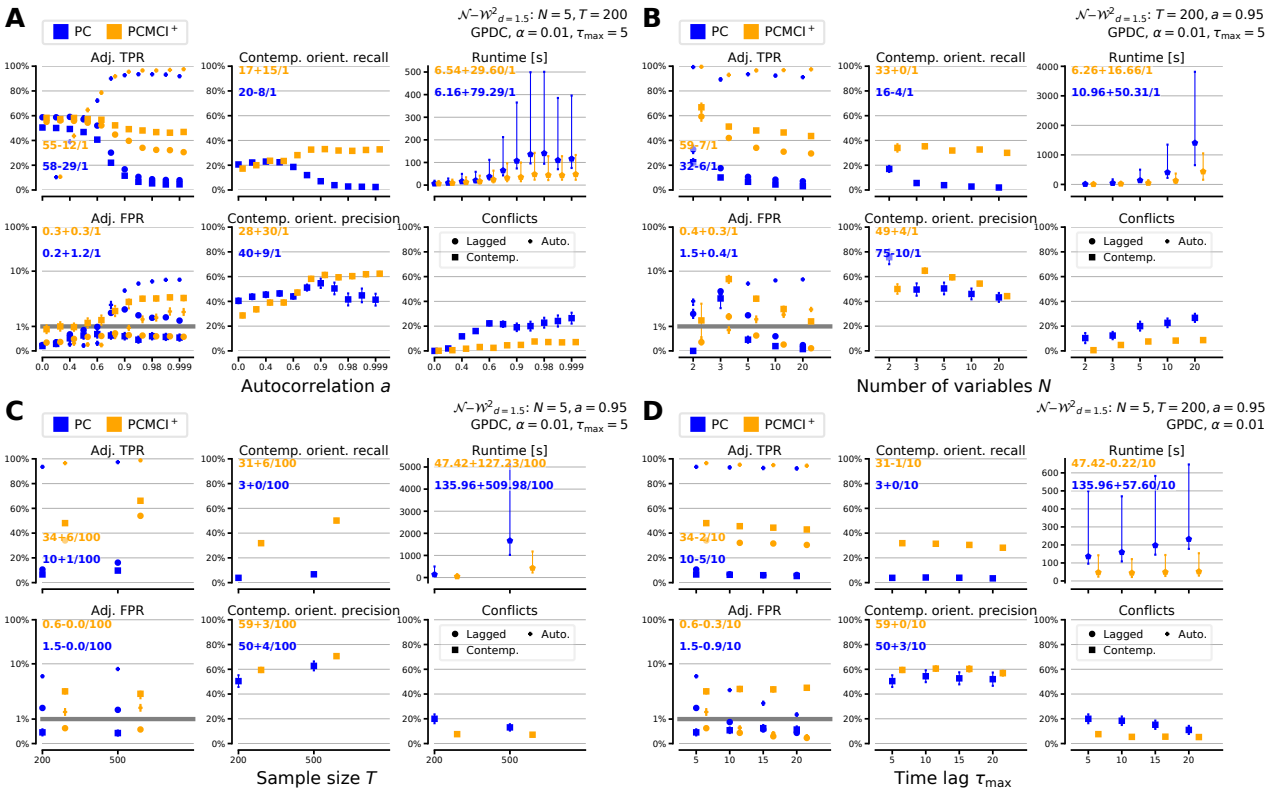

Figure S3: Numerical experiments with nonlinear mixed noise setup for varying (**A**) autocorrelation strength a (**B**) number of variables N (**C**) sample size T and (**D**) maximum time lag $\tau_{\max}$. All remaining setup parameters indicated in the top right. Errorbars show std. errors or the 90% range (for runtime). The insets show ANOVA statistics.

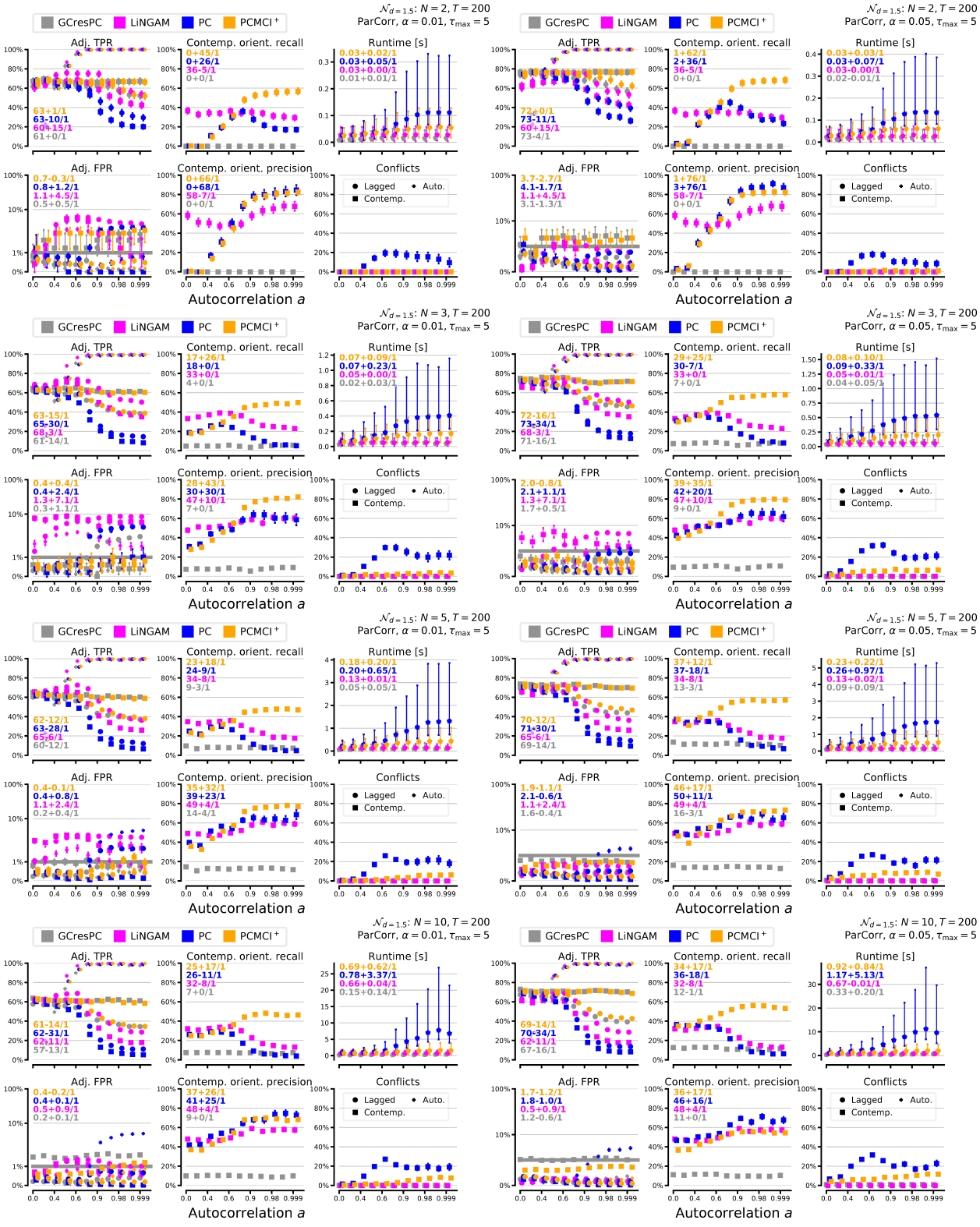

Figure S4: Numerical experiments with linear Gaussian setup for varying autocorrelation a and $T = 200$. The left (right) column shows results for significance level $\alpha = 0.01$ ($\alpha = 0.05$). The rows depict results for $N = 2, 3, 5, 10$ (top to bottom). All model and method parameters are indicated in the upper right of each panel.

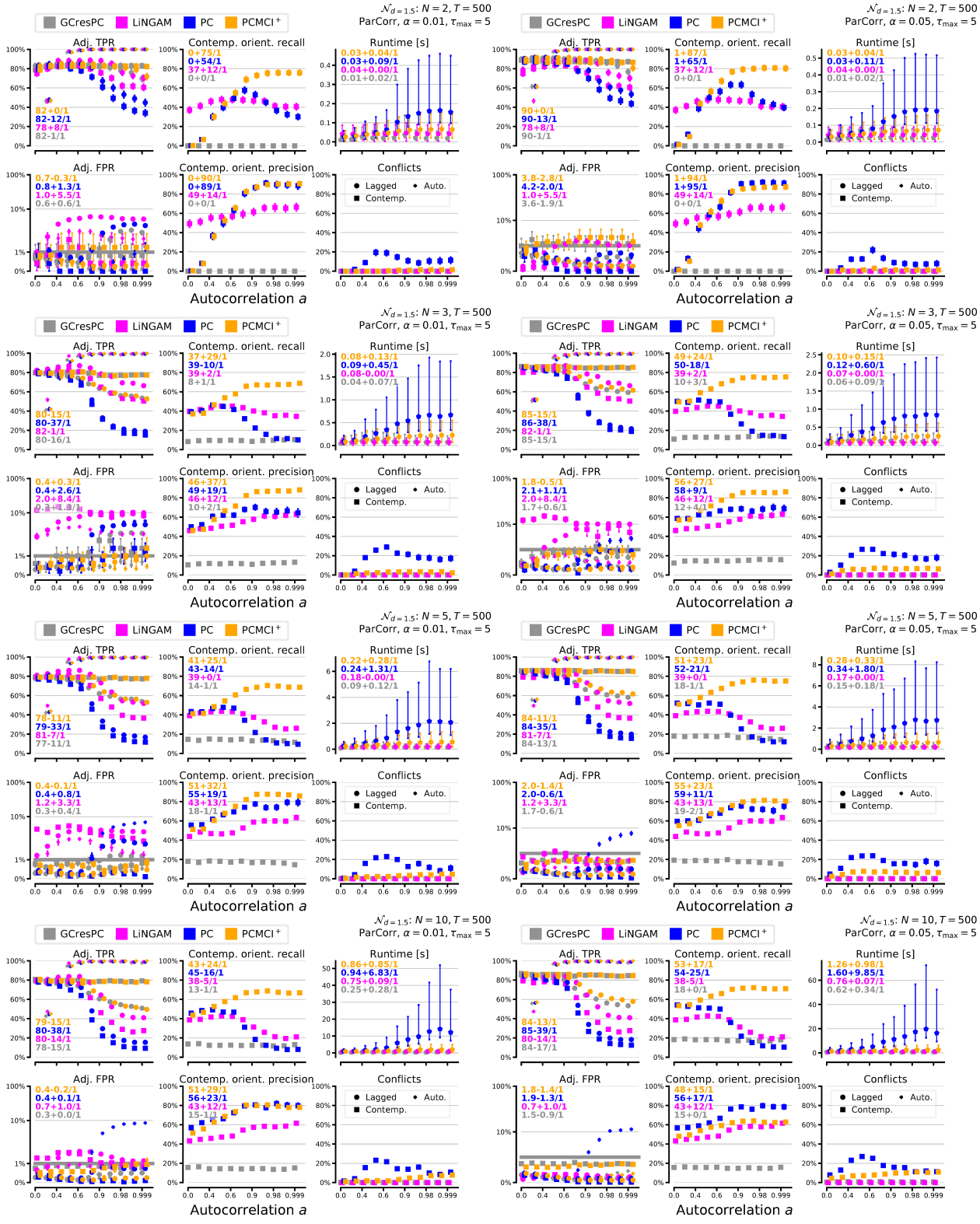

Figure S5: Numerical experiments with linear Gaussian setup for varying autocorrelation a and $T = 500$. The left (right) column shows results for significance level $\alpha = 0.01$ ($\alpha = 0.05$). The rows depict results for $N = 2, 3, 5, 10$ (top to bottom). All model and method parameters are indicated in the upper right of each panel.

