Causal Discovery for time series from multiple datasets with latent contexts

Abstract

Causal discovery from time series data is a typical problem setting across the sciences. Often, multiple datasets of the same system variables are available, for instance, time series of river runoff from different catchments. The local catchment systems then share certain causal drivers, such as time-dependent large-scale weather over all catchments, but differ in other catchment-specific drivers, such as the altitude of the catchment. These drivers can be called temporal and spatial contexts, respectively, and are often partially unobserved. Pooling the datasets and considering the joint causal graph among system, context, and certain auxiliary variables enables us to overcome such latent confounding of system variables. In this work, we present a non-parametric time series causal discovery method, *J(oint)-PCMCI*$^+$, that efficiently learns such joint causal time series graphs when both observed and latent contexts are present, including time lags. We present asymptotic consistency results and numerical experiments demonstrating the utility and limitations of the method.

1 INTRODUCTION

Causal discovery from observational data has gained widespread interest in recent years. Next to score-based methods Chickering [2002], Granger causality Granger [1969], and the more recent restricted structural causal models (SCM) framework Peters et al. [2017], Spirtes and Zhang [2016], the constraint-based approach[Spirtes et al., 2000] to this discovery task exploits conditional independencies in the data to constrain causal graphs and can flexibly handle nonlinear dependencies.

Most real world data comes in the form of *time series*, which provide opportunities and challenges for causal discovery [Runge et al., 2019a]. While the inherent time order implies certain causal directions, time-series data typically violates the *i.i.d.* assumption usually made in conditional independence testing. Therefore, specific algorithms that target the challenges of time-series data have become an increasingly popular sub-category within causal discovery, for instance, versions of the PC algorithm and FCI [Entner and Hoyer, 2010b, Malinsky and Spirtes, 2018], or the PCMCI framework [Runge et al., 2019b, Runge, 2020, Gerhardus and Runge, 2020].

The methods mentioned above consider single multivariate time series datasets and aggregate samples across time. Another relevant development has been the incorporation of multiple datasets and modeling their different contexts [Mooij et al., 2020, Huang et al., 2020], which can also be framed as a data-fusion problem [Pearl and Bareinboim, 2011, Bareinboim and Pearl, 2016].

In the following, we illustrate the main ideas of this paper on the example of time series datasets of river runoff from different catchment systems [Wagener et al., 2007]. If we assume all of these to come from the same (stationary) distribution, we can just concatenate (pool) the data to obtain a larger sample and, hence, more reliable causal discoveries among the system variables. But multiple datasets can also be used to de-confound relationships: The local catchment systems often share certain causal drivers, such as time-dependent large-scale weather dynamics over all catchments, that can be called temporal contexts and cause latent confounding between two or more system variables, if they are unobserved. If we now assume that such a time-dependent latent confounder is the same across all datasets, we can condition on the time index (or add a so-called *time-dummy variable*) by aggregating samples across datasets, instead of aggregating across time. This then yields a joint graph across all time points.

But datasets not only share common causal drivers, they also differ in other dataset-specific drivers, such as the altitude or vegetation-type of the catchment, which can be called

Submitted to the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023). **To be used for reviewing only**.

spatial contexts. As spatial contexts are constant across time, they do not constitute a confounding *within* each dataset. However, in the pooled data they vary across datasets and lead to confounding in the joint graph across datasets. If spatial contexts are observed, they can be included as variables in the analysis and if they are unobserved, they can be de-confounded by the same idea as time-dependent confounders by assuming that they are constant across all time points and conditioning on the dataset index (or adding a so-called *space-dummy variable*, not to be confused with physical space). The idea of conditioning on time and space is heavily employed in *fixed-effect panel regression models* in econometrics [Angrist and Pischke, 2009] and here we consider these for causal discovery.

Next to deconfounding system variables, observed context variables can help to orient causal links: Consider two system variables $X \circ\!\!-\!\!\circ Y$ whose causal direction cannot be identified by Markov equivalence. In the joint graph we could add a context variable C and assume (or learn) that $C \to X \circ\!\!-\!\!\circ Y$. Then applying the collider or orientation rules [Meek, 1995] allows to infer the causal direction between X and Y.

