Causal Discovery for time series from multiple datasets with latent contexts (Supplementary Material)

Wiebke Günther[1] **Urmi Ninad**[2,1] **Jakob Runge**[1,2]

[1]German Aerospace Center, Institute of Data Science, 07745 Jena, Germany
[2]Technische Universität Berlin, Dept. of Electrical Engineering and Computer Science, 10623 Berlin, Germany

In this Supplementary Material, we provide some subtleties of the target graph of J-PCMCI$^+$, the dummy-projection and deletion operations, the embedding and representation of the dummy variable and how to relax some of the assumptions. Furthermore, we give background on the challenge of determinism within causal discovery, illustration on how context nodes can help with orienting additional edges, proofs for the main theorems, additional pseudocode and details on the simplified experimental setup as well as additional plots for the numerical experiments.

A CODE

The code to reproduce the experimental results can be found under the following url `https://github.com/guenwi/J-PCMCIplus`. The method (J-PCMCI$^+$) will also be made available as part of the tigramite package (`https://github.com/jakobrunge/tigramite`).

B MORE ON DUMMY-PROJECTION AND -DELETION

B.1 THE TARGET GRAPH OF J-PCMCI+

We define the "target graph" as our ultimate object of interest, which is the causal graph between the system nodes. By extension, it is implied that the links between the context nodes as well as those between the dummy and system nodes aren't of interest, the latter additionally so because the dummy variable is not a causal variable. To make this more tangible, we provide additional illustration of the target graph in relation to the dummy-projection and dummy-deleted version of the ground truth graph in figure 1.

B.2 DUMMY CONFOUNDING

A misleading fact about the dummy projection, and also of $\mathcal{G}_{alg}$, which is the result of algorithm J-PCMCI$^+$ (J-PC, respectively), is that it can contain system variables confounded by the dummy, that do not correspond to actual latent confounding, see figure 3 for a visualization of one such case. However, as we prove in Section 4.3, this is not a concern because we are interested in the true causal graph over the system variables together with edges from context to system variables, and for this task conditioning on such a dummy that isn't a true confounder doesn't lead to wrong inferences.

Furthermore, note that we include the time-dummy D_{time} only once into the time series graph. Since the time-dummy does not contain information on the specific value of the unobserved context variables but only encodes expert knowledge on their structure, we are not able to discover at which specific lag the causal relationship between the latent temporal context variables and the system variables occurs. However, we are able to find whether the system variables are influenced by a temporal context variable or not. See figure 2 for a visualization.

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

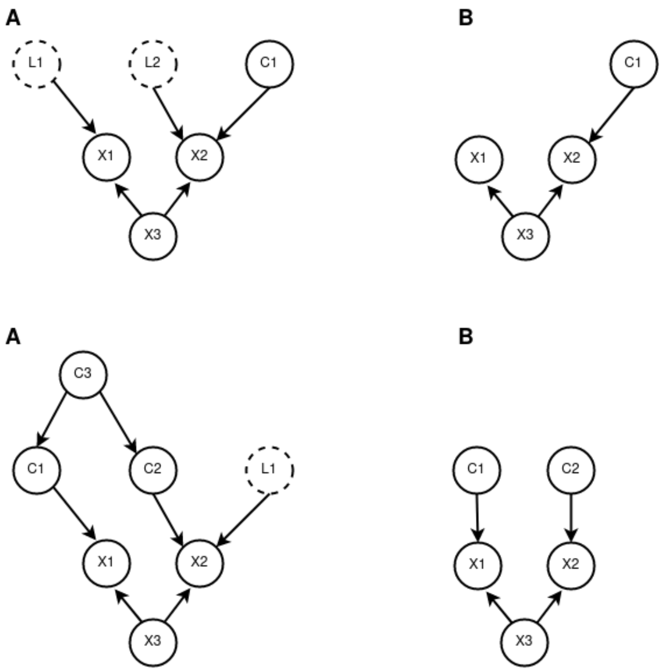

Figure 1: Visualization of the summary graphs of SCM (1) (A), as well as the corresponding target graph (B). The node D in the dummy projected graph can be either the time dummy D_{time} or the space dummy D_{space}. Latent context nodes are visualized using dashed circles, dashed arrows denote deterministic dependencies (not part of dummy projection).

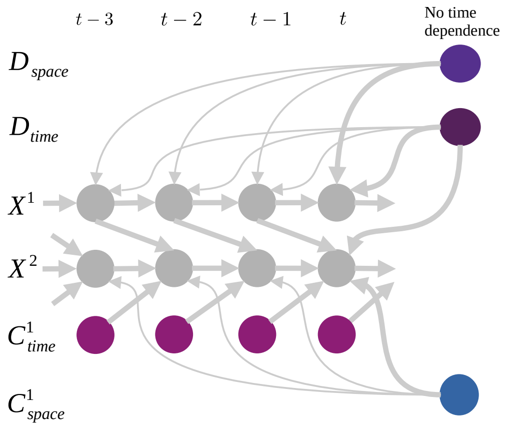

Figure 2: Unrolled time series graph with system variables X^1, X^2, an observed temporal context variable C^1_{time} with (possibly lagged) links to system variables, an observed spatial context variable C^1_{space} as a single node as it's constant over time, space and time dummy D_{space} and D_{time} as single nodes. Note that this graph omits links between context variables and between dummy and context variables. For better readability, we used thinner arrows for lagged links from the context and dummy variables to the system variables.

B.3 EMBEDDING OF THE DUMMY VARIABLES

Here, we provide further detail on how to represent or encode the dummy variable. The choice of embedding matters most for testing conditional independence between dummy and system variables. Potential choices of embeddings include one-hot encoding, which we use, or using the integers that denote the time or data set index directly (as done in [Huang et al., 2020]), among others.

In CD-NOD [Huang et al., 2020], the auxiliary variable corresponds to the domain or time index, i.e., the dummy takes values in $\{1, \dots, n_C\}$, where n_C is the number of contexts. To then be able to test for marginal and conditional independence between a system variable X and the dummy D, Huang et al. [2020] employ the KCI test [Zhang et al., 2011] since the functional relationship between D and X is highly non-linear. In case of non-stationary data, they also assume that all temporal context variables are a smooth function of the time-dummy, thus they also need to keep the time order in their embedding of the dummy (which we do not do). When using a one-hot encoded dummy in combination with a partial correlation test for testing whether a system variable depends on the context, we are essentially testing for differences in mean of that system variable as the context changes, since the partial correlation coefficient reduces to the point-biserial correlation coefficient Sheskin [2020] if one of the variables is dichotomous. If we would adapt the CI test, the same could be achieved with an integer-embedding of the dummy.

Furthermore, the choice of embedding also has implications on how easy it is to regress out context information from the system variables when testing system-system adjacencies conditional on the dummy. Using the one-hot encoding of the dummy values, we are centering the system data within each dataset or across time. This is very related to the well-established technique of fixed effects panel regression.

B.4 RELAXING THE NO-MEDIATION ASSUMPTION

It is possible to relax the no-mediation assumption, which is part of Assumption 2. However, this will result in the context-system links as discovered by J-PCMCI$^+$ representing ancestral rather than direct causal relationships. A consequence of that is that the graph $\mathcal{G}_{alg}$ that corresponds to the (time-series) graph resulting from algorithm J-PCMCI$^+$ is not identical to the dummy-projected graph $\mathcal{G}_D$ of the ground truth graph $\mathcal{G}$. In particular $\mathcal{G}_{alg}$ will have more system-context links than those in $\mathcal{G}_D$. Such link could appear because of an observed context variable that is indeed a parent of the system variable, i.e., links that also appear in $\mathcal{G}_D$. However, it could also happen that a latent context L is a mediator between an observed context variable and a system variable. Since we cannot condition on the dummy in the first step of the algorithm (due to the deterministic relationship between dummy and context), the algorithm will not remove this link. However, it is not contained in the dummy projected ground truth graph because there is no actual link to system from the observed context variable.

Consequently, the consistency theorem 2 does no longer hold since it relies on the dummy-projection of the ground truth graph which does not include ancestral links between context and system variables that are mediated by a latent context. The theorem would be adapted to a new definition of the dummy-projection that includes such links.

C FAITHFULNESS VIOLATION DUE TO DETERMINISM

Causal inference on data containing deterministic relationships and the challenges thereof have been dealt with previously, e.g. [Daniusis et al., 2012], [Lemeire et al., 2011]. Let us look at an example which was taken from Lemeire et al. [2012] to illustrate the challenge that is introduced by deterministic relationships.