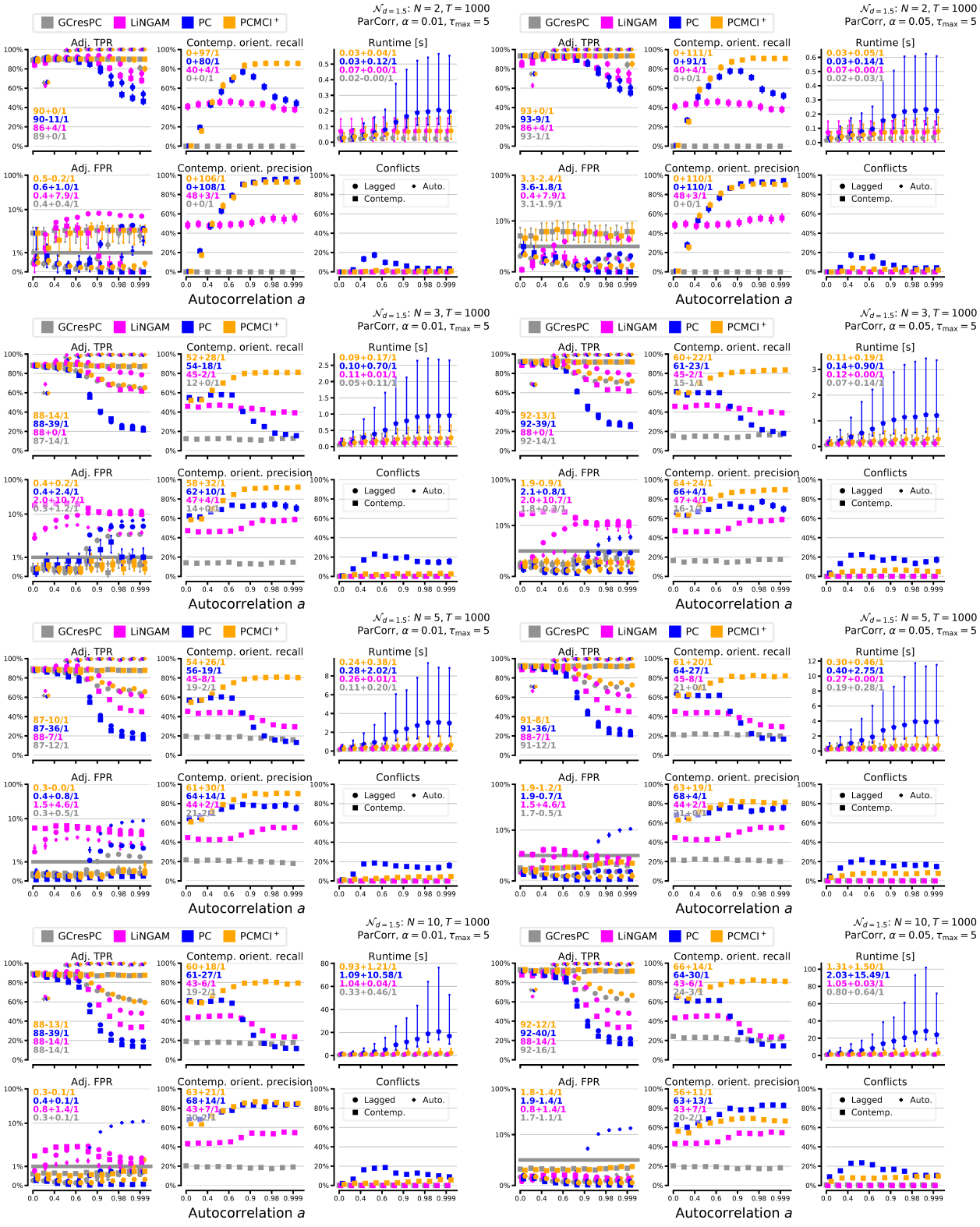

Figure S6: Numerical experiments with linear Gaussian setup for varying autocorrelation a and $T = 1000$. The left (right) column shows results for significance level $\alpha = 0.01$ ($\alpha = 0.05$). The rows depict results for $N = 2, 3, 5, 10$ (top to bottom). All model and method parameters are indicated in the upper right of each panel.

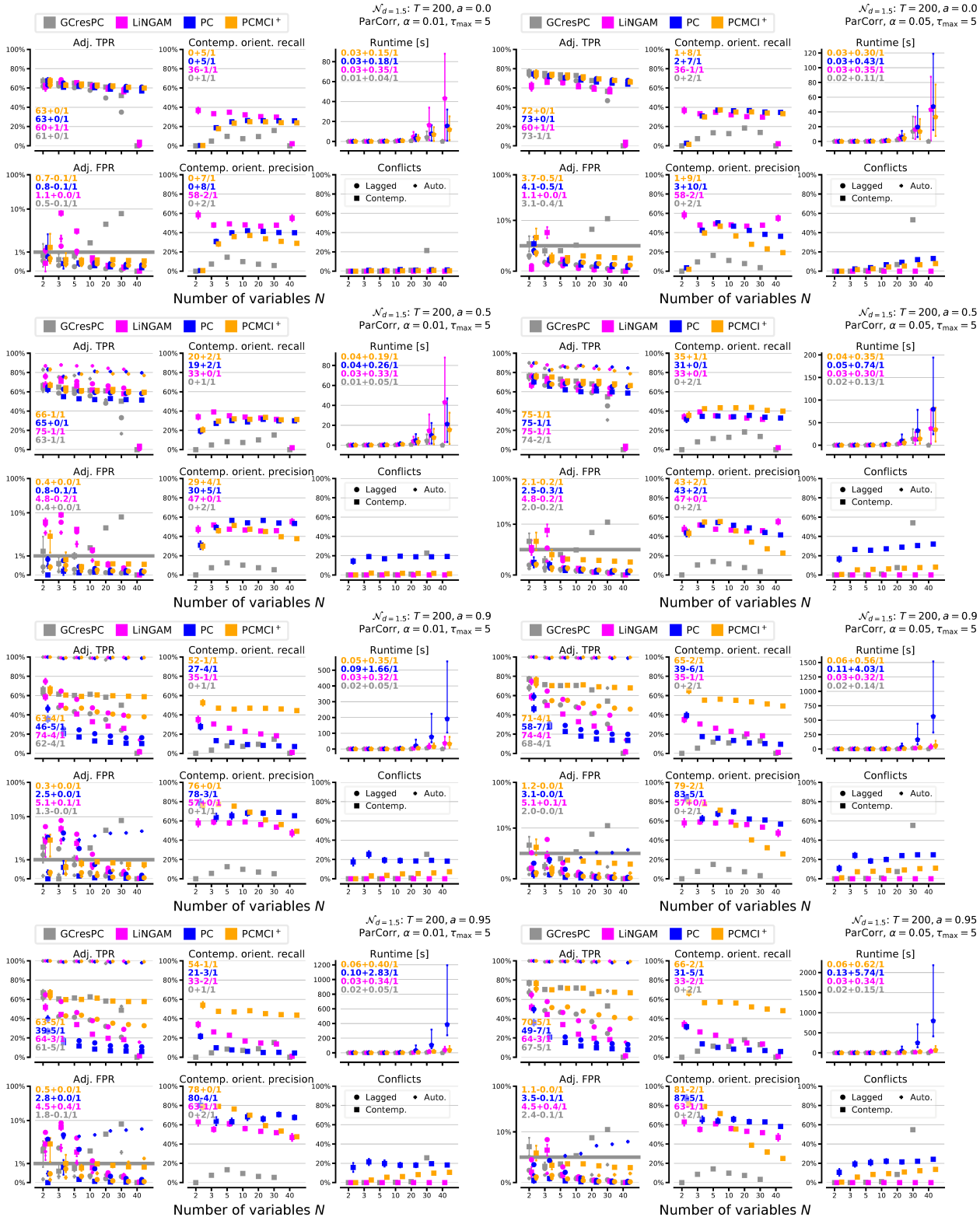

Figure S7: Numerical experiments with linear Gaussian setup for varying number of variables N and $T = 200$. The left (right) column shows results for significance level $\alpha = 0.01$ ($\alpha = 0.05$). The rows depict results for increasing autocorrelations a (top to bottom). All model and method parameters are indicated in the upper right of each panel.

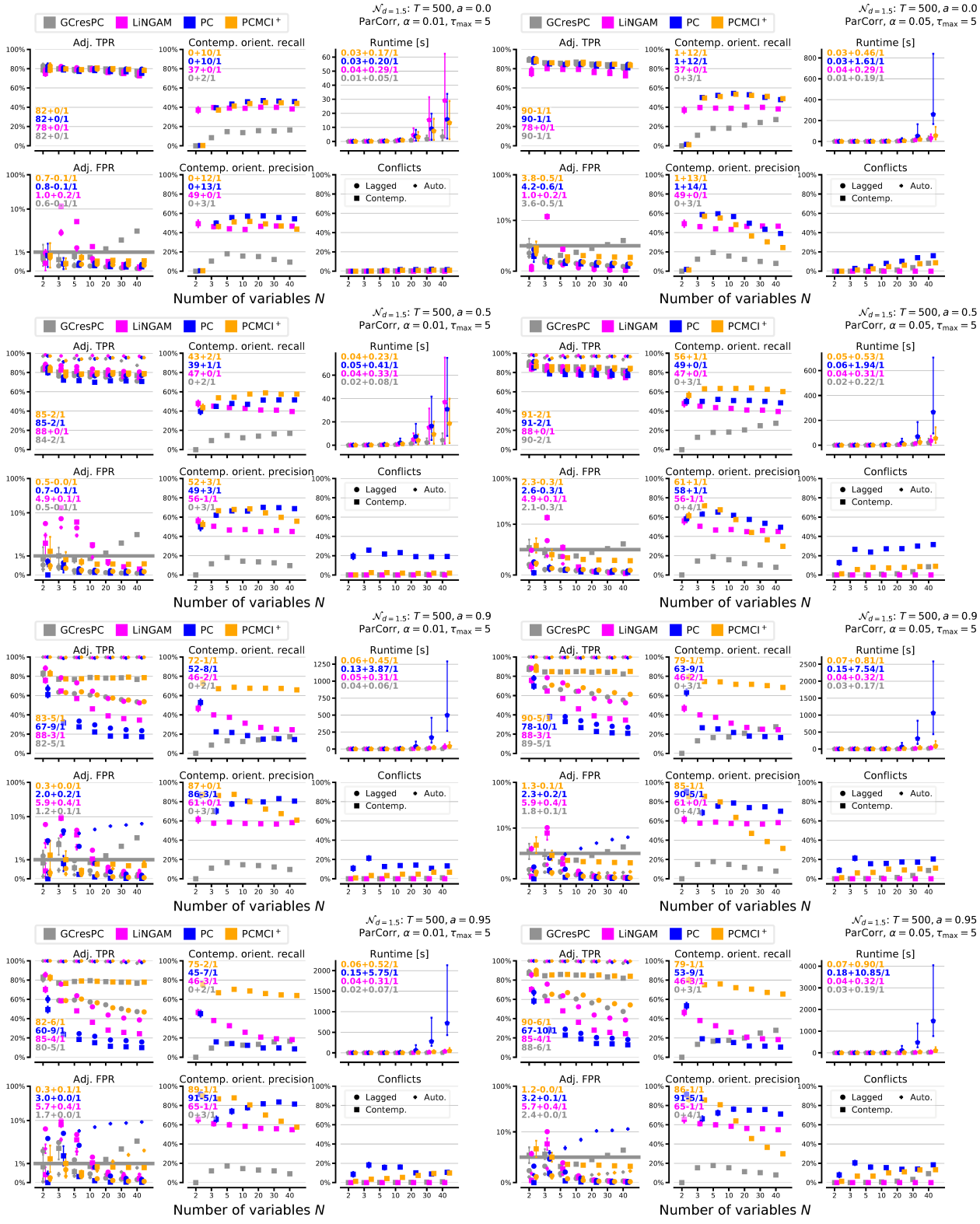

Figure S8: Numerical experiments with linear Gaussian setup for varying number of variables N and $T = 500$. The left (right) column shows results for significance level $\alpha = 0.01$ ($\alpha = 0.05$). The rows depict results for increasing autocorrelations a (top to bottom). All model and method parameters are indicated in the upper right of each panel.