Our approach partially follows Mooij et al. [2020]. Their approach is to pool data from different contexts, for instance, observational and interventional data, and do causal discovery on the pooled dataset, called *joint causal inference* (JCI). In particular, they established a general framework to (i) interpret contexts as auxiliary variables that describe the context of each dataset, (ii) pool all the data from different contexts while keeping the contextual information of the data by including the auxiliary context variables into a single dataset, and (iii) apply standard causal discovery to all data jointly, incorporating appropriate background knowledge on the causal relationships involving the context variables.

Our aim is to extend the PCMCI$^+$ time series causal discovery algorithm [Runge, 2020] to the case of datasets from multiple dataset- or time dependent contexts. We term this technique *J(oint)-PCMCI$^+$* because it combines the two JCI-ideas mentioned above, i.e., pooling datasets from multiple contexts, and adding observed context variables to the graph. We go beyond JCI by using time- and space-dummy variables to account for latent context variables that confound system variables, that are conceptually similar to surrogate variables in [Huang et al., 2020]. This approach combines the advantages of PCMCI$^+$ regarding detection power and false positive control in the presence of strong autocorrelation [Runge, 2020] with the advantage of the JCI framework.

2 RELATED WORK

In causal inference, the idea of context variables has been explored under different notions : "policy variables" Spirtes et al. [2000], "force variables" Pearl [1993], "decision variables" in influence diagrams Dawid [2002], "selection variables" in selection diagrams Bareinboim and Pearl [2013], and "environment variables" Peters et al. [2016]. Mooij et al. [2020] established a general framework to combine data from multiple contexts with traditional causal discovery techniques.

In Huang et al. [2020], the problem of *heterogeneous* data, which might correspond to varying dataset collection conditions (analogous to contexts in JCI), as well as the problem of *non-stationary* data was addressed in a framework called CD-NOD. Here, changing causal mechanisms across time or datasets were interpreted as confounding of the system by an unobserved *pseudo-confounder*, so named because it can be written as a function of the dataset or the time index. This confounding was then addressed by introducing a *surrogate variable* that captures changes of causal mechanisms, thereby deconfounding the pseudo-confounded system variables. Further, they employed the information of changing causal mechanisms to infer additional causal directions than standard causal discovery allows for by formalizing independence of cause and mechanism Peters et al. [2017].

Constraint-based causal discovery methods for time-series data historically began with Granger causality Granger [1969], and since has been addressed in Entner and Hoyer [2010a], Malinsky and Spirtes [2018], Runge et al. [2019b], Runge [2020], Gerhardus and Runge [2020] to cover nonlinear relationships, contemporaneous as well as lagged links, latent confounders and highly auto-correlated data. For an overview, see Runge et al. [2019a].

3 THEORETICAL FOUNDATIONS

Within the JCI framework [Mooij et al., 2020], the causal relations of a system and its context are represented by a joint (or meta) structural causal model (SCM). A *system variable* is a time-dependent random variable whose distribution can change across datasets. In the following, a *temporal context variable* is a time-dependent random variable [1] that remains the same across datasets. A *spatial context variable* is a random variable that is constant over time and within a dataset but can change across datasets. The information on which variables belong to the system or to the context is given as a domain assumption. We now formulate an assumption on the data-generating mechanism that always holds, unless stated otherwise.

Assumption 0. *("Joint time-dependent SCM") The underlying data-generating mechanism across datasets $d \in \mathcal{D}$ with $|\mathcal{D}| = M$ is a simple time-dependent structural causal model (SCM) involving the time-dependent system variables $\mathbf{X}_t = \{X_t^i\}_{i \in \mathcal{I}}$ at time t as well as context*

[1] The term *random variable* is used for context variables in the sense that it is done in Mooij et al. [2020].