Let $X \rightarrow Y \rightarrow Z$ be the ground truth causal graph over the variables X, Y, Z, where there is a deterministic relation between Y and X, i.e. there exists a function $f(\cdot)$ s.t. $Y = f(X)$. Therefore, $Y \perp\!\!\!\perp Z|X$ since X contains all information about Y.

This illustrates that deterministic relations generate additional independencies beyond those implied by the Markov condition. In other words, the true DAG is not faithful to the joint probability distribution of X, Y, Z. To describe these additional independencies, the D-separation criterion has been introduced [Geiger et al., 1990]. It is worth noting that the PC algorithm will not be sound when deterministic relationships are present. In the above example, the algorithm will remove both the edge between Y and Z, and also between X and Z. This will happen because the conditional independence $Y \perp\!\!\!\perp Z|X$ (wrongly) suggests that X separates Y and Z. On the other hand, $X \perp\!\!\!\perp Z|Y$ due to d-separation.

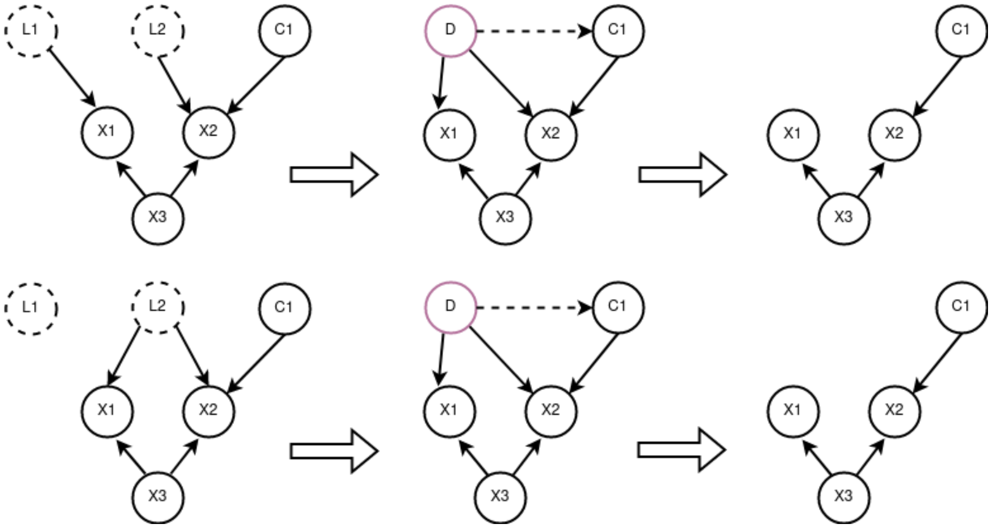

Figure 3: Visualization of the dummy projection operator (middle) and the dummy deletion (right) on summary graphs of SCM (1). The node D can be either the time dummy D_{time} or the space dummy D_{space}. Latent context nodes are visualized using dashed circles, dashed arrows denote deterministic dependencies (not part of dummy projection). The first row indicates that a dummy confounder does not correspond to a real latent confounder, while in the second row it does.

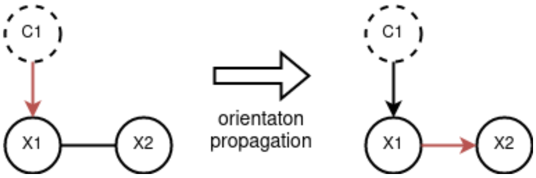

Figure 4: Visualization of how the context variables help orient additional edges by making use of the assumption that context nodes are exogenous to the system (left) and the standard rules of orientation propagation (right).

D CONTEXT NODES HELP IN ORIENTING EDGES

We quickly recap how context variables help in orienting additional system-system links. We consider the situation where $C \rightarrow X - Y$ and there is no edge between C and Y. Then $C \rightarrow X - Y$ forms an unshielded triple. Then we can use standard collider orientation rules to orient the edge between X and Y. In more detail:

(i) If Y and C are independent given a set of variables that does not include X, then the triple is a V-structure, and we have $X \leftarrow Y$.

(ii) Otherwise, if Y and C are independent given a set of variables including X, then we have $X \rightarrow Y$.

See figure 4 for a visualization.

E PROOFS

E.1 PROOF OF THEOREM 1

Before we get to the proof of Theorem 1, we note that the following useful lemma holds for the non-time series case.

Lemma 1. *For two system variables X and Y, it holds for any $S \subset \mathbf{X}$*

$$X \perp\!\!\!\perp Y | S \cup \{D\} \quad \Longleftrightarrow \quad X \perp\!\!\!\perp Y | S \cup \mathbf{C} \cup \mathbf{L}.$$

For a definition of the sets $\mathbf{X}$, $\mathbf{C}$ and $\mathbf{L}$, please refer to the main text. Note that in the non-time series case, there is no time dummy, therefore $D := D_{\text{space}}$.

Proof. This was already shown in the proof of Theorem 1 by Huang et al. [2020] but, for convenience, we repeat the arguments here. All system variables can be expressed as functions of $\mathbf{C}$, $\mathbf{L}$, and the noise. Therefore, the conditional distribution of the system given the dummy (i.e. the distribution within each data set) $P(\mathbf{X}|D)$ is determined by the joint distribution of the noise, and the observed and latent context variables $\mathbf{C} \cup \mathbf{L}$. This implies $P(X, Y|S \cup \mathbf{C} \cup \mathbf{L} \cup \{D\}) = P(X, Y|S \cup \mathbf{C} \cup \mathbf{L})$ where $S \subset \mathbf{X}$ (since the noise is independent of D). Then by recalling the weak union property of conditional independece as well using the fact that $\mathbf{L}$ and $\mathbf{C}$ are deterministic functions of D, it follows that, $X \perp\!\!\!\perp Y|S \cup \{D\}$, i.e. $P(X, Y|S \cup \{D\}) = P(X|S \cup \{D\})P(Y|S \cup \{D\})$ is equivalent to $P(X, Y|S \cup \mathbf{C} \cup \mathbf{L}) = P(X|S \cup \mathbf{C} \cup \mathbf{L})P(Y|S \cup \mathbf{C} \cup \mathbf{L})$.

□

We now recall theorem 1.

Theorem 1 (Non-time series consistency result)**.** *Denote the output of J-PC (Algorithm 1 in the main text) as $\mathcal{G}_{alg}$. Under the assumptions 1, 2, 3, 4, and assuming consistent conditional independence tests are used, the dummy deletion of $\mathcal{G}_{alg}$ corresponds to the dummy-deleted ground truth graph as the number of data sets M tends to infinity.*

Proof. Let us denote the skeleton of the projected ground truth graph with deleted dummy nodes by $\mathcal{G}^*$. Similarly, we denote the skeleton of the dummy-deleted output of the algorithm by $\hat{\mathcal{G}}^*$. We call their dummy-projected version of the ground truth graph $\mathcal{G}$, and the output of the algorithm (which is essentially a dummy projection) $\mathcal{G}_{alg}$.

First, we prove soundness of the algorithm, in other words we need to show that $\hat{\mathcal{G}}^* = \mathcal{G}^*$.

The soundness of context-system links follows from the soundness of the PC algorithm on the subset of system and observed context nodes. Let $X \in \mathbf{X}$ and $C \in \mathbf{C}$. The algorithm removes a link iff $X \perp\!\!\!\perp C|S$ where $S \subset \mathbf{X} \cup \mathbf{C}$. Then Faithfulness (w.r.t. ground truth graph to $P_m(X, C, L)$) implies that all links not in $\hat{\mathcal{G}}^*$ are also not in $\mathcal{G}^*$.
We also need to show that any context-system links that are not in $\mathcal{G}^*$ are also not in $\hat{\mathcal{G}}^*$. If the link between X and C is not in $\mathcal{G}^*$, then $X \perp\!\!\!\perp C|S$ where $S \subset \mathbf{X} \cup \mathbf{C} \cup \mathbf{L}$. Using the assumption that system and context (Assumption 3) are not confounded by latent variables and latent context nodes cannot be mediators between system and context, this is equivalent to $X \perp\!\!\!\perp C|S$ where $S \subset \mathbf{X} \cup \mathbf{C}$. This is tested at some iterative step of the PC-algorithm, and consequently the link is removed.