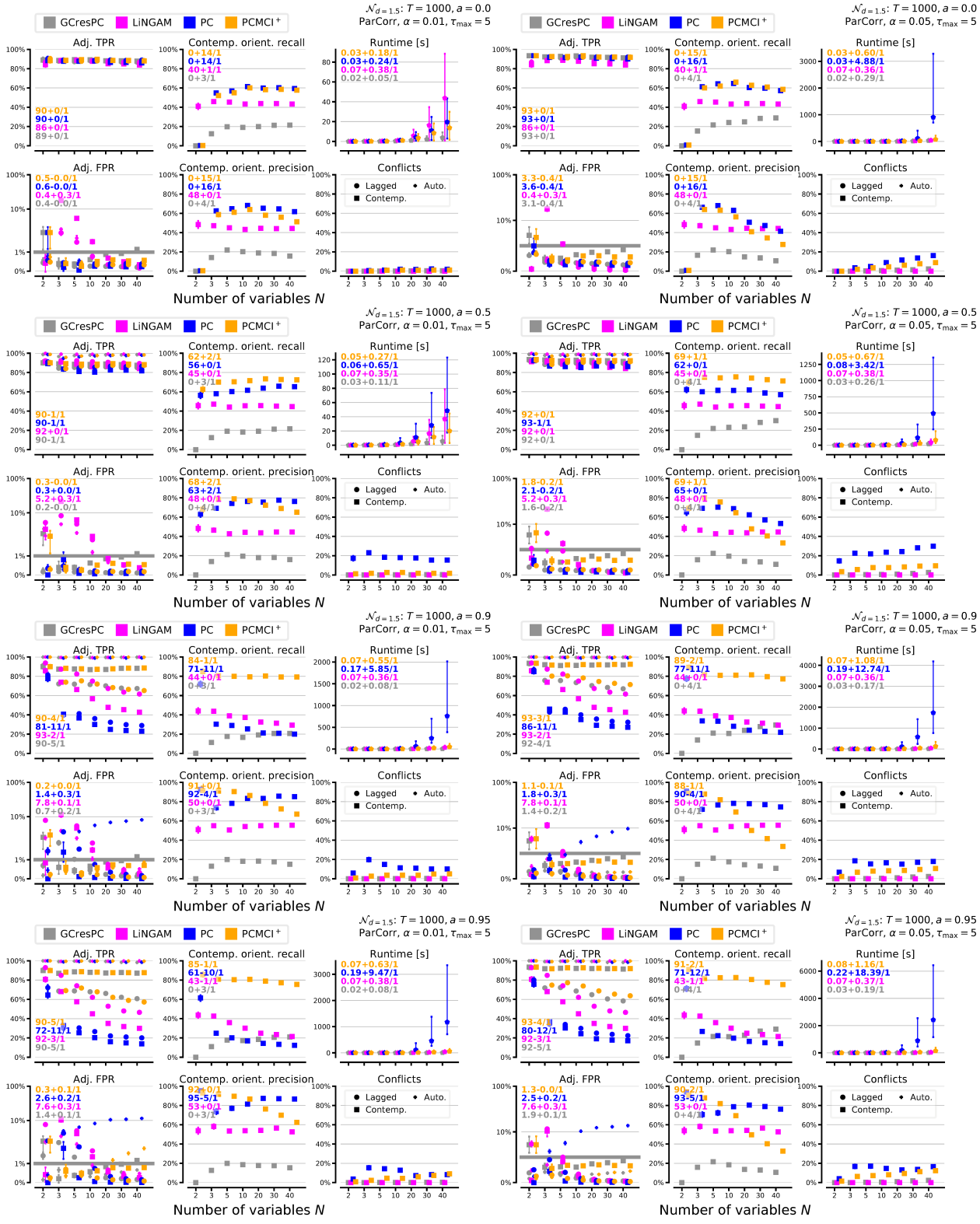

Figure S9: Numerical experiments with linear Gaussian setup for varying number of variables N and $T = 1000$. The left (right) column shows results for significance level $\alpha = 0.01$ ($\alpha = 0.05$). The rows depict results for increasing autocorrelations a (top to bottom). All model and method parameters are indicated in the upper right of each panel.

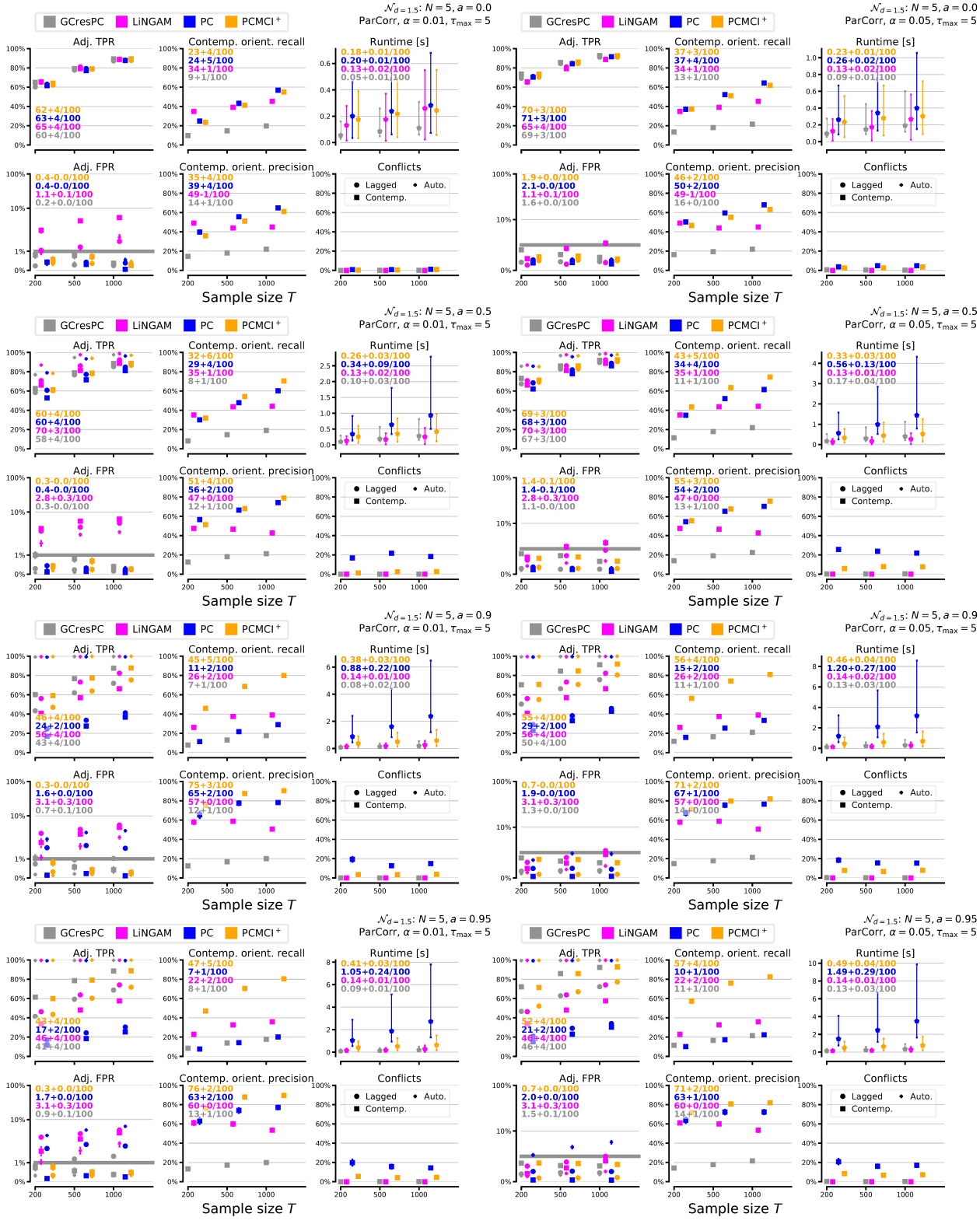

Figure S10: Numerical experiments with linear Gaussian setup for varying sample size T for $N = 5$. The left (right) column shows results for significance level $\alpha = 0.01$ ($\alpha = 0.05$). The rows depict results for increasing autocorrelations a (top to bottom). All model and method parameters are indicated in the upper right of each panel.