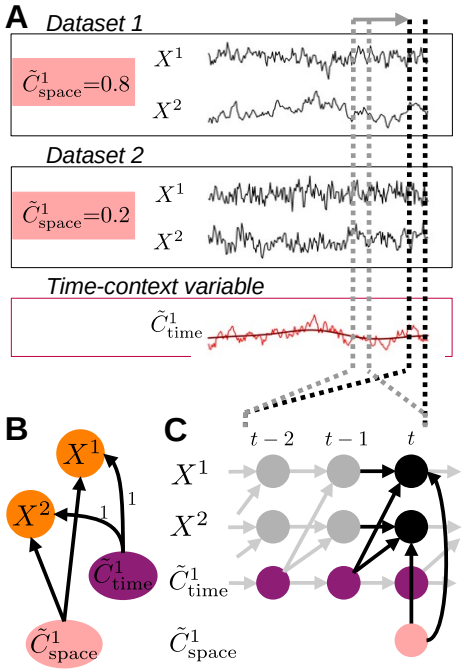

Figure 1: **Causal discovery with temporal- and spatial-contexts. (A)** Two datasets of system variables X^1, X^2 may be confounded by the same temporal context $\tilde{C}^1_{\text{time}}$, but differ in dataset-specific characteristics that are constant over time, here autocorrelation, which can be represented by a spatial context $\tilde{C}^1_{\text{space}}$. J(oint)-PCMCI$^+$ learns the causal relations which can be represented in **(B)** a summary causal graph (link labels denote time lags) or **(C)** a time series graph. Samples of the nodes are pooled from both datasets over the (user-defined) stationary part of the time series (grey dotted lines) leading to a repeating structure of the time series graph (grey links). Context nodes can also help orienting links since they can create colliders. If contexts are unobserved, J-PCMCI$^+$ can utilize temporal- or spatial dummy variables to de-confound system variable relationships.

variables $\tilde{\mathbf{C}} = \tilde{\mathbf{C}}_{time} \dot\cup \tilde{\mathbf{C}}_{space}$ *with temporal contexts* $\tilde{\mathbf{C}}_{time,t} = \{\tilde{C}^k_t\}_{k\in\mathcal{K}_{\text{time}}}$ *and (time-independent) spatial-contexts* $\tilde{\mathbf{C}}_{space} = \{\tilde{C}^l\}_{l\in\mathcal{K}_{\text{space}}}$*, for* $i \in \mathcal{I}$*:*

$$
\begin{aligned}
X^{i,d}_t &:= f^i(\text{Pa}_X(X^{i,d}_t), \text{Pa}_{\tilde{C}_{\text{time}}}(X^{i,d}_t), \text{Pa}_{\tilde{C}_{\text{space}}}(X^{i,d}_t), \eta^{i,d}_t),\\
\tilde{C}^k_{\text{time},t} &:= g^k(\text{Pa}_{\tilde{C}_{\text{time}}}(\tilde{C}^k_{\text{time},t}), \eta^k_{\text{time},t}), \quad k \in \mathcal{K}_{\text{time}}\\
\tilde{C}^{l,d}_{\text{space}} &:= h^l(\text{Pa}_{\tilde{C}_{\text{space}}}(\tilde{C}^l_{\text{space}}), \eta^{l,d}_{\text{space}}), \quad l \in \mathcal{K}_{\text{space}}
\end{aligned}
\tag{1}
$$

where the exogenous noise variables $(\eta^{i,d}_t, \eta^k_{\text{time},t}, \eta^{l,d}_{\text{space}})$ *are jointly independent and* $\eta^{i,d}_t$ *are identically distributed across time and space,* $\eta^k_{\text{time},t}$ *are identically distributed across time, and* $\eta^{l,d}_{\text{space}}$ *are identically distributed across space.* Pa_X *denotes the causal parents within* $\mathbf{X}$*, and analogously for* $\text{Pa}_{\tilde{C}_{\text{time}}}$ *and* $\text{Pa}_{\tilde{C}_{\text{space}}}$*.*

Here an SCM is called *simple* if it is uniquely solvable with respect to any subset of variables in the SCM [Bongers et al., 2021]. In our setup, this implies that the time-dependent system is stable. As discussed above, we also allow for latent context variables. Non-stationary data could be modeled by intervening on a temporal context node simultaneously across all time points.