Since the dummy-deleted graphs do not contain any links to the dummy, we need to show soundness for the system-system links. However, within that it is needed that we find the correct dummy-system links. In other words we first show that if the link $D - X$ is not in $\mathcal{G}_{alg}$, then it also is not in $\mathcal{G}$.
If the link $D - X$ is not in $\mathcal{G}_{alg}$, then $X \perp\!\!\!\perp D|S \cup \text{Pa}_C(X)$ where $S \subset \mathbf{X}$. This implies that for all latent context nodes L holds $X \perp\!\!\!\perp L|S \cup \text{Pa}_C(X)$ since L can be expressed as a function of D, i.e. there exists a function g with $L = g(D)$. Therefore by Faithfulness and the non-invertibility of the function g, there is also no link between X and L in the ground truth graph, and thus also no link to the dummy in its projected version $\mathcal{G}$.
For the other direction, we need to show that if the link $D - X$ is not in $\mathcal{G}$, then it also is not in $\mathcal{G}_{alg}$. If the link $D - X$ is not in $\mathcal{G}$, then for all latent nodes L it holds $L - X$ is not in the ground truth graph. By the Causal Markov Condition, it holds $X \perp\!\!\!\perp L|\text{Pa}(X)$ for all L. This also implies that $\text{Pa}_L(X)$ is empty. We also know that X can be expressed as a function of the context nodes $\mathbf{C}$, $\mathbf{L}$ and the noise. This means, conditional on $\text{Pa}(X)$, X only depends on the noise. The noise is independent of D, thus $X \perp\!\!\!\perp D|\text{Pa}(X)$. Therefore, the algorithm will remove this dummy-system link.

Now, we prove soundness for the system-system links. The algorithm removes a link iff $X \perp\!\!\!\perp Y|S \cup \text{Pa}_{CD}(X)$ where $S \subset \mathbf{X}$, $\text{Pa}_{CD}(X)$ dummy and contextual parents of X. Note that by Assumption 4 there exist functions g^i, h^j s.t. $L^i = g^i(D)$ and $C^j = g^j(D)$, and thus $P(X, Y|S \cup \mathbf{C} \cup \mathbf{L} \cup \{D\}) = P(X, Y|S \cup \{D\})$.
So, if $D \in \text{Pa}_{CD}(X)$ this yields, together with Lemma 1

$$X \perp\!\!\!\perp Y|S \cup \text{Pa}_{CD}(X) \implies X \perp\!\!\!\perp Y|S \cup \{D\} \implies X \perp\!\!\!\perp Y|S \cup \mathbf{C} \cup \mathbf{L}.$$

And Faithfulness (of ground truth graph to $P_m(X, C, L)$) implies that this link is not in $\mathcal{G}^*$. If $D \notin \text{Pa}_{CD}(X)$, Faithfulness is directly applicable.

It remains to show that system-system links not in $\mathcal{G}^*$ are also not in $\hat{\mathcal{G}}^*$. If the link between X and Y is not in $\mathcal{G}^*$, then $X \perp\!\!\!\perp Y|S$ where $S \subset \mathbf{X} \cup \mathbf{C} \cup \mathbf{L}$, and also $X \perp\!\!\!\perp Y|S \cup \text{Pa}_C(X) \cup \text{Pa}_L(X)$ where $S \subset \mathbf{X}$. Again, we distinguish two cases:

First note that by what we proved above $\text{Pa}_{CD}(X)$ in the dummy-projection $\mathcal{G}$ is a subset of $\text{Pa}_{CD}(X)$ in the dummy-projection $\mathcal{G}_{alg}$ (i.e. it might not contain the dummy if $\mathcal{G}_{alg}$ has a dummy link to X). If $D \notin \text{Pa}_{CD}(X)$ in $\mathcal{G}$, then there exists $V \in \mathbf{C} \cup \mathbf{X}$ with $X \perp\!\!\!\perp L|V$ (since L is a function of D), i.e. $\text{Pa}_L(X) = \emptyset$, and thus $X \perp\!\!\!\perp Y|S \cup \text{Pa}_C(X)$, but also $X \perp\!\!\!\perp Y|S \cup \text{Pa}_C(X) \cup \{D\}$. So, in any case (if D is a parent of X in $\mathcal{G}_{alg}$ or not) the algorithm removes the link.
On the other hand, if $D \in \text{Pa}_{CD}(X)$ in $\mathcal{G}$, we use that $X \perp\!\!\!\perp Y|S \cup \text{Pa}_C(X) \cup \text{Pa}_L(X)$ is equivalent to $X \perp\!\!\!\perp Y|S \cup \mathbf{C} \cup \mathbf{L}$ which is equivalent to $X \perp\!\!\!\perp Y|S \cup \{D\}$. If this holds, then also $X \perp\!\!\!\perp Y|S \cup \{D\} \cup \text{Pa}_C(X)$, which implies $X \perp\!\!\!\perp Y|S \cup \text{Pa}_{CD}(X)$. So, also in this case the algorithm will remove the link. This concludes the soundness proof.

Completeness follows from soundness of the context-system, dummy-system and system-system links proved above, and the completeness of the PC-algorithm under tiered background knowledge Andrews et al. [2020]. □

E.2 PROOF OF THEOREM 2

We extend Lemma 1 to the time series case.

Lemma 2. *Let $\mathbf{X}$ be time series data. Define $\mathcal{B}^-_{XY} := (\hat{\mathcal{B}}^-_t(Y_t) \setminus \{X_{t-\tau}\}), \hat{\mathcal{B}}^-_{t-\tau}(X_{t-\tau})$ where $\hat{\mathcal{B}}^-_t(X_t)$ denotes the lagged adjacency set resulting from the lagged skeleton phase of PCMCI$^+$ (Algorithm 1 in Runge [2020]). For two system variables $X_{t-\tau}$ and Y_t, it holds for any $S \subset \mathbf{X}$*

$$X_{t-\tau} \perp\!\!\!\perp Y_t|S, \mathcal{B}^-_{XY}, D_{\textit{time}}, D_{\textit{space}} \quad \Longleftrightarrow \quad X_{t-\tau} \perp\!\!\!\perp Y_t|S, \mathcal{B}^-_{XY}, \mathbf{C}, \mathbf{L},$$

and

$$X_{t-\tau} \perp\!\!\!\perp Y_t|S, \mathcal{B}^-_{XY}, D_{\textit{space}} \quad \Longleftrightarrow \quad X_{t-\tau} \perp\!\!\!\perp Y_t|S \cup \mathbf{C}_{\textit{space}}, \mathcal{B}^-_{XY}, \mathbf{L}_{\textit{space}},$$

as well as

$$X_{t-\tau} \perp\!\!\!\perp Y_t|S, \mathcal{B}^-_{XY}, D_{\textit{time}} \quad \Longleftrightarrow \quad X_{t-\tau} \perp\!\!\!\perp Y_t|S, \mathcal{B}^-_{XY}, \mathbf{C}_{\textit{time}}, \mathbf{L}_{\textit{time}}.$$

Proof. The following equation follows exactly in the same way as Lemma 1. Since the observed and latent context variables are either space- or time-dependent, this also works for the space and time dimension separately. □

Theorem 2 (Time series consistency result)**.** *Denote the time series graph output of J-PCMCI$^+$ (Algorithm 2 in the main text) as $\mathcal{G}_{alg}$. Under assumptions 1, 2, 3, 4, and assuming consistent conditional independence tests are used, the dummy deletion of $\mathcal{G}_{alg}$ corresponds to the target graph (definition 1) as the number of data sets M and the number of times steps T tend to infinity.*

Proof. Let us denote the skeleton of the projected ground truth time series graph with deleted dummy nodes by $\mathcal{G}^*$. Similarly, we denote the skeleton of the dummy-deleted time series graph output of the algorithm by $\hat{\mathcal{G}}^*$. We call their dummy-projected version of the ground truth graph $\mathcal{G}$, and the output of the algorithm (which is essentially a dummy projection) $\mathcal{G}_{alg}$.

Soundness:
First, note that the lagged phase returns a set that always contains the parents of X^j_t by Lemma S1 in Runge [2020]. This still holds if latent context confounders are present, only additional links are possible.