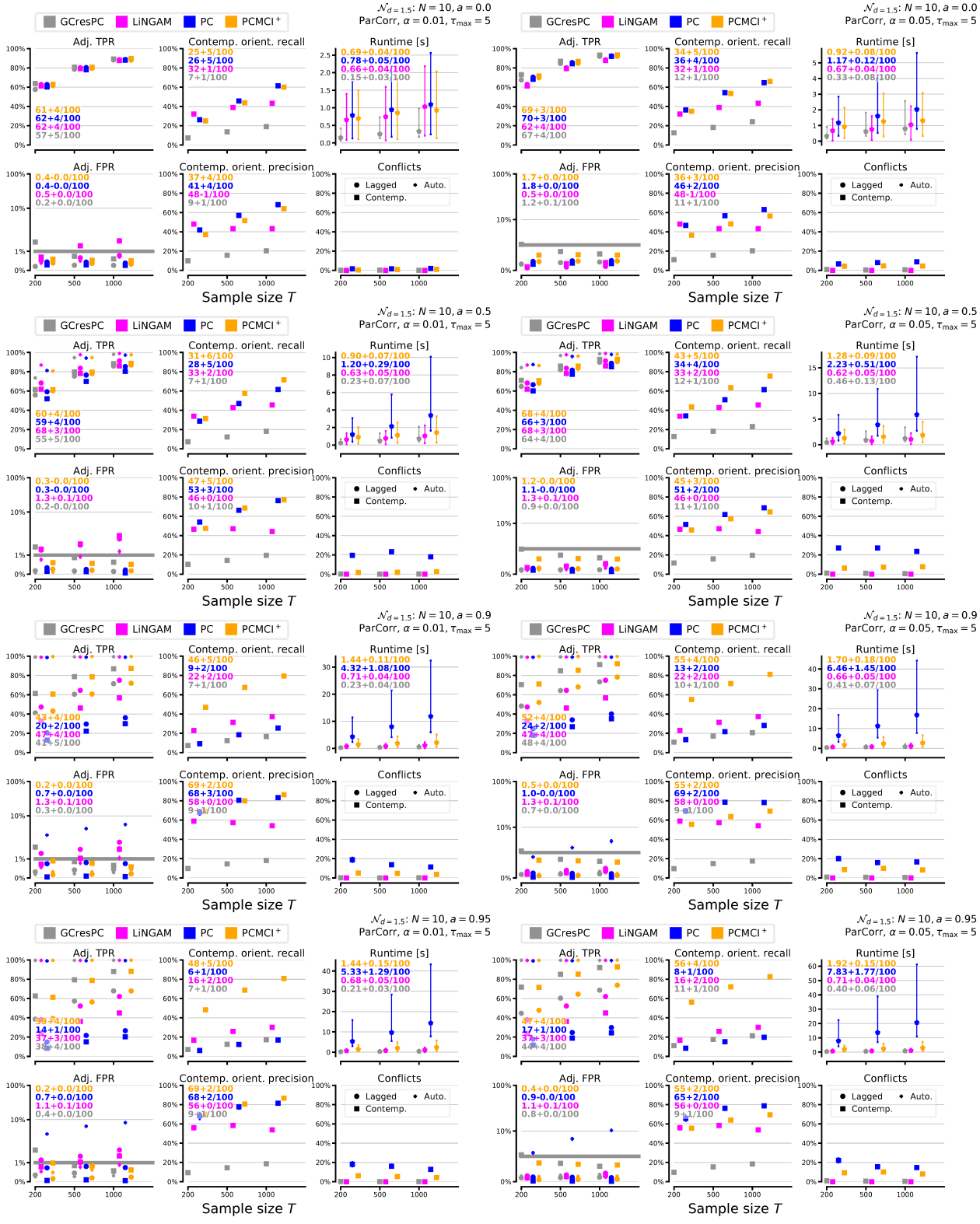

Figure S11: Numerical experiments with linear Gaussian setup for varying sample size T for $N = 10$. The left (right) column shows results for significance level $\alpha = 0.01$ ($\alpha = 0.05$). The rows depict results for increasing autocorrelations a (top to bottom). All model and method parameters are indicated in the upper right of each panel.

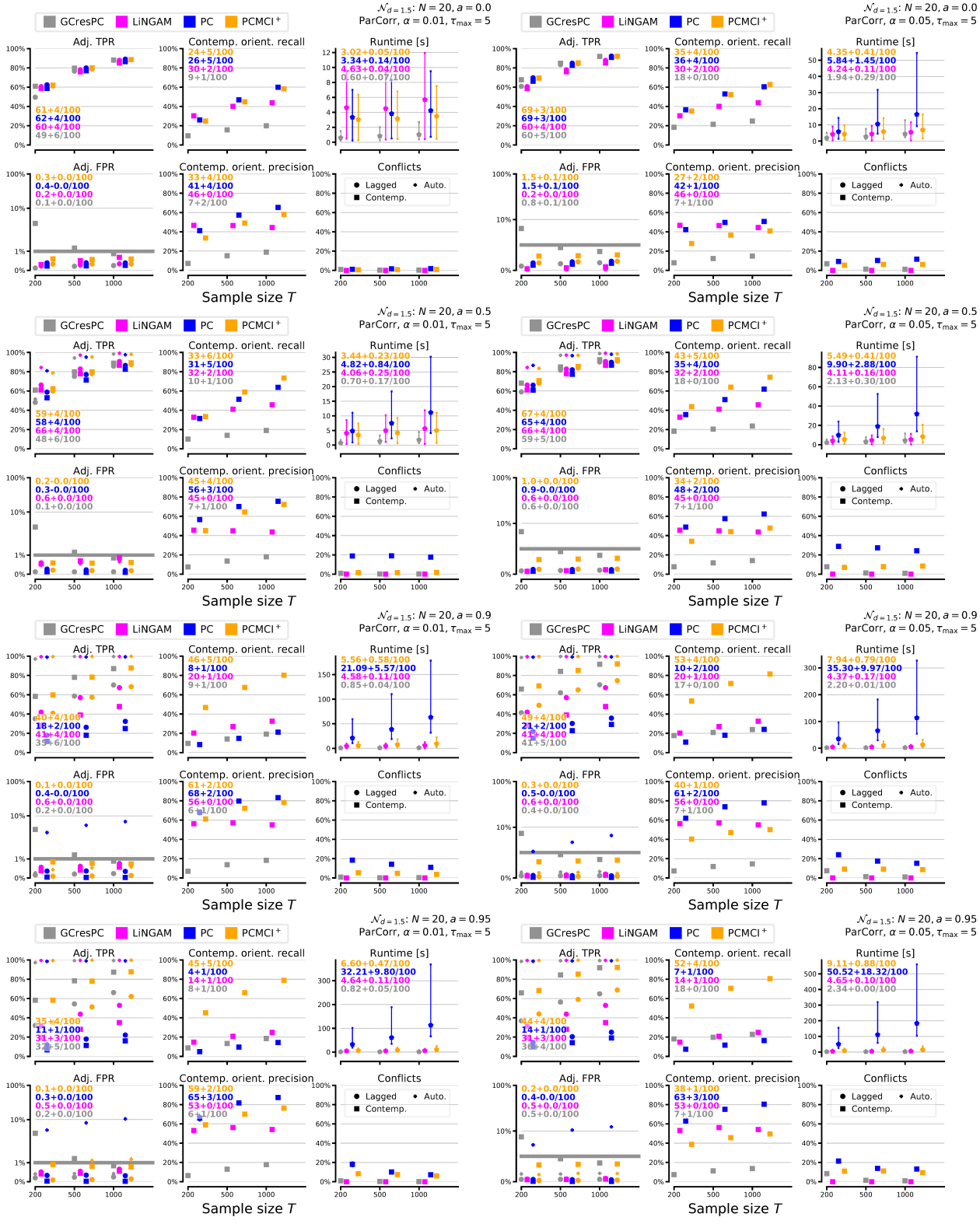

Figure S12: Numerical experiments with linear Gaussian setup for varying sample size T for $N = 20$. The left (right) column shows results for significance level $\alpha = 0.01$ ($\alpha = 0.05$). The rows depict results for increasing autocorrelations a (top to bottom). All model and method parameters are indicated in the upper right of each panel.

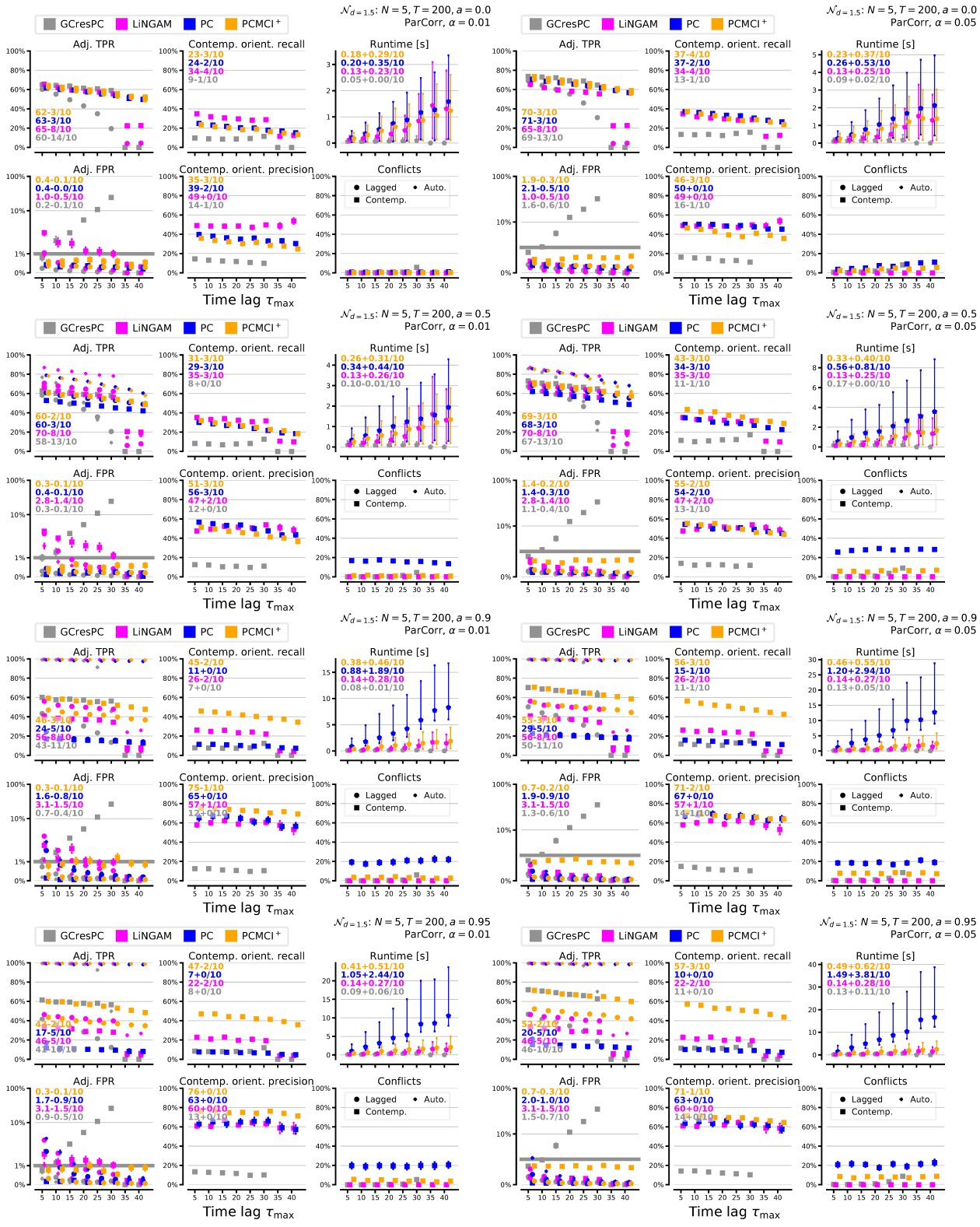

Figure S13: Numerical experiments with linear Gaussian setup for varying maximum time lag $\tau_{\max}$ and $T = 200$. The left (right) column shows results for significance level $\alpha = 0.01$ ($\alpha = 0.05$). The rows depict results for increasing autocorrelations a (top to bottom). All model and method parameters are indicated in the upper right of each panel.

Figure S14: Numerical experiments with linear Gaussian setup for varying maximum time lag $\tau_{\max}$ and $T = 500$. The left (right) column shows results for significance level $\alpha = 0.01$ ($\alpha = 0.05$). The rows depict results for increasing autocorrelations a (top to bottom). All model and method parameters are indicated in the upper right of each panel.

Figure S15: Numerical experiments with linear Gaussian setup for varying maximum time lag $\tau_{\max}$ and $T = 1000$. The left (right) column shows results for significance level $\alpha = 0.01$ ($\alpha = 0.05$). The rows depict results for increasing autocorrelations a (top to bottom). All model and method parameters are indicated in the upper right of each panel.