Assumption 1. *("Markov, Faithfulness and partial causal sufficiency") Let* M *be a SCM of the form (1) with graph* $\mathcal{G}$*, the joint distribution* $P_M(X, \tilde{C})$ *induced by the SCM satisfies the Markov Property with respect to the graph* $\mathcal{G}$*. Additionally,* $P_M(X, \tilde{C})$ *is faithful to the graph* $\mathcal{G}$ *of* M*. The collection of datasets from the joint SCM* (1) *is assumed to contain data from all system nodes* $\mathbf{X}_t$*, but may contain unobserved context nodes* $\mathbf{L}$ *and observed context nodes* $\mathbf{C}$*,*

$$\tilde{\mathbf{C}} = \mathbf{L} \dot\cup \mathbf{C}\,. \tag{2}$$

Subsequently, $\mathbf{L}$ *and* $\mathbf{C}$ *can also be written as a disjoint union over their spatial and temporal components.*

Note, that in the setup of (1) we have included the assumption that context variables are exogenous to the system, see *JCI assumption 1* in Mooij et al. [2020]. Furthermore, we assume that no latent context confounds an observed context and a system variable. Since a latent context is the only possible latent confounder in our setup, this assumption is equivalent to *JCI assumption 2* Mooij et al. [2020]. We make an additional assumption in view of efficiency:

Assumption 2. *("Context-system links") JCI assumption 1 and 2 hold. Further, a causal link between an observed context and a system variable is not mediated by latent context variables.*

Finally, we adapt the *pseudo-causal sufficiency assumption* of Huang et al. [2020] for our case as follows:

Assumption 3. *("Context-determinism") The spatial context variables are deterministic functions of the dataset index. The temporal context variables are deterministic functions of the time-index. We assume these functions to be non-invertible.*

The SCM (1) yields a joint time series causal graph $\mathcal{G}$ (see Fig. 1) over all datasets d. The joint graph $\mathcal{G}$ is related to the target of our discovery task:

Definition 1. *("Target graph") The target graph of J-PCMCI+ is the induced subgraph of* $\mathcal{G}$ *over the system nodes together with the observed context nodes and their edges to the system nodes.*

4 METHOD

The general idea of our method is to include context nodes in the time series graph motivated by the JCI approach. In

order to deal with latent context variables, we introduce dummy variables (Sect. 4.1) before presenting J-PCMCI$^+$ (Sect. 4.2) and state consistency results in Sect. 4.3.

4.1 DUMMY VARIABLES AS PROXIES FOR LATENT CONFOUNDERS

Let $\mathcal{G} = (V, E)$ be the time series graph corresponding to SCM (1), where V denote the vertices and E edges between vertices. Here, $V = \mathbf{X} \cup \mathbf{C} \cup \mathbf{L}$, where $\mathbf{X}$ (resp. $\mathbf{C}$ and $\mathbf{L}$) refers to vertices at all time points. The set of edges E can be written as a disjoint union between edges E_L, where at least one of the corresponding nodes is in $\mathbf{L}$, and its complement E_O, which consists of edges that only connect observed (system and/or context) nodes. That is, $E = E_L \dot{\cup} E_O$. Further, E_O is itself a disjoint union of E_S, which are the edges where at least one of the two nodes is in $\mathbf{X}$ and its complement E_C, i.e., $E_O = E_S \dot{\cup} E_C$.

Definition 2. *(Space dummy variable) The space dummy variable D_{space}, henceforth referred to as space dummy, is a variable that labels datasets.*

Without prior expert knowledge, this labelling is arbitrary. For instance, the simplest embedding is $D_{\text{space}} \in \{1, \dots, M\}$. Alternatively, in a one-hot-encoded embedding, $i \in \{1, \dots, M\}$ denotes the position of the 1 in an M-dimensional vector where all other entries are 0. In the following, we work with a one-hot-encoded space dummy. However, we note that the question of which embedding to choose for the space dummy is far from settled, and requires further expert knowledge about the particular setup and the type of conditional independence to be used in the causal discovery algorithm. Also note, there is no one-to-one relation between datasets and spatial contexts, i.e., two datasets can have same value for a spatial context.