Now, we show soundness of the system-context links. If $X^j_t - C^i_{t-\tau}$ not in $\hat{\mathcal{G}}^*$, then by Faithfulness also $X^j_t - C^i_{t-\tau}$ not in $\mathcal{G}^*$.
For the other direction,if the link between X^j_t and $C^i_{t-\tau}$ is not in $\mathcal{G}^*$, due to the Causal Markov Condition it holds $(X^i_{t-\tau}, W^-_t) \perp\!\!\!\perp C^j_t|\text{Pa}(C^j_t)$. Define $W^-_t := (\hat{\mathcal{B}}^-_t(X^j_t) \setminus \{C^i_{t-\tau}\}), \hat{\mathcal{B}}^-_{t-\tau}(C^i_{t-\tau}) \setminus \text{Pa}(X^j_t)$ as in Runge [2020] where $\hat{\mathcal{B}}^-_t(X_t)$ denotes the lagged adjacency set resulting from the lagged skeleton phase of PCMCI$^+$ (Algorithm 1 in Runge [2020]). Using the weak union property of conditional independence this implies $X^i_{t-\tau} \perp\!\!\!\perp C^j_t|\text{Pa}(C^j_t), W^-_t$ which is, by definition of W^-_t equivalent to $X^i_{t-\tau} \perp\!\!\!\perp C^j_t|\text{Pa}(C^j_t), \hat{\mathcal{B}}^-_t(C^j_t) \setminus \{X^i_{t-\tau}\}, \hat{\mathcal{B}}^-_{t-\tau}(X^i_{t-\tau})$. Note that $\text{Pa}(C^j_t) \subset \mathbf{C} \cup \mathbf{L}$, however by Assumptions 2, there exist no latent confounders or mediators between system and context, thus we also find a set $S \subset \mathbf{C}$ s.t. $X^i_{t-\tau} \perp\!\!\!\perp C^j_t|S, \hat{\mathcal{B}}^-_t(C^j_t) \setminus \{X^i_{t-\tau}\}, \hat{\mathcal{B}}^-_{t-\tau}(X^i_{t-\tau})$. This is tested at some iterative step of the algorithm and the link is removed.

Even though, eventually we are only interested in the soundness of system-context and system-system links, we need to establish that the dummy-system links within $\mathcal{G}_{alg}$ correspond to those in the dummy-projected ground truth graph $\mathcal{G}$. In the following, D can either denote D_{time} or D_{space}.
Now, we show if the link $D - X^j_t$ is not in $\mathcal{G}_{alg}$, then it also is not in $\mathcal{G}$. If the link $D - X^j_t$ is not in $\mathcal{G}_{alg}$, then

$D \perp\!\!\!\perp X^j_t | \mathbf{S}, \hat{\mathcal{B}}^C_t(X^j_t)$. This conditional independence also holds for all latent context nodes L since it can be expressed as a non-invertible function of D. Therefore by Faithfulness and the non-invertibility of this function, there is also no link between X^j_t and L in the ground truth graph, and thus also no link to the dummy in its projected version $\mathcal{G}$.

For the other direction, let us define $W^-_t := \hat{\mathcal{B}}^-_t(X^j_t) \setminus \text{Pa}(X^j_t)$, the set W^-_t does not contain parents of X^j_t, it also does not contain any latent nodes. If the link $D - X^i_t$ is not in $\mathcal{G}$, then for all latent nodes L it holds $L - X^i_t$ is not in the ground truth graph. Thus by the Causal Markov Condition $(L, W^-_t) \perp\!\!\!\perp X^j_t | \text{Pa}(X^j_t)$, and by the weak union property and using the definition of W^j_t, we get $L \perp\!\!\!\perp X^j_t | \text{Pa}(X^j_t), \hat{\mathcal{B}}^-_t(X^j_t)$ for all $L \in \mathbf{L}$. This also implies that $\text{Pa}_L(X^j_t) = \emptyset$.
Also, similarly to the non time series case, X^j_t can be expressed as a function of the context nodes $\mathbf{C}$, $\mathbf{L}$ and the noise (and auto-correlation which is accounted for by $\hat{\mathcal{B}}^-_t(X^j_t)$). This means, conditional on $(\text{Pa}(X^j_t) \setminus \mathbf{L}) \cup \hat{\mathcal{B}}^-_t(X^j_t)$, X^j_t only depends on the noise. The noise is independent of D, thus $X^j_t \perp\!\!\!\perp D | \text{Pa}(X^j_t)$. Therefore, the algorithm will remove this dummy-system link.

Next, we show the soundness of the discovery of the system-system links.
We first show, if the link $X^i_{t-\tau} - X^j_t$ not in $\hat{\mathcal{G}}^*$ then it is also not in $\mathcal{G}^*$. Essentially, this follows with the same arguments as in non time series case combined with Faithfulness, but we will go through the arguments in more detail now.
To simplify the notation, we make the abbreviation $\mathcal{B} := \hat{\mathcal{B}}^-_t(X^j_t) \setminus \{X^i_{t-\tau}\}, \hat{\mathcal{B}}^-_{t-\tau}(X^i_{t-\tau})$. The algorithm removes the link between $X^i_{t-\tau}$ and X^j_t if and only if

$$X^i_{t-\tau} \perp\!\!\!\perp X^j_t | S, \mathcal{B}, \text{Pa}_{CD}(X^i_{t-\tau}, X^j_t)$$

for some $S \in \hat{\mathcal{A}}_t(X^j_t)$.
If $D_{\text{time}}, D_{\text{space}} \notin \text{Pa}_{CD}(X^i_{t-\tau}, X^j_t)$: Faithfulness is directly applicable.
Let now D be either D_{time} or D_{space}. If $D \in \text{Pa}_{CD}(X^i_{t-\tau}, X^j_t)$ this yields, together with Lemma 2

$$X \perp\!\!\!\perp Y | S, \mathcal{B}, \text{Pa}_{CD}(X, Y) \implies X \perp\!\!\!\perp Y | S, \mathcal{B}, \{D\} \implies X \perp\!\!\!\perp Y | S, \mathcal{B}, \mathbf{C}, \mathbf{L}.$$

and we can apply the Faithfulness argument.

Now we show the other direction, i.e. if $X^i_{t-\tau} - X^j_t$ not in $\mathcal{G}^* \implies X^i_{t-\tau} - X^j_t$ not in $\hat{\mathcal{G}}^*$. For that, we define $W^-_t := (\hat{\mathcal{B}}^-_t(X^j_t) \setminus \{X^i_{t-\tau}\}, \hat{\mathcal{B}}^-_{t-\tau}(X^i_{t-\tau}), \text{Pa}_C(X^i_{t-\tau})) \setminus \text{Pa}(X^j_t)$ similar to Runge [2020]. This set does not contain any parents of X^j_t and by the assumption also $X^i_{t-\tau}$ is not a parent of X^j_t. Furthermore, we assume that for $\tau = 0$, X^i_t is not a descendant of X^j_t (can be always achieved by exchanging the roles of X^i_t and X^j_t).
Then the Causal Markov Condition implies $(X^i_{t-\tau}, W^-_t) \perp\!\!\!\perp X^j_t | \text{Pa}(X^j_t)$ Using the weak union property this implies $X^i_{t-\tau} \perp\!\!\!\perp X^j_t | \text{Pa}(X^j_t), W^-_t$ which is, by definition of W^-_t, equivalent to

$$X^i_{t-\tau} \perp\!\!\!\perp X^j_t | \text{Pa}(X^j_t), \hat{\mathcal{B}}^-_t(X^j_t) \setminus \{X^i_{t-\tau}\}, \hat{\mathcal{B}}^-_{t-\tau}(X^i_{t-\tau}), \text{Pa}_C(X^i_{t-\tau}) \tag{1}$$

Note that the conditioning set potentially also contains nodes from $\mathbf{L}$ (but only in $\text{Pa}(X^j_t)$). This also implies

$$X^i_{t-\tau} \perp\!\!\!\perp X^j_t | \text{Pa}(X^j_t) \setminus \mathbf{L}, \hat{\mathcal{B}}^-_t(X^j_t) \setminus \{X^i_{t-\tau}\}, \hat{\mathcal{B}}^-_{t-\tau}(X^i_{t-\tau}), \text{Pa}_C(X^j_{t-\tau}), \{D_{\text{time}}, D_{\text{space}}\}, \tag{2}$$

and similarly if $\text{Pa}(X^j_t)$ only contains nodes from $\mathbf{L}_{\text{space}}$ which can be accounted for by additionally conditioning on D_{space} (same argument holds if we replace space by time). If there are no latent nodes in $\text{Pa}(X^j_t)$, then (1) is the same as $X^i_{t-\tau} \perp\!\!\!\perp X^j_t | \text{Pa}_{XC}(X^j_t), \hat{\mathcal{B}}^-_t(X^j_t) \setminus \{X^i_{t-\tau}\}, \hat{\mathcal{B}}^-_{t-\tau}(X^i_{t-\tau})$, in our algorithm we either test this or (2) and therefore remove the link.
If there are latent nodes in $\text{Pa}(X^j_t)$, then $D \in \text{Pa}_{CD}(X^j_t)$ within $\mathcal{G}$ and thus also in $\text{Pa}_{CD}(X^j_t)$ within $\mathcal{G}_{alg}$, so we test (2) and remove the link.