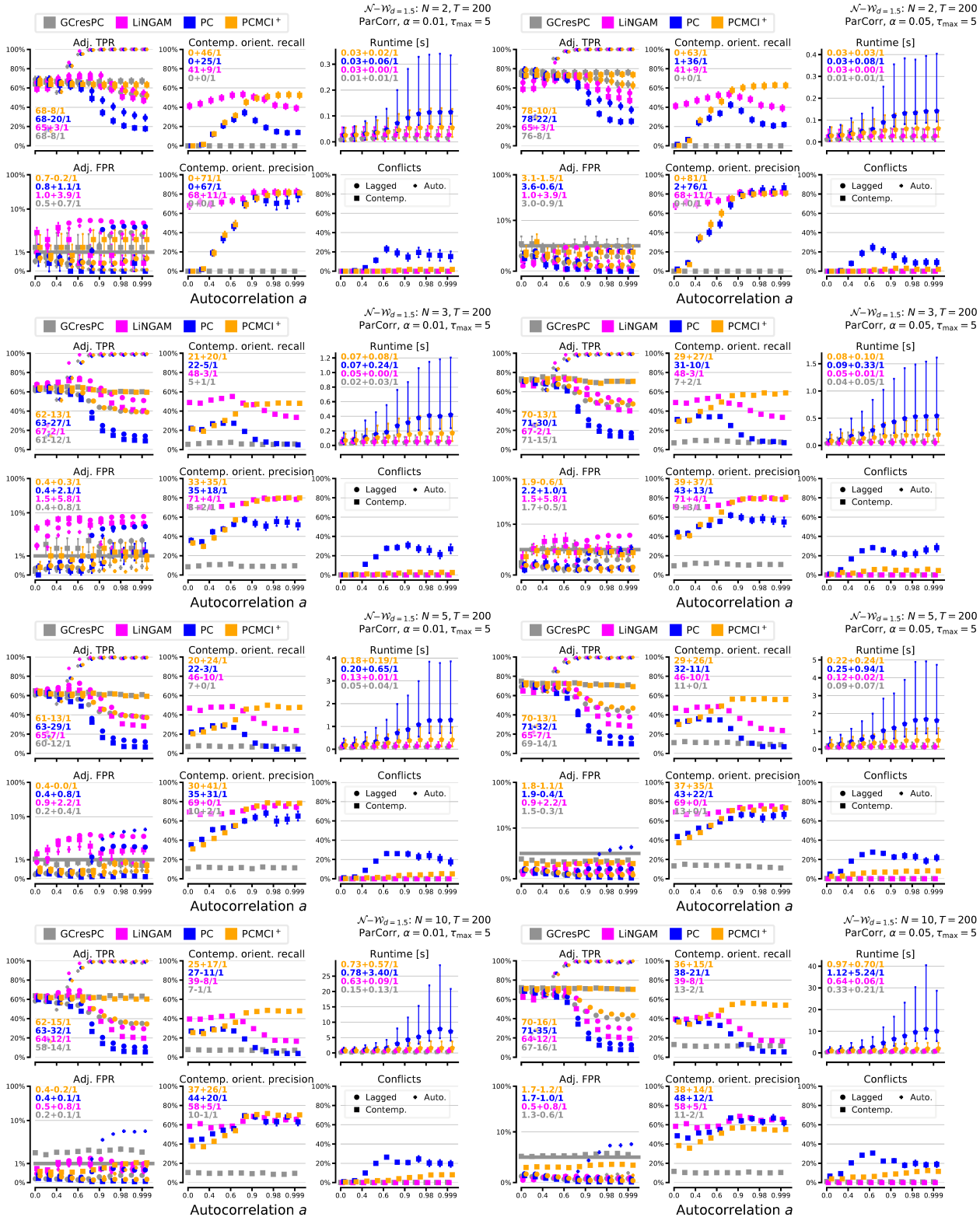

Figure S16: Numerical experiments with linear mixed noise setup for varying autocorrelation a and $T = 200$. The left (right) column shows results for significance level $\alpha = 0.01$ ($\alpha = 0.05$). The rows depict results for $N = 2, 3, 5, 10$ (top to bottom). All model and method parameters are indicated in the upper right of each panel.

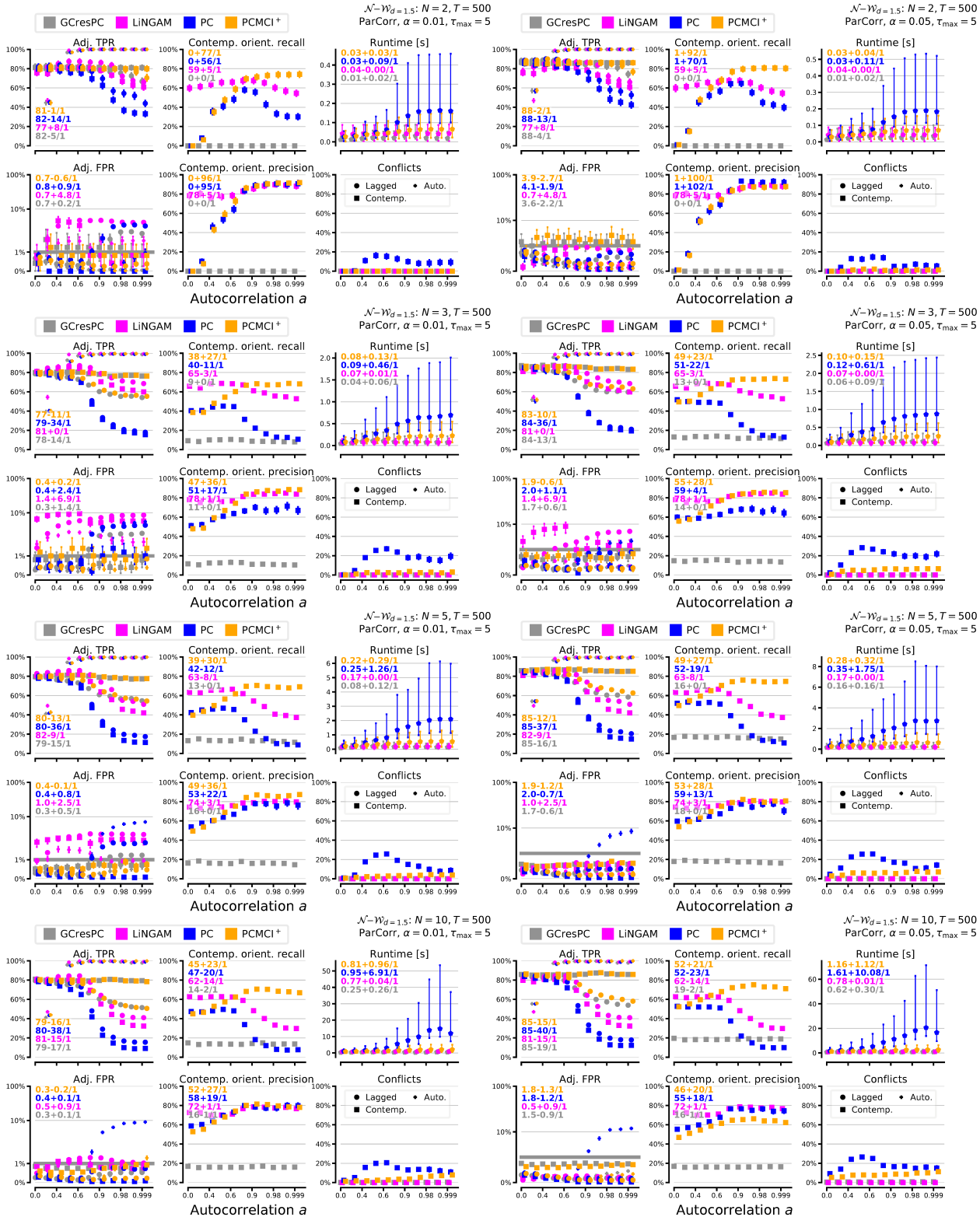

Figure S17: Numerical experiments with linear mixed noise setup for varying autocorrelation a and $T = 500$. The left (right) column shows results for significance level $\alpha = 0.01$ ($\alpha = 0.05$). The rows depict results for $N = 2, 3, 5, 10$ (top to bottom). All model and method parameters are indicated in the upper right of each panel.

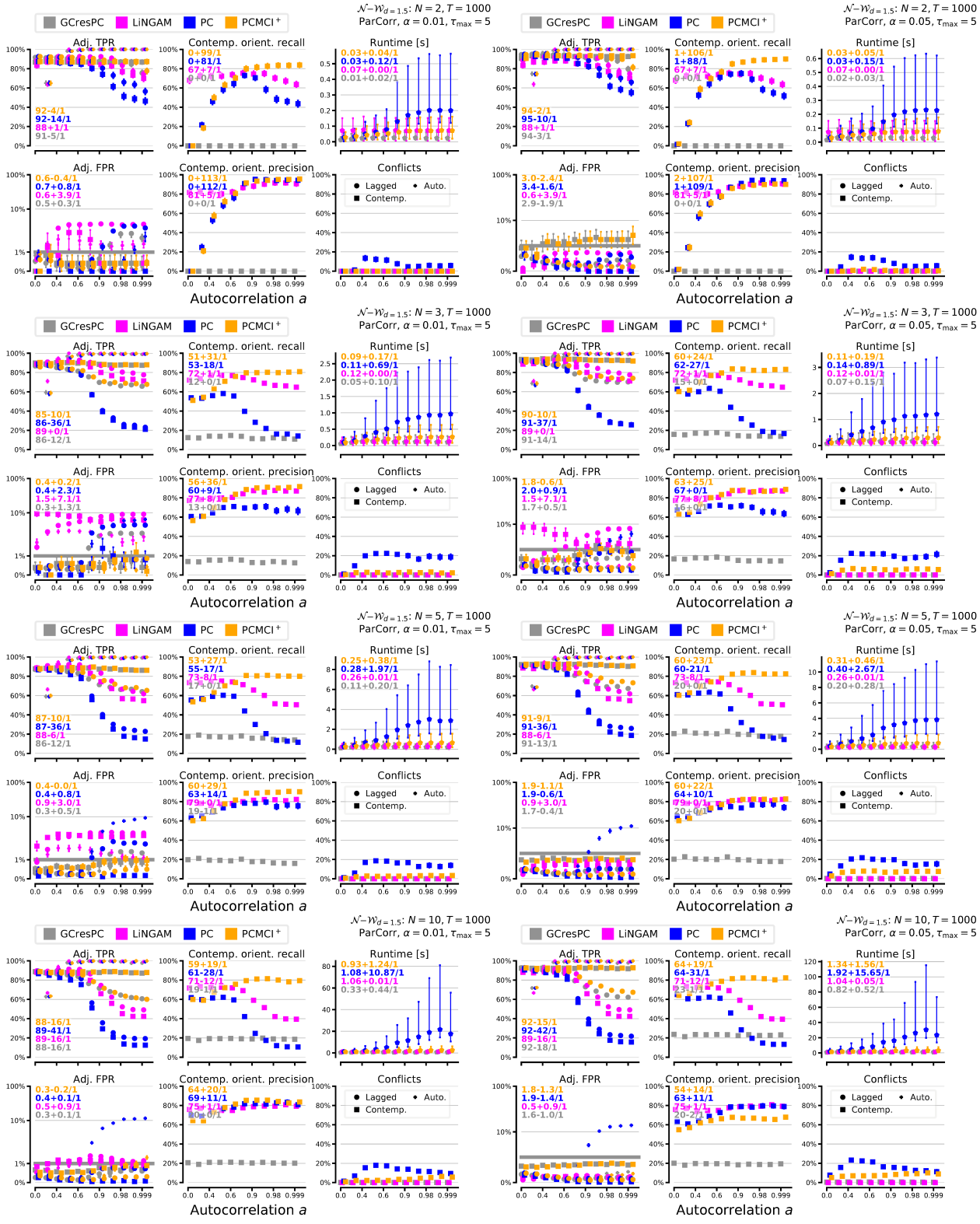

Figure S18: Numerical experiments with linear mixed noise setup for varying autocorrelation a and $T = 1000$. The left (right) column shows results for significance level $\alpha = 0.01$ ($\alpha = 0.05$). The rows depict results for $N = 2, 3, 5, 10$ (top to bottom). All model and method parameters are indicated in the upper right of each panel.