Definition 3. *(Time dummy variable) The time dummy variable D_{time}, henceforth referred to as time dummy, is a variable that labels each time-step in the time-series data.*

Here too, we arbitrarily choose the embedding for D_{time} to be a one-hot-encoding into a T-dimensional vector, where T is the length of the time-series.

Definition 4. *(Dummy projection) We define the dummy-projection of the graph $\mathcal{G} = (\mathbf{X} \cup \mathbf{C} \cup \mathbf{L}, E_L \cup E_O)$ to be the graph $\mathcal{G}_D = (\mathbf{X} \cup \mathbf{C} \cup \{D_{\text{space}} \cup D_{\text{time}}\}, \tilde{E})$, where edges $\tilde{E}$ are defined as:*

$$\tilde{E} = \{(D_{\text{space}}, v) | (u, v) \in E,\ \forall\, u \in \mathbf{L}_{space} \text{ and } v \in \mathbf{X}\}$$
$$\cup \{(D_{\text{time}}, v) | (u, v) \in E,\ \forall\, u \in \mathbf{L}_{time} \text{ and } v \in \mathbf{X}_t\} \cup E_S.$$

We further specify that D_{space} and D_{time} nodes only have contemporaneous links to system variables, see figure 1 in SM for details. Further, note that in the dummy projection, we have omitted the edges E_C, i.e., the edges between the observed context variables, since these relationships are not of interest for the target graph (definition 1).

Finally, we introduce the dummy-deleted graph. For a visualization of the dummy projection and deletion operations, see figure 2 in SM.

Definition 5. *(Dummy deletion) Let $\mathcal{G}_D$ be the dummy projection of the graph $\mathcal{G}$. The dummy-deleted graph $\mathcal{G}_{D_{del}}$ is the graph where the dummy variables and any outgoing edge therefrom is removed.*

Under assumption 3, for $C_s \in \mathbf{C}_{space}$ and $L_s \in \mathbf{L}_{space}$, we can always find (not necessarily unique) functions g_C and g_L with $C_s = g_C(D_{\text{space}})$ and $L_s = g_L(D_{\text{space}})$, and analogously for the temporal counterparts.

The introduction of the dummy variables into the causal discovery task is for the purpose of removing the confounding effect of latent context variables on a pair of system variables. Latent system variables and their confounding cannot be handled with this method. As we will see in Section 4.2, our method yields a graph between $\mathbf{X}$, $\mathbf{C}$ and D, which is not exactly the dummy projection of the true graph $\mathcal{G}$, but whose dummy deletion is the target graph (definition 1).

Interpretation of the dummy-projection Some caution has to be applied in the interpretation of links between the dummy and system nodes in the dummy-projected graph. These links are only placeholders for the links between unobserved context nodes and system nodes in the ground truth graph. The dummy is not a causal variable itself! Further, note that in definition 4, we did not include links between the dummy and observed context nodes in the dummy projection since these links are deterministic by assumption 3 and always present and thus not informative.

Including both observed context and dummy variables A natural question at this point might be, why include the observed context variables at all in the causal discovery task, even though the dummy can remove all influence of context variables because of the general way in which is it defined (definitions 2, 3). Theoretically, there would indeed be nothing wrong with excluding contexts altogether. However, the dummy variable is not interpretable, and not useful when the goal is to learn the causal relationship between specific context and system variables. Further, as we will see below, in the first step our casual discovery algorithm learns the context-system adjacencies, the relationship between which may be mathematically simpler than those between the highly general dummy and system variables. In the second step, it learns the dummy-system adjacencies given the context parents learnt in the first step. This helps infer influence on the system variable of the dataset label that cannot be explained by the context variables, i.e., in essence we learn hidden contexts. Finally, as also pointed out by Mooij

et al. [2020], the separate observed context nodes help in orienting adjacencies between system variables. Refer also figure 1 for a visualization.

A naive implementation to learn the causal relationships between context-system and dummy-system, where the context and dummy variables are treated on the same footing when testing adjacencies to the system, would not be correct as the *causal faithfulness assumption* would be violated. This is because the relationship between context and dummy variables is deterministic. The two-step procedure outlined above circumvents erroneous inferences of adjacencies due to faithfulness violation. For details, see sect.4.2 and SM.