For system-system links completeness follows as in Runge [2020]. The context-system links are already correctly oriented by the exogeneity assumption (context cannot be a descendent of system). □

Corollary 1. *If some of the observed context variables are treated as unobserved, and the assumptions 1, 2, 3, 4 still hold, our method J-PCMCI$^+$ will recover the correct system-system adjacencies.*

Proof. This follows directly from theorem 2. □

F PSEUDOCODE

We present the pseudocodes for `poolData`, `partialSkeletonPC` and `partialContempSkeletonPCMCI+` below.

Algorithm 1: poolData (for non-time-series data)

For the time-series case we rely on the functionality supplied in Tigramite [Runge et al., 2019] to handle time-series data from multiple data-sets while keeping the time structure (in particular while using the sliding window approach to cunstruct time-series data for the lagged variables $X^i_{-\tau}$ for $\tau > 0$

```
Data: $M$ data-sets $\mathbf{X}$ containing observations of the same system (and for time-series case: temporal context) variables,
      $M$ observations of context variables $\mathbf{C}$ (one per data-set), optional: dummy variable with $M$ distinct values
Result: one data-set containing the pooled data
Let $N$ denote the number of system variables
Let $K$ denote the number of (observed) context variables
Let $T_m$ denote the sample-size of the system variables with $m = 1, \dots, M$
for $i$ in $1, \dots, N$ do
    concatenate $(X^{i,(m)})_{m=1,\dots,M}$
end
for $j$ in $1, \dots, K$ do
    construct array of context variable $C^j$ by repeating its $M$ values $T_m$ times
end
if data for the dummy variable $D$ is provided then
    construct array for the dummy variable $D$ by repeating its $M$ values $T_m$ times
end
return $(X^1, \dots, X^N, C^1, \dots, C^K, D)$
```

Algorithm 2: partialSkeletonPC

CI(X, Y, S) is some suitable conditional independence test

```
Data: Data $\mathbf{X}$, significance level $\alpha$, node pairs to consider $\mathcal{P}$, link knowledge $\mathcal{B}$
Result: graph $\mathcal{G}$
Form a graph $\mathcal{G}$ with information from $\mathcal{B}$, connect all other nodes with undirected links
Set $p = 0$
while any adjacent pairs $(X, Y)$ in $\mathcal{P}$ satisfy $|\mathcal{A}(X) \setminus \{Y\}| \geq p$ do
    Select an adjacent pair $(X, Y)$ from $\mathcal{P}$ with $|\mathcal{A}(X) \setminus \{Y\}| \geq p$
    Select $S \subset \mathcal{A}(X) \setminus \{Y\}$ with $|S| = p$
    p-value $\leftarrow$ CI$(X, Y, S)$
    if p-value $> \alpha$ then
        Delete link $X - Y$ from $\mathcal{G}$
        Store (unordered) sepset $(X, Y) = \mathcal{S}$
    end
end
return $\mathcal{G}$, sepset
```

Algorithm 3: partialContempSkeletonPCMCI+, small adaption of Algorithm 2 in Runge [2020]
$\mathrm{CI}(X, Y, S)$ is some suitable conditional independence test

```
Data: $M$ time-series data-sets $X^{(m)} = (X^{1,(m)}, \dots, X^{N,(m)})$ which can contain system, context and also dummy
      variables, indices of system variables $J$, max. time lag $\tau_{\text{max}}$, significance threshold $\alpha_{\text{PC}}$, $\hat{\mathcal{B}}_t^-(X_t^j)$ for all
      $X_t^j \in \mathbf{X}_t = (X_t^1, \dots, X_t^N)^1$, contextual parents $\mathcal{C}(X)$, pairs to consider $\mathcal{P}$
Result: graph $\mathcal{G}$, sepset
Form time series graph $\mathcal{G}$ with lagged links from $\hat{\mathcal{B}}_t^-(X_t^j)$ for all $X_t^j \in \mathbf{X}_t$, fully connect all contemporaneous system
  variables, i.e. add $X_t^i - X_t^j$ for all $X_t^i \neq X_t^j \in \mathbf{X}_t$ with $i, j \in J$, and set links between context and system according
  to $\mathcal{C}(X)$
Initialize contemporaneous adjacencies $\hat{\mathcal{A}}(X_t^j) := \hat{\mathcal{A}}_t(X_t^j) = \{X_t^i \neq X_t^j \in \mathbf{X}_t | X_t^i - X_t^j \text{ in } \mathcal{G}\}$
Let $p = 0$
while any adjacent pairs $(X_{t-\tau}^i, X_t^j)$ for $\tau \geq 0$ in $\mathcal{G}$ from $\mathcal{P}$ satisfy $|\hat{\mathcal{A}}(X_t^j) \setminus \{X_{t-\tau}^j\}| \geq p$ do
    Select new adjacent pair $(X_{t-\tau}^i, X_t^j)$ from $\mathcal{P}$ for $\tau \geq 0$ satisfying $|\hat{\mathcal{A}}(X_t^j) \setminus \{X_{t-\tau}^j\}| \geq p$
    while $(X_{t-\tau}^i, X_t^j)$ are adjacent in $\mathcal{G}$ and not all $S \subset \hat{\mathcal{A}}(X_t^j) \setminus \{X_{t-\tau}^j\}$ with $|S| = p$ have been considered do
        Choose new $S \subset \hat{\mathcal{A}}(X_t^j) \setminus \{X_{t-\tau}^j\}$ with $|S| = p$
        Set $\mathbf{Z} = (S, \hat{\mathcal{B}}_t^-(X_t^j) \setminus \{X_{t-\tau}^j\}, \hat{\mathcal{B}}_{t-\tau}^-(X_{t-\tau}^j)$
        $(X_{t-\tau}^i, X_t^j, \mathbf{Z}) \leftarrow \mathrm{poolData}((X_{t-\tau}^{i,(m)}, X_t^{j,(m)}, \mathbf{Z}^{(m)})_{m=1,\dots,M})$
        $(\text{p-value}, I) \leftarrow \mathrm{CI}(X_{t-\tau}^i, X_t^j, \mathbf{Z})$
        $I^{\min}(X_{t-\tau}^i, X_t^j) = \min(|I|, I^{\min}(X_{t-\tau}^i, X_t^j))$
        if p-value $> \alpha_{PC}$ then
            Delete link $X_{t-\tau}^i \to X_t^j$ for $\tau > 0$ (or $X_t^i - X_t^j$ for $\tau = 0$) from $\mathcal{G}$
            Store (unordered) sepset $(X_{t-\tau}^i, X_t^j) = \mathcal{S}$
        end
        Let $p = p + 1$ and re-compute $\hat{\mathcal{A}}(X_t^j)$ from $\mathcal{G}$ and sort by $I^{\min}(X_{t-\tau}^i, X_t^j)$ from largest to smallest
    end
end
return $\mathcal{G}$, sepset
```

G SIMPLIFIED EXPERIMENTAL SETUP

We want to understand the shape of the adjacency-FPR surface of our method better. From the numerical results of the standard setup it seems that for a fixed samplesize T the FPR goes up with the number of datasets. On the other hand, for a fixed number of datasets M, FPR goes down with increasing samplesize. A similar pattern is visible when simply applying PCMCI$^+$ on data where dummy variables have been included.

To this end, we simplify our experimental setup in the following way. We sample data from a specific version of the SCM (1):

$$\begin{aligned}
X_t^0 &:= 0.5X_t^1 + 0.5C_{\text{space}}^0 + 0.5C_{\text{space}}^1 + 0.5C_{\text{time},t-1}^0 + 0.5C_{\text{time},t-1}^1 + \eta^0 \\
X_t^1 &:= 0.5X_{t-1}^1 + 0.5C_{\text{space}}^0 + 0.5C_{\text{space}}^1 + 0.5C_{\text{time},t-1}^0 + 0.5C_{\text{time},t-1}^1 + \eta^1 \\
C_{\text{space}}^0 &:= \eta_{\text{space}}^0 \\
C_{\text{space}}^1 &:= \eta_{\text{space}}^1 \\
C_{\text{time},t}^0 &:= \eta_{\text{time}}^0 \\
C_{\text{time},t}^1 &:= \eta_{\text{time}}^1,
\end{aligned} \tag{3}$$

where C_{space}^1 and C_{time}^1 are unobserved, and all other variables are observed. On the system data of this SCM we apply a modified version of PCMCI$^+$ where we always including the dummy variables in the conditioning sets of all conditional independence tests. By doing so, we are able to see what effect conditioning on the dummies has on the FPR, see figure 5. In the FPR-plot, we see a similar pattern as is visible in the more involved experimental setup. Generally speaking, for a fixed samplesize T the FPR goes up with the number of datasets M while, for a fixed number of datasets M, it goes down with increasing samplesize T. This means, in the large sample limit (of both M and T) we can expect consistent results. However, if we only have a small samplesize T, there are potentially inflated false positives.