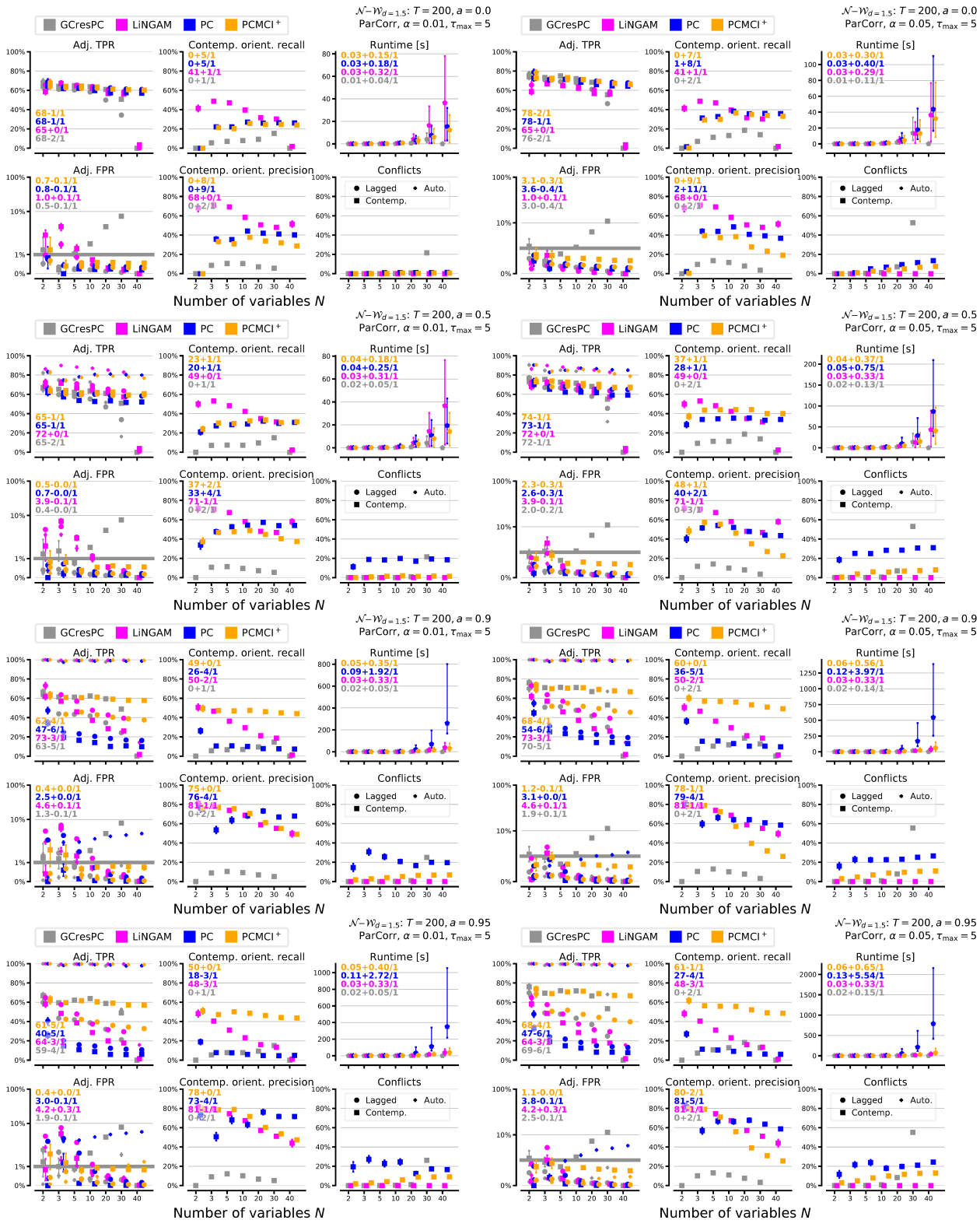

Figure S19: Numerical experiments with linear mixed noise setup for varying number of variables N and $T = 200$. The left (right) column shows results for significance level $\alpha = 0.01$ ($\alpha = 0.05$). The rows depict results for increasing autocorrelations a (top to bottom). All model and method parameters are indicated in the upper right of each panel.

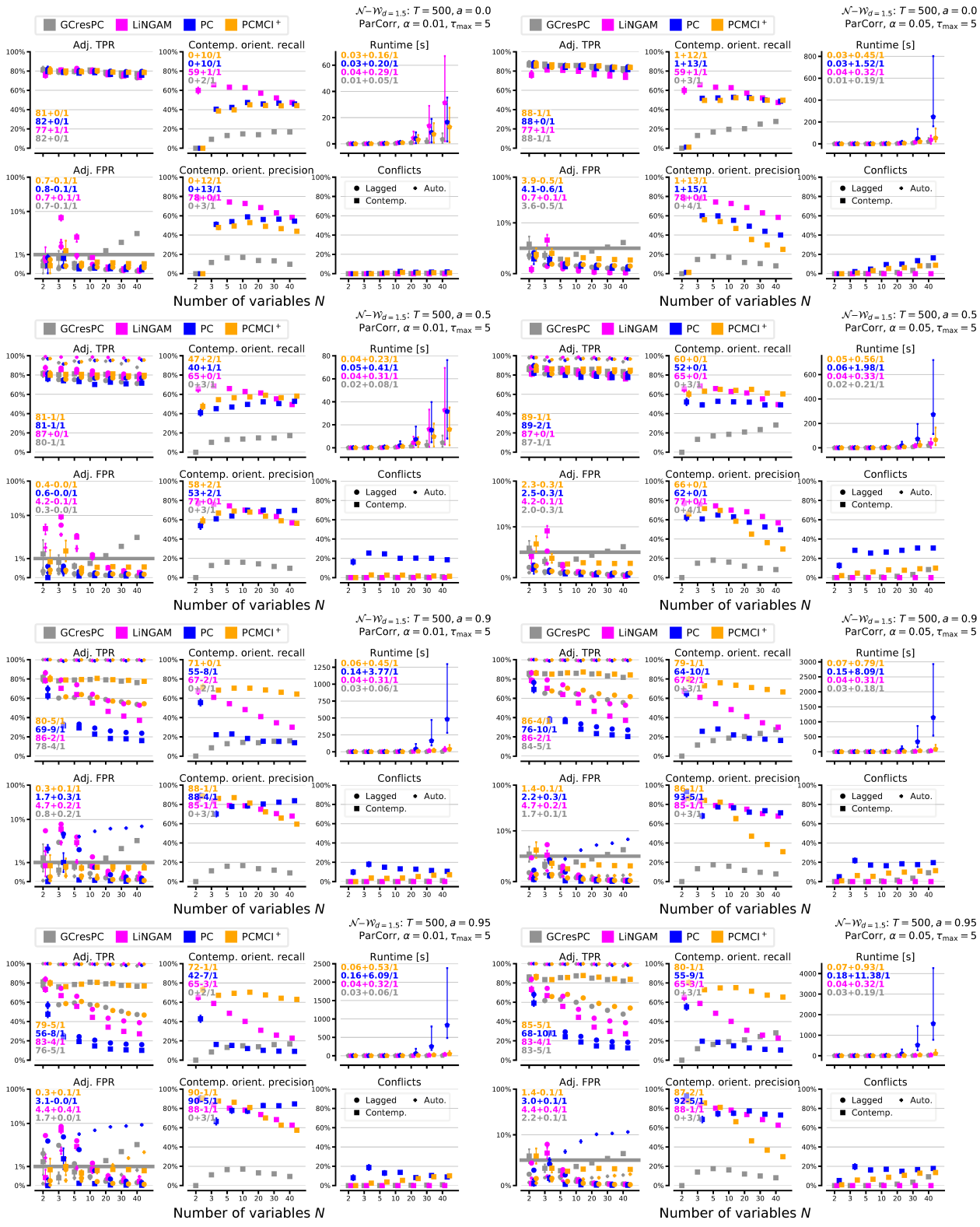

Figure S20: Numerical experiments with linear mixed noise setup for varying number of variables N and $T = 500$. The left (right) column shows results for significance level $\alpha = 0.01$ ($\alpha = 0.05$). The rows depict results for increasing autocorrelations a (top to bottom). All model and method parameters are indicated in the upper right of each panel.

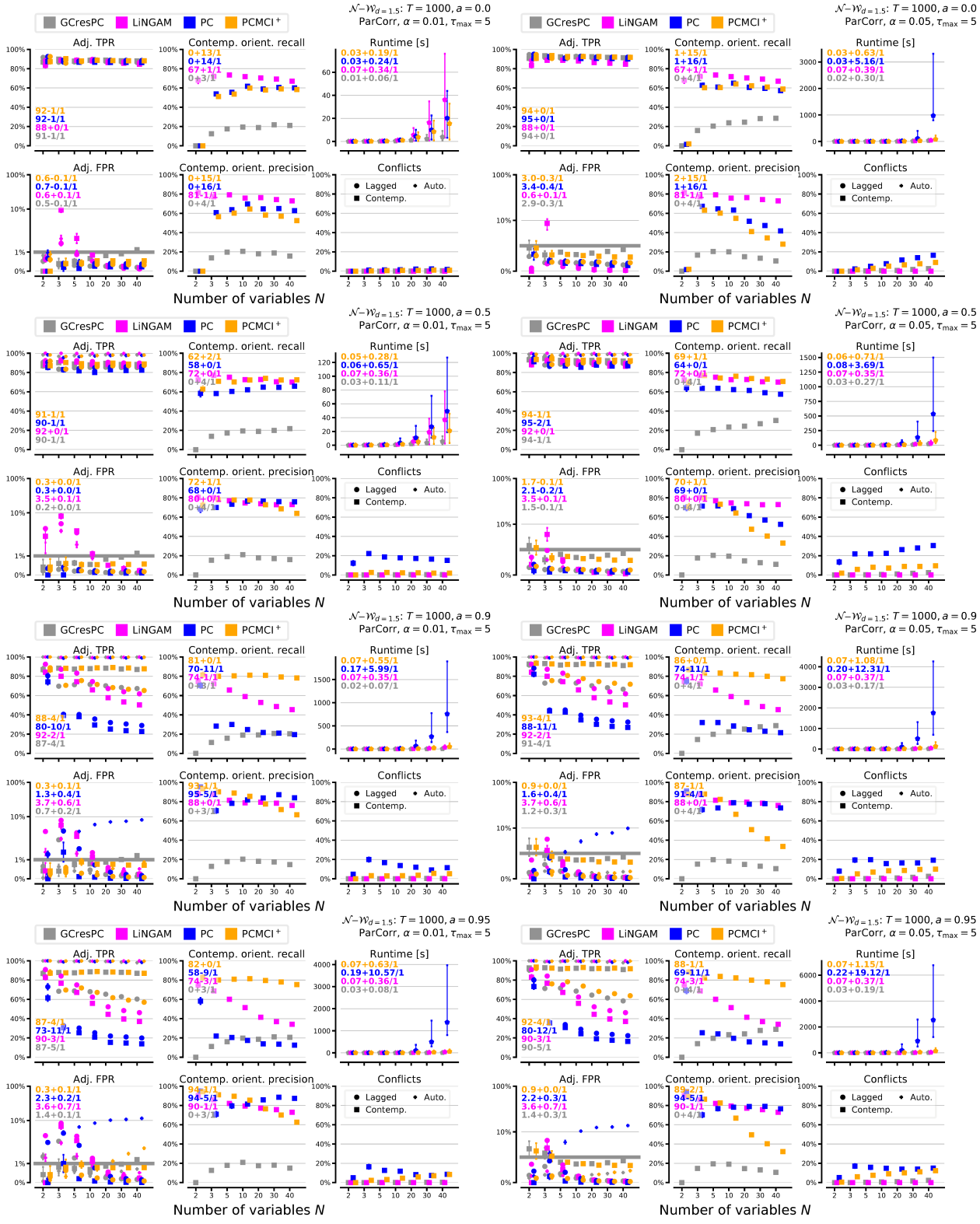

Figure S21: Numerical experiments with linear mixed noise setup for varying number of variables N and $T = 1000$. The left (right) column shows results for significance level $\alpha = 0.01$ ($\alpha = 0.05$). The rows depict results for increasing autocorrelations a (top to bottom). All model and method parameters are indicated in the upper right of each panel.