4.2 ALGORITHM

In the pooled dataset, we include one variable for the space and time dummy each, as well as for spatial context variables, at time t. These nodes can only have contemporaneous links to system nodes since they either do not change over time or contain no information about the temporal structure. That is, only temporal context variable can have a lagged influence on the system variables, see also figure 1 and SM (figure 1). To be able to deal with observed contexts and dummy variables, that are essentially placeholders for the unobserved context variables, our method first discovers links between system and observed context nodes while ignoring the dummy nodes, and in the next step discovers links between dummy and system nodes. Finally, using the information on the contextual parents of each system node, we do causal discovery on the system node pairs. In the following, we detail this procedure for the non-time series and time-series case.

Non-time series case: To ease the explanations for the time-series case, where both spatial and temporal context variables can occur, we first focus on the non-time-series case, where only spatial context variables can occur. Consequently, we only have to consider the space dummy. Note that assumption 0 can be simplified to the non-time series straightforwardly. We will combine the well-known PC algorithm Spirtes et al. [2000] with both observed and dummy context variables. Pseudocode for this method is provided in Algorithm 1. In the standard setting, the PC algorithm is a constraint-based causal discovery algorithm for the causal sufficient case that relies on the Markov and Faithfulness assumption. In its first stage (skeleton discovery), adjacencies are learned based on iteratively testing conditionally independence of pairs of variables at some significance level α. Afterwards, the links are oriented based on a set of rules. We will focus on how its skeleton phase needs to be adapted.

1. In the first step, we discover context-system links. We iteratively test independence between the following node pairs (X^i, C^j), and (C^j, X^i) for all i, j while conditioning on subsets of the union over system and observed context nodes $\mathbf{X} \cup \mathbf{C}$. In other words, we initialize a fully connected graph between the system and context variables, eliminate the edges between context variables, and run the skeleton phase. By ignoring the dummy node in this step, we are able to circumvent the faithfulness violation that stems from the fact that every observed context node is a deterministic function of the dummy. Then, we combine the found adjacencies with the exogeneity of the context to the system to infer the observed contextual parents $\text{Pa}_C(X^i)$ of each system variable X^i.

2. In the second step, we focus on the discovery of dummy-system links. In particular, we test independence between D and each $X^i \in \mathbf{X}$ conditional on subsets of $\mathbf{X}$ and the found contextual parents $\text{Pa}_C(X^i)$. Combined with the expert knowledge that the dummy cannot be a descendent of a system variable, this gives us the dummy parents of X^i. We denote the set of dummy and contextual parents of X^i by $\text{Pa}_{CD}(X^i)$.

3. Finally, we run the skeleton phase of the full PC algorithm on $\mathbf{X} \cup \mathbf{C} \cup \{D\}$ while incorporating the background knowledge of the links from $\text{Pa}_{CD}(X^i)$ to X^i, and no context-context and context-dummy links.

Since context-system and dummy-system links are oriented by assumption, the orientation phase, see Meek [1995] for rules, needs to be applied to orient the system variables only. Note, however that we are taking between context-system or dummy-system edges into account whenever they form an unshielded triple with two system variables, i.e. $C \rightarrow X^i \circ\!\!-\!\!\circ X^j$ or $D \rightarrow X^i \circ\!\!-\!\!\circ X^j$. This allows to orient more edges than only considering triples of system variables.

Time series case: Next, we combine the PCMCI$^+$ algorithm Runge [2020] with observed context and dummy variables. To recall briefly, PCMCI$^+$ is a causal discovery algorithm for time series data, that allows for both contemporaneous and lagged links and assumes causal sufficiency. It consists of two steps. The first *PC$_1$ lagged phase* infers a superset of the lagged parents together with the parents of contemporaneous ancestors. Next, the *MCI contemporaneous phase* starts with links found in the previous step and all possible contemporaneous links, it then conducts momentary conditional independence (MCI) with a modified conditioning set learned in the previous step to increase detection power.