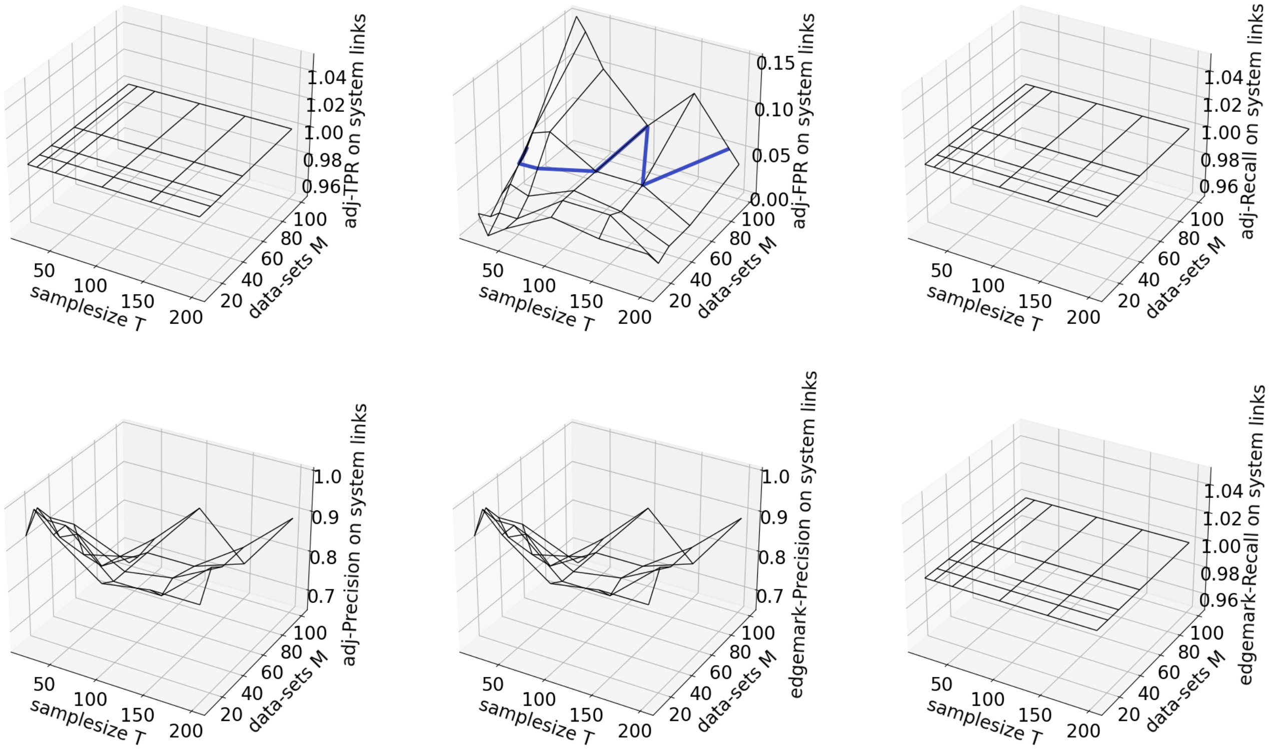

Figure 5: Discovery results of system-system links in the simplified experimental setup (section G) for varying sample sizes T, and number of datasets M.

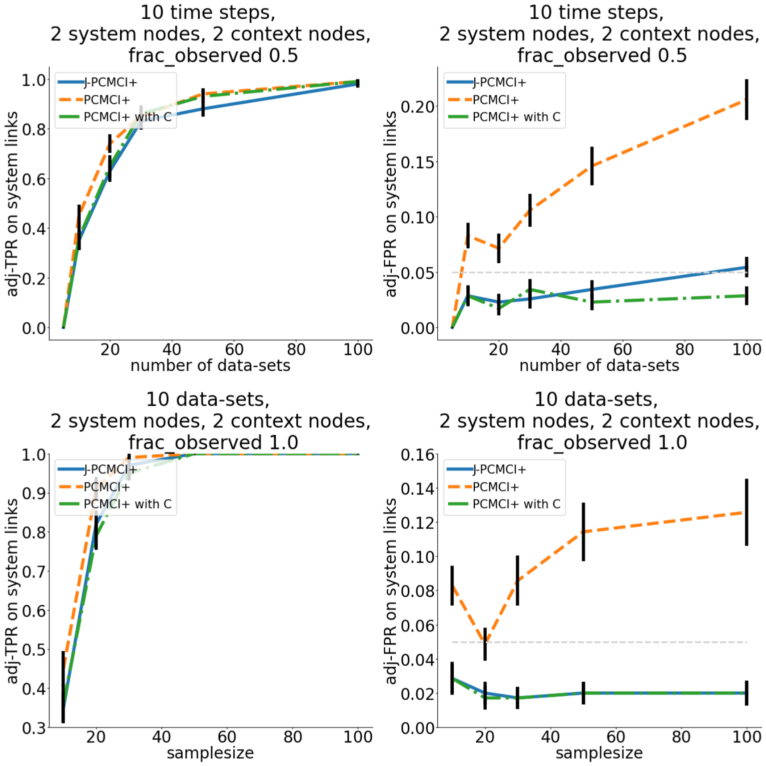

Figure 6: Discovery results of system-system links for varying sample sizes T, and fixed $M = 10$ (top row), and varying number of contexts M, and fixed $T = 10$ (bottom row). The data is generated according to the SCM described in section H. In this setting all of the context nodes are observed. We compare our method (J-PCMCI+) to PCMCI$^+$ using all data of observed nodes (PCMCI+ with C) and only using data of system variables (PCMCI+).

H NONLINEAR EXPERIMENTAL SETUP

We extend the simplified experimental setup of section G a bit to allow for nonlinear mechanisms. In this way, we are able to demonstrate that our method can be flexibly combined with any CI test. In this setup, we use a CI test based on Gaussian process regression and a distance correlation (GPDC).

$$\begin{aligned} X^0_t &:= 0.3(X^1_t)^2 + 0.5C^0_{\text{space}} - 0.2(C^0_{\text{time},t-1})^2 + \eta^0 \\ X^1_t &:= 0.5X^1_{t-1} - 0.5(C^0_{\text{space}})^2 + 0.3(C^0_{\text{time},t-1})^2 + \eta^1 \\ C^0_{\text{space}} &:= \eta^0_{\text{space}} \\ C^0_{\text{time},t} &:= \eta^0_{\text{time}} \end{aligned} \tag{4}$$

We show the results in figure 6.