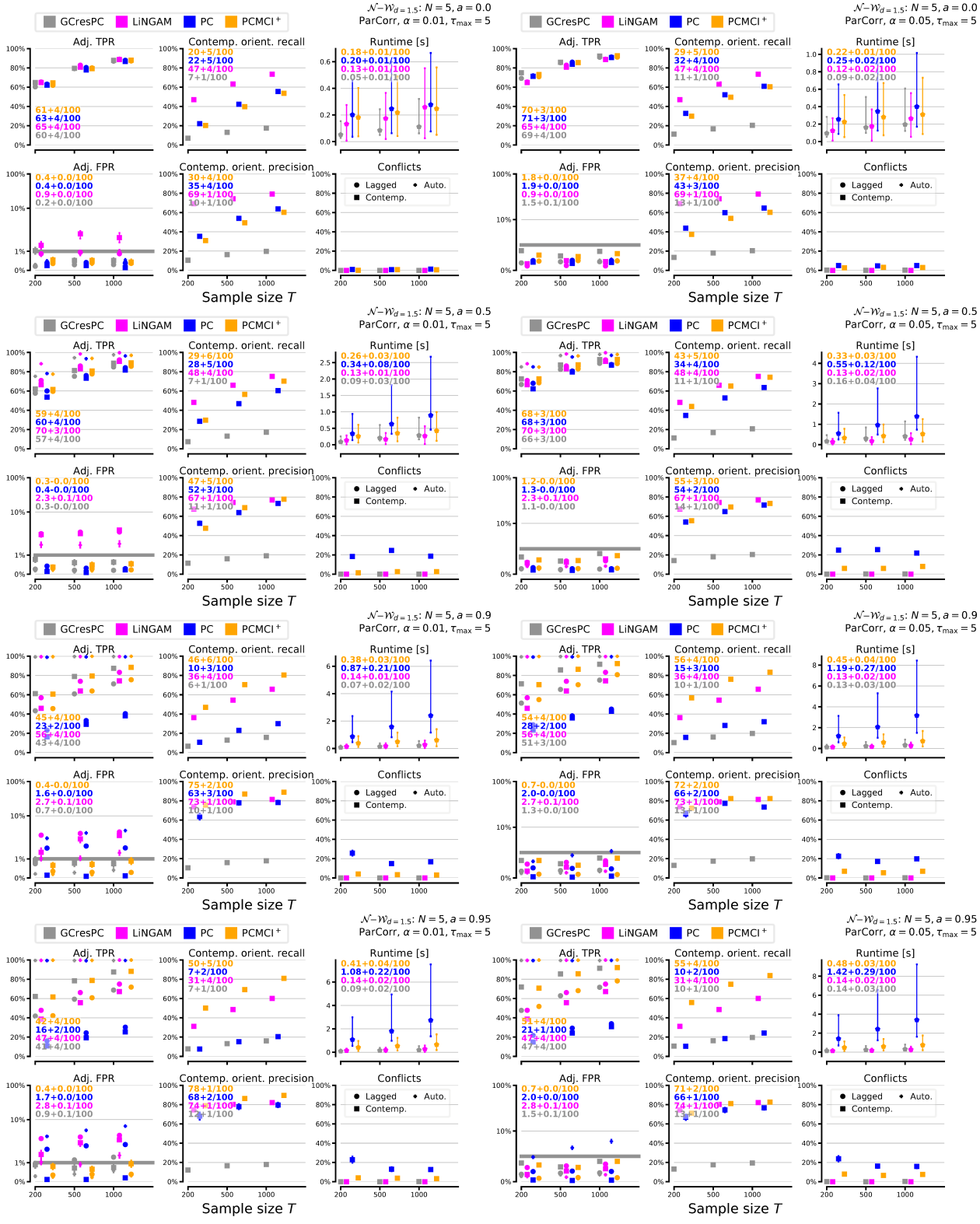

Figure S22: Numerical experiments with linear mixed noise setup for varying sample size T for $N = 5$. The left (right) column shows results for significance level $\alpha = 0.01$ ($\alpha = 0.05$). The rows depict results for increasing autocorrelations a (top to bottom). All model and method parameters are indicated in the upper right of each panel.

Figure S23: Numerical experiments with linear mixed noise setup for varying sample size T for $N = 10$. The left (right) column shows results for significance level $\alpha = 0.01$ ($\alpha = 0.05$). The rows depict results for increasing autocorrelations a (top to bottom). All model and method parameters are indicated in the upper right of each panel.

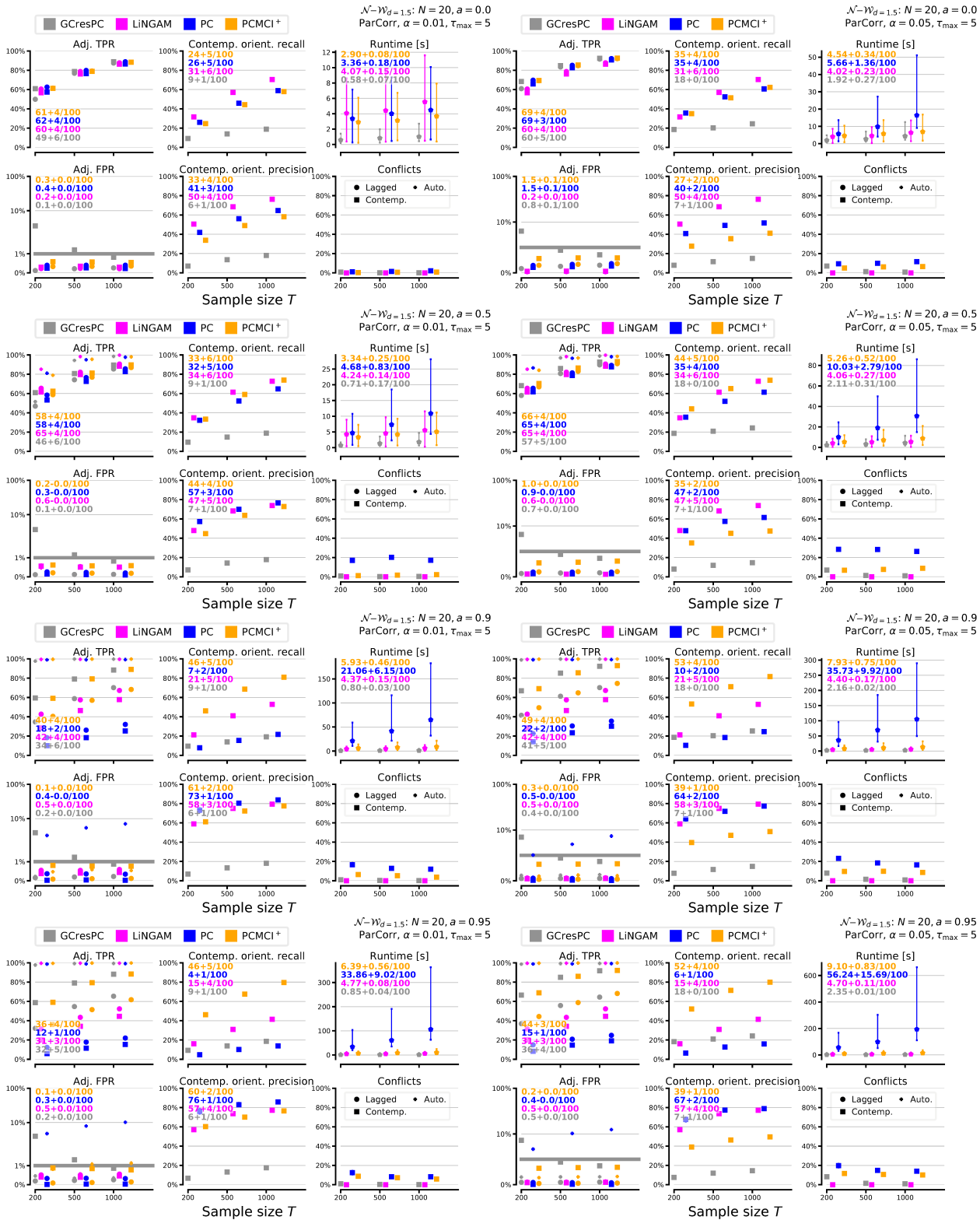

Figure S24: Numerical experiments with linear mixed noise setup for varying sample size T for $N = 20$. The left (right) column shows results for significance level $\alpha = 0.01$ ($\alpha = 0.05$). The rows depict results for increasing autocorrelations a (top to bottom). All model and method parameters are indicated in the upper right of each panel.