I ADDITIONAL PLOTS

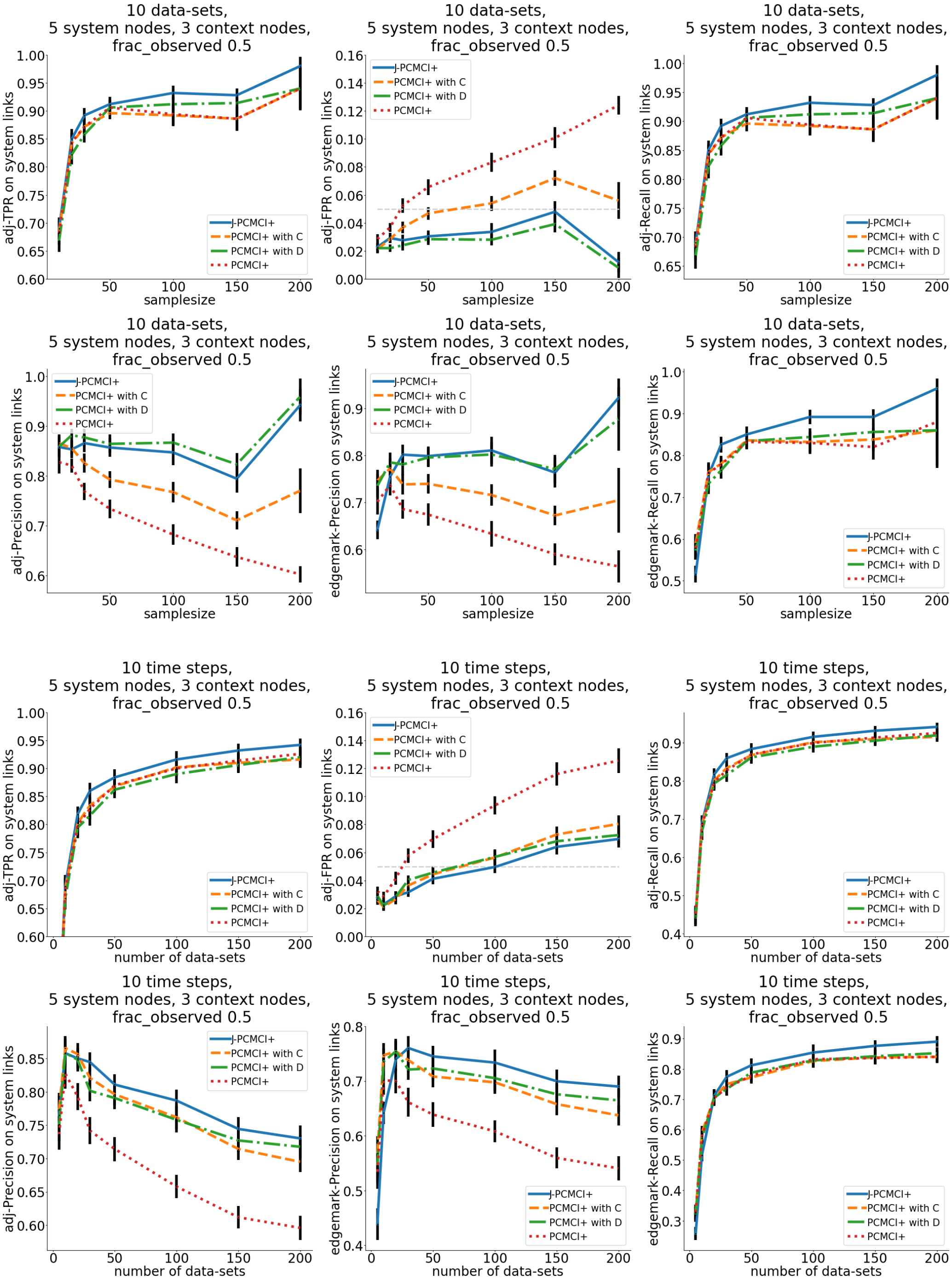

Figure 7: Discovery results of system-system links for varying sample sizes T, and fixed $M = 10$ (top two rows), and varying number of contexts M, and fixed $T = 10$ (bottom two rows). All other setup parameters are set as the defaults described in the main text. In this setting half of the context nodes are observed. We compare our method (J-PCMCI+) to PCMCI$^+$ using all data of observed nodes (PCMCI+ with C), using all data of system variables and including dummies (PCMCI+ with D), and only using data of system variables (PCMCI+).

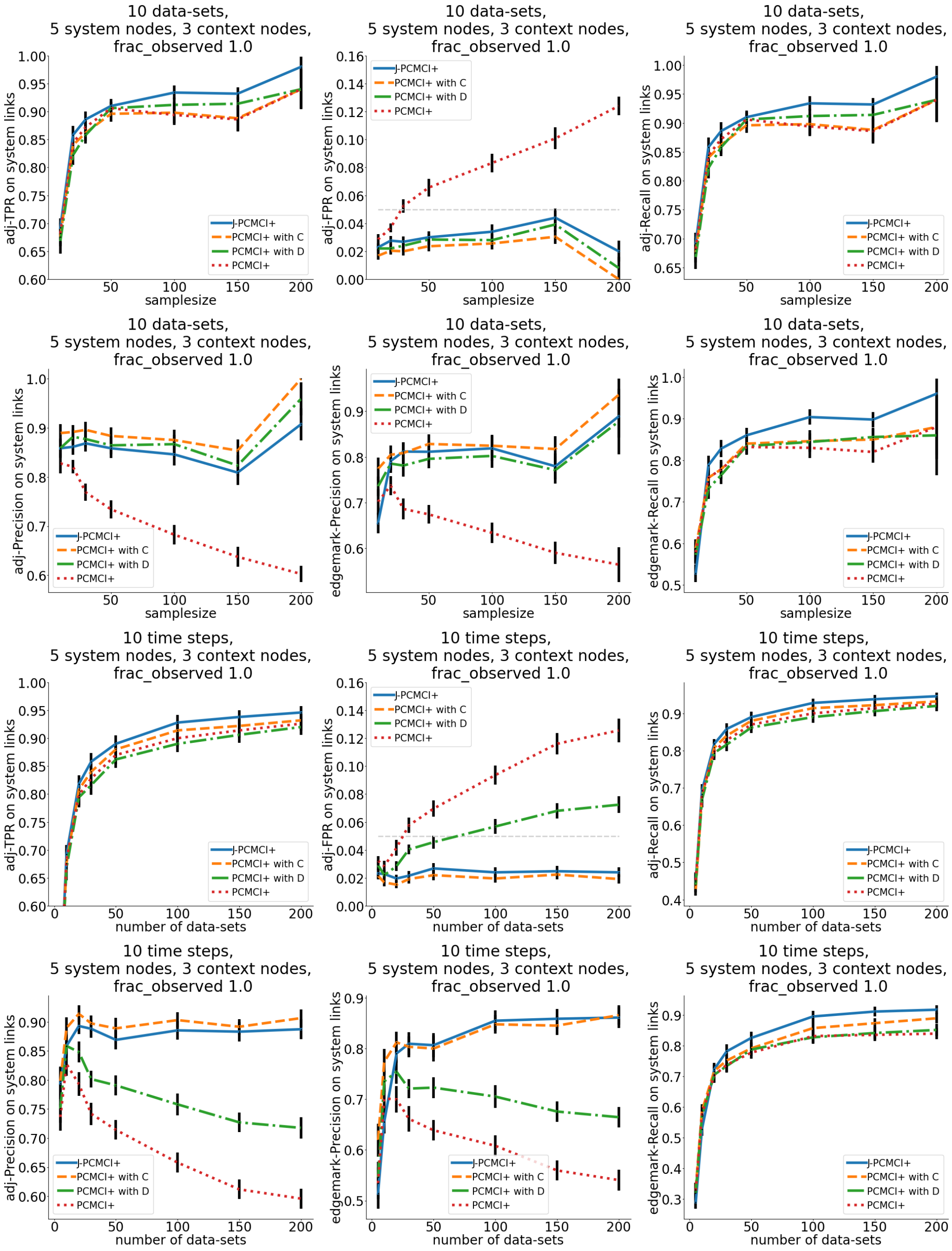

Figure 8: Discovery results of system-system links for varying sample sizes T, and fixed $M = 10$ (top two rows), and varying number of contexts M, and fixed $T = 10$ (bottom two rows). All other setup parameters are set as the defaults described in the main text. In this setting all of the context nodes are observed. We compare our method (J-PCMCI+) to PCMCI$^+$ using all data of observed nodes (PCMCI+ with C), using all data of system variables and including dummies (PCMCI+ with D), and only using data of system variables (PCMCI+).

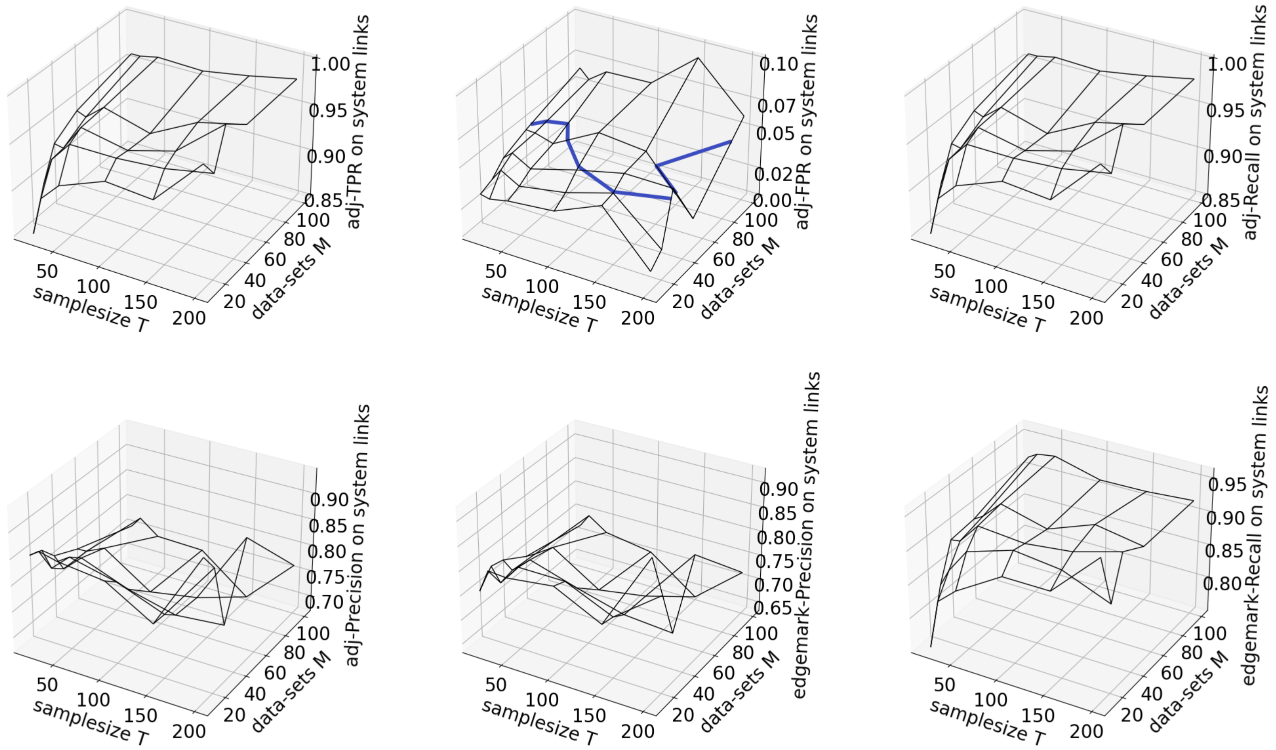

Figure 9: Discovery results of our method (J-PCMCI+) on system-system links for varying sample sizes T, and number of contexts M. All other setup parameters are set as the defaults described in the main text. In this setting half of the context nodes are observed. We show the contour line corresponding to the significance level α in the adjacency-FPR plot.

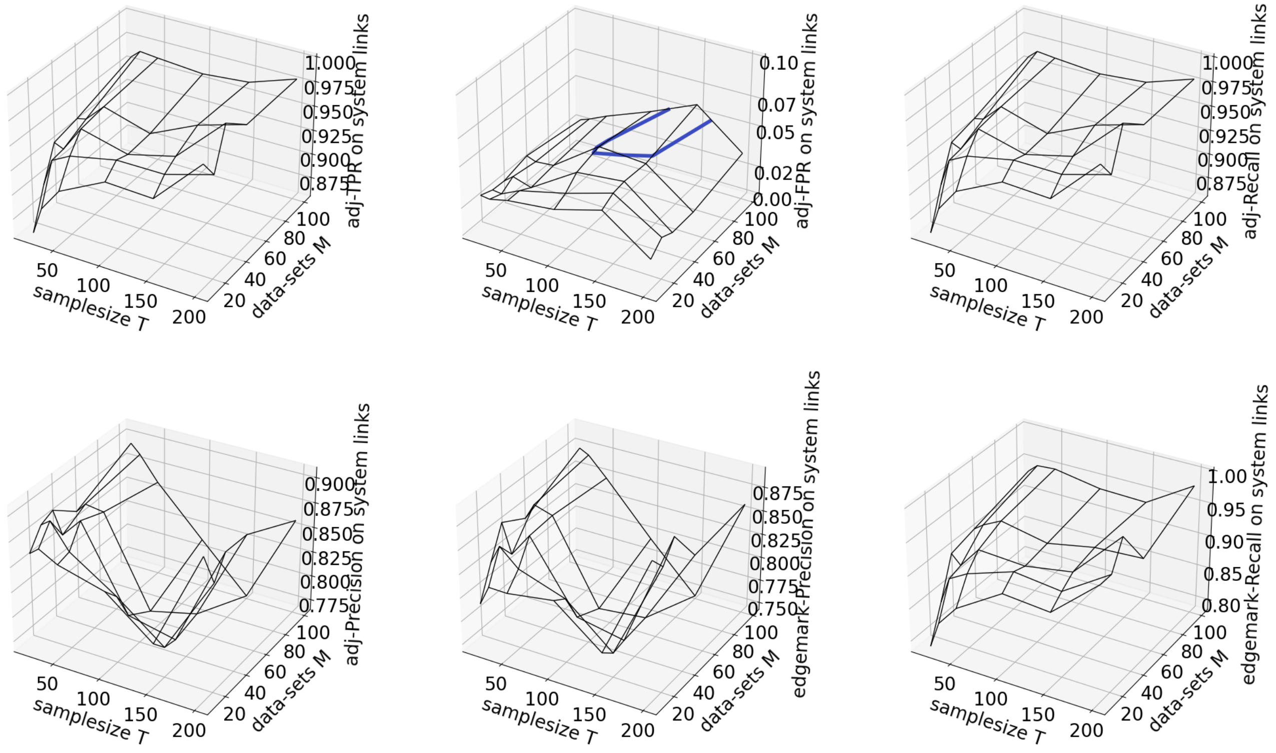

Figure 10: Discovery results of our method (J-PCMCI+) on system-system links for varying sample sizes T, and number of contexts M. All other setup parameters are set as the defaults described in the main text. In this setting all the context nodes are observed.We show the contour line corresponding to the significance level α in the adjacency-FPR plot.

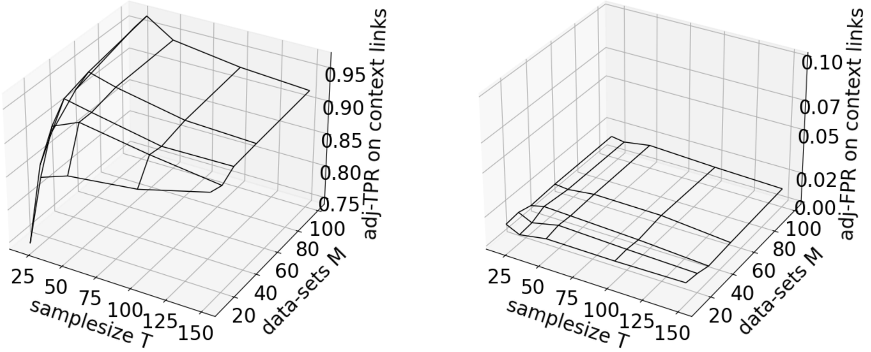

Figure 11: Discovery results of our method (J-PCMCI+) on context-system links for varying sample sizes T, and number of contexts M. All other setup parameters are set as the defaults described in the main text. In this setting all the context nodes are observed.

References

Bryan Andrews, Peter Spirtes, and Gregory F Cooper. On the completeness of causal discovery in the presence of latent confounding with tiered background knowledge. In *International Conference on Artificial Intelligence and Statistics*, pages 4002–4011. PMLR, 2020.

Povilas Daniusis, Dominik Janzing, Joris Mooij, Jakob Zscheischler, Bastian Steudel, Kun Zhang, and Bernhard Schölkopf. Inferring deterministic causal relations. *arXiv preprint arXiv:1203.3475*, 2012.

Dan Geiger, Thomas Verma, and Judea Pearl. Identifying independence in bayesian networks. *Networks*, 20(5):507–534, 1990.

Biwei Huang, Kun Zhang, Jiji Zhang, Joseph Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. *The Journal of Machine Learning Research*, 21(1):3482–3534, 2020.

Jan Lemeire, Stijn Meganck, Francesco Cartella, Tingting Liu, and Alexander R Statnikov. Inferring the causal decomposition under the presence of deterministic relations. In *ESANN*, 2011.

Jan Lemeire, Stijn Meganck, Francesco Cartella, and Tingting Liu. Conservative independence-based causal structure learning in absence of adjacency faithfulness. *International Journal of Approximate Reasoning*, 53(9):1305–1325, 2012.

Jakob Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Conference on Uncertainty in Artificial Intelligence*, pages 1388–1397. PMLR, 2020.

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

David J Sheskin. *Handbook of parametric and nonparametric statistical procedures*. crc Press, 2020.

K Zhang, J Peters, D Janzing, and B Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, pages 804–813. AUAI Press, 2011.