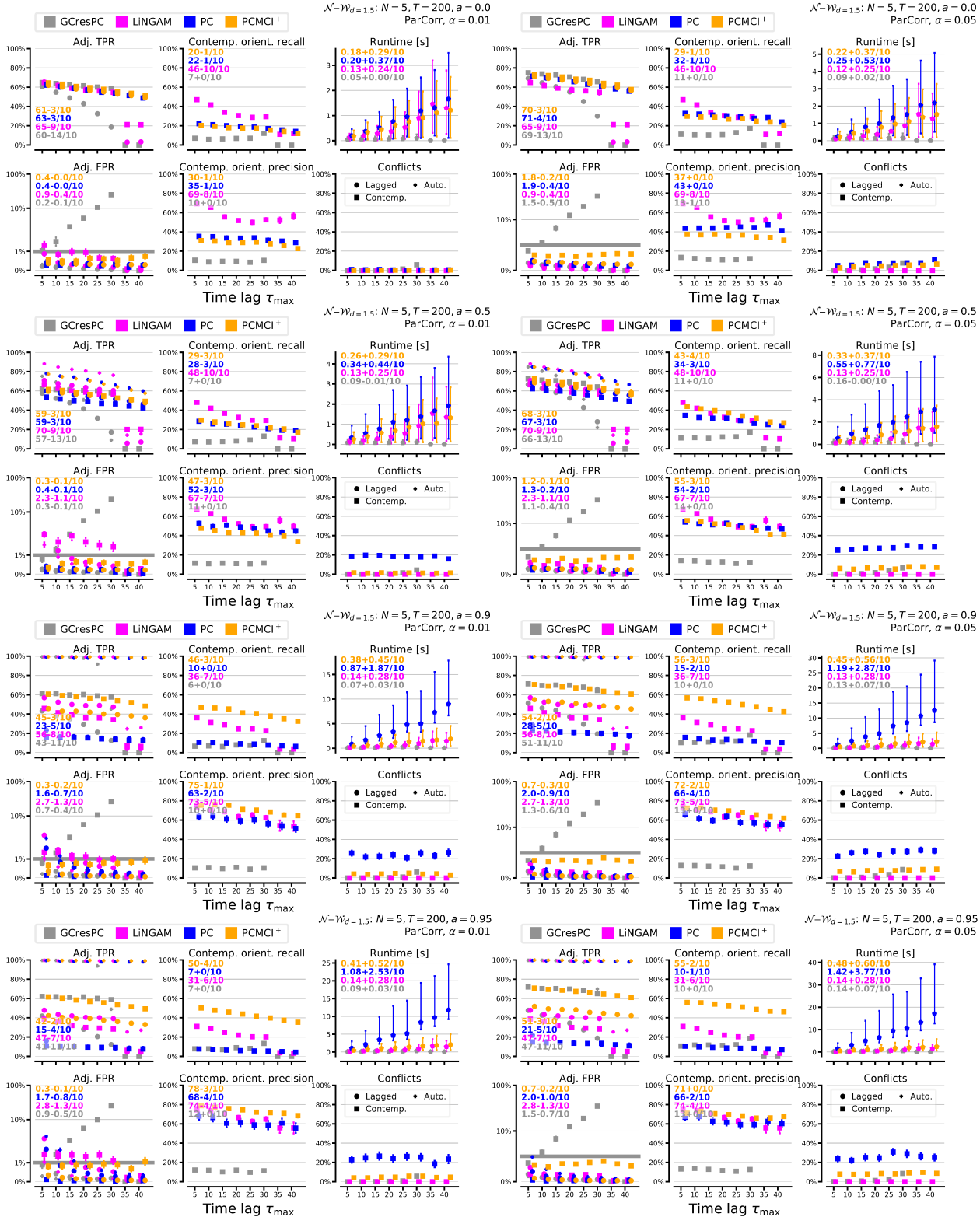

Figure S25: Numerical experiments with linear mixed noise setup for varying maximum time lag $\tau_{\max}$ and $T = 200$. The left (right) column shows results for significance level $\alpha = 0.01$ ($\alpha = 0.05$). The rows depict results for increasing autocorrelations a (top to bottom). All model and method parameters are indicated in the upper right of each panel.

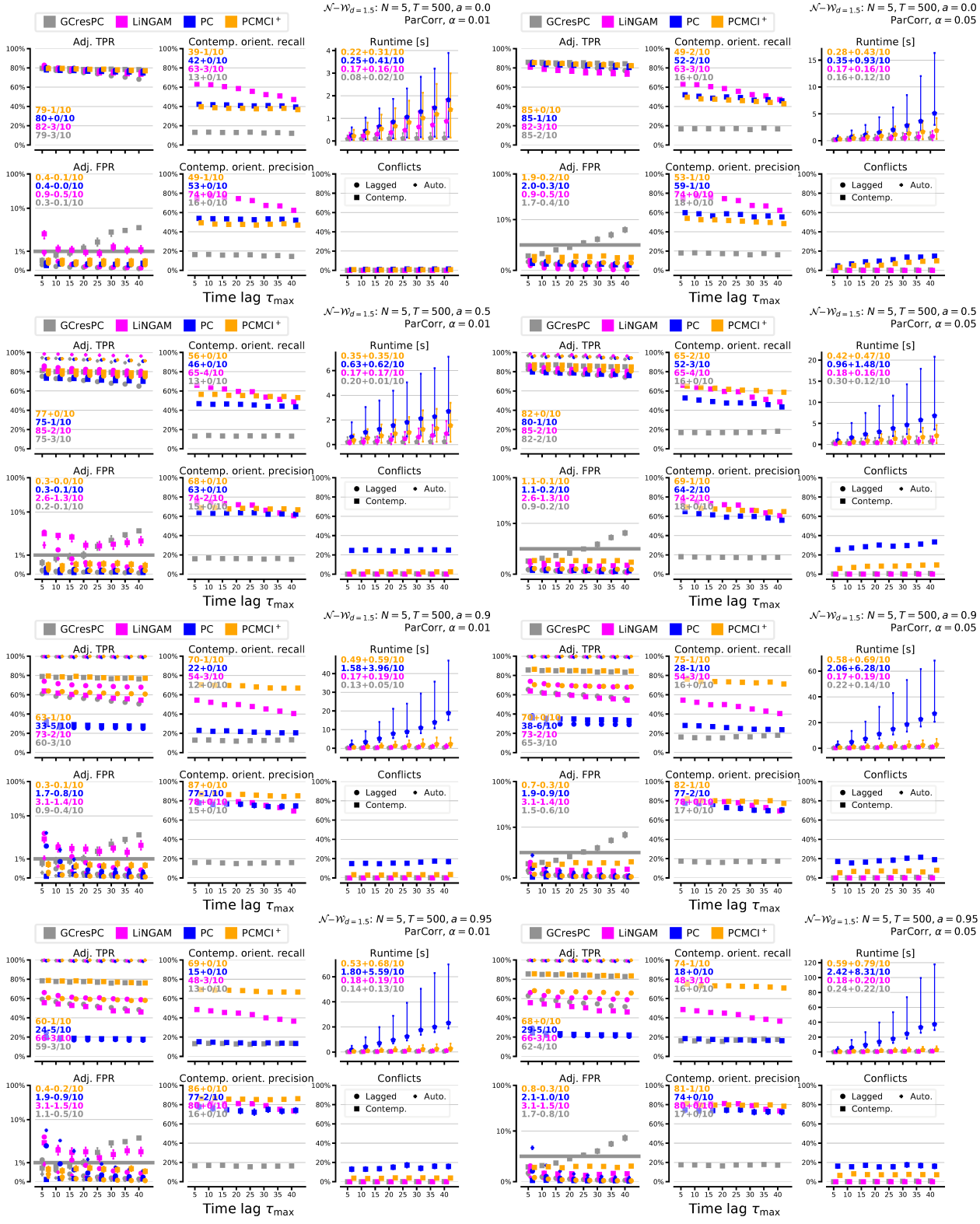

Figure S26: Numerical experiments with linear mixed noise setup for varying maximum time lag $\tau_{\max}$ and $T = 500$. The left (right) column shows results for significance level $\alpha = 0.01$ ($\alpha = 0.05$). The rows depict results for increasing autocorrelations a (top to bottom). All model and method parameters are indicated in the upper right of each panel.

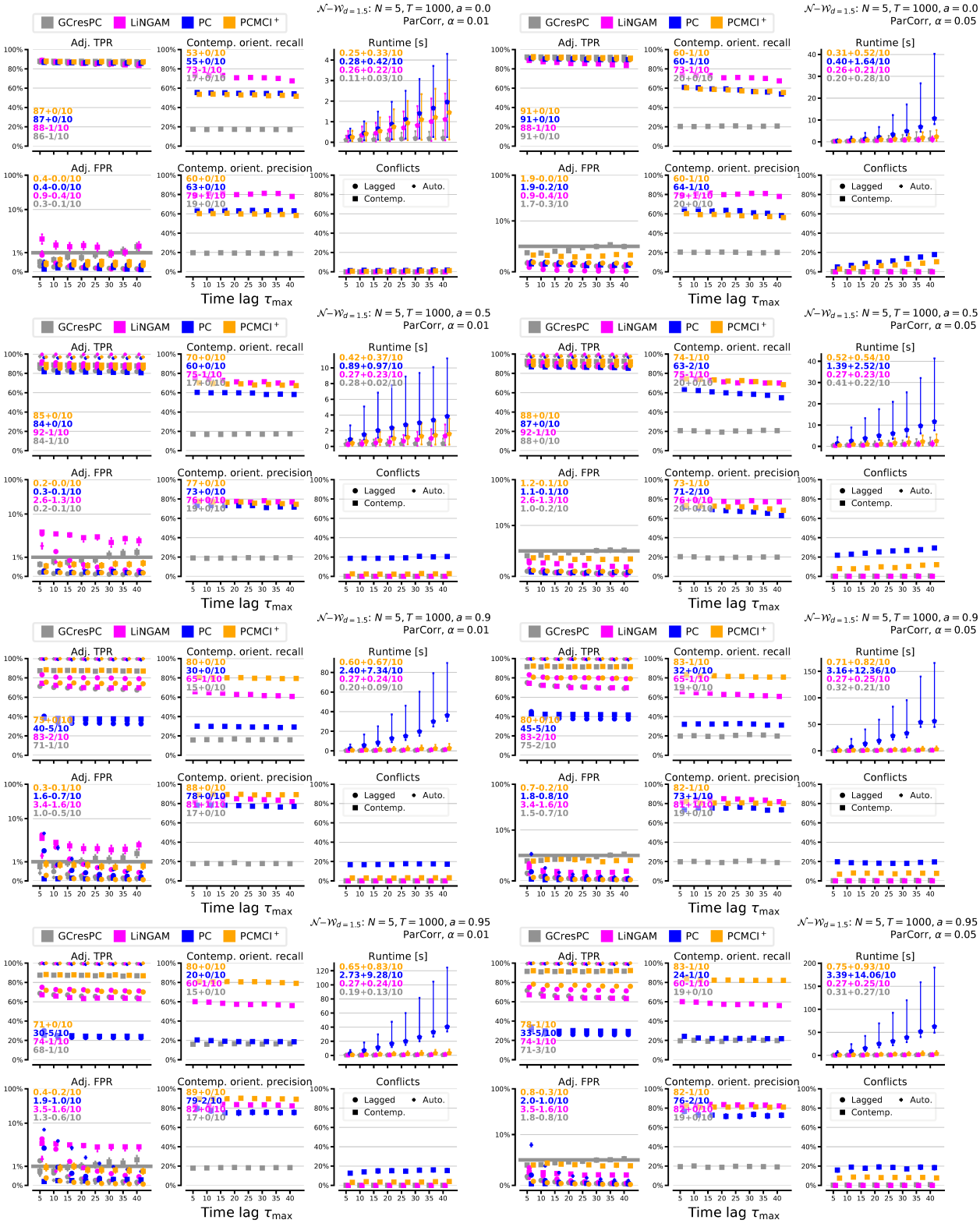

Figure S27: Numerical experiments with linear mixed noise setup for varying maximum time lag $\tau_{\max}$ and $T = 1000$. The left (right) column shows results for significance level $\alpha = 0.01$ ($\alpha = 0.05$). The rows depict results for increasing autocorrelations a (top to bottom). All model and method parameters are indicated in the upper right of each panel.

References

[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.

[Hyvärinen et al., 2010] Hyvärinen, A., Zhang, K., Shimizu, S., and Hoyer, P. O. (2010). Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11(May):1709–1731.

[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.

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