Our method consists of four main steps: one PC$_1$ lagged phase and three MCI phases. In the first step, supersets of the lagged parents of the system and observed temporal context nodes are discovered by running the PC$_1$ lagged phase on this subset of nodes. Next, the MCI test is run on pairs of system and context nodes conditional on subsets of system and context, i.e. perform MCI tests for pairs

Algorithm 1: J-PC (for non-time series), pseudocode for poolData, and partialSkeletonPC is provided in SM

Data: Background knowledge on context-system link orientation $\mathcal{E}$, observational data $(\mathbf{X}^{(m)})_{m=1,\dots,M}$ in M dataset, observed context variables $(\mathbf{C}^{(m)})_{m=1,\dots,M}$ for each dataset, dummy variable D with distinct values for each dataset, significance level α

Result: graph $\mathcal{G}$

$(\mathbf{X}, \mathbf{C}, D) \leftarrow \text{poolData}((\mathbf{X}^{(m)}, \mathbf{C}^{(m)})_{m=1,\dots,M}, D)$

Set $\mathcal{P}_C := \{(X, C), (C, X) | X \in \mathbf{X}, C \in \mathbf{C}\}$,

$\mathcal{P}_D := \{(X, D) | X \in \mathbf{X}\}, \mathcal{P}_S := \{(X, Y) | X, Y \in \mathbf{X}\}$

Set $data_C := (\mathbf{X}, \mathbf{C})$ and $data_D := (\mathbf{X}, \mathbf{C}, D)$

Set $\mathcal{C} = \emptyset$

for *index in [C, D]* **do**

  $\mathcal{G} \leftarrow \text{partialSkeletonPC}(data_{\text{index}}, \alpha, \mathcal{P}_{\text{index}}, \mathcal{C})$

  for X *in* $\mathbf{X}$ **do**

    Orient context-system edge as in $\mathcal{E}$

    Add contextual parents $\text{Pa}_{\text{index}}(X)$ as in $\mathcal{G}$ to $\mathcal{C}$

  end

end

$\mathcal{G} \leftarrow \text{partialSkeletonPC}(data_D, \alpha, \mathcal{P}_S, \mathcal{C})$

Orient system edges using PC-orientation rules

return $\mathcal{G}$

$((C^j_{t-\tau}, X^i_t))_{\tau>0}$, (C^j_t, X^i_t), (X^i_t, C^j_t) for all i, j,

$$C^i_{t-\tau} \perp\!\!\!\perp X^j_t | \mathbf{S}, \hat{\mathcal{B}}^-_t(X^j_t) \setminus \{C^i_{t-\tau}\}, \hat{\mathcal{B}}^-_{t-\tau}(C^i_{t-\tau})$$

with $\mathbf{S}$ being a subset of the contemporaneous adjacencies $\mathcal{A}_t(X^j_t)$ and $\hat{\mathcal{B}}^-_t(X^j_t)$ are the lagged adjacencies from step one. If C is a spatial context variable, we only have to test the contemporaneous pairs (C^j_t, X^i_t), (X^i_t, C^j_t) for all i, j. In the third step, MCI tests on all system-dummy pairs conditional on the superset of lagged links, the discovered contemporaneous context adjacencies, as well as on subsets of contemporaneous system links, are performed, i.e. test for (D, X^i_t), (X^i_t, D) for all i, i.e.

$$D \perp\!\!\!\perp X^j_t | \mathbf{S}, \hat{\mathcal{B}}^C_t(X^j_t)$$

where $\mathbf{S} \subset \mathcal{A}_t(X^i_t)$ and $\hat{\mathcal{B}}^C_t(X^j_t)$ are the lagged and contextual adjacencies found in the previous step. Using assumption 2, context node is the parent in all system-context links. Finally, in the fourth step, we perform MCI tests on all system pairs conditional on discovered lagged, context and dummy adjacencies, as well as on subsets of contemporaneous system links and orientation phase. In more detail, we perform MCI test for pairs $((X^j_{t-\tau}, X^i_t))_{\tau>0}$, (X^i_t, X^j_t) for all i, j, i.e.

$$X^i_{t-\tau} \perp\!\!\!\perp X^j_t | \mathbf{S}, \hat{\mathcal{B}}^{CD}_t(X^j_t) \setminus \{X^i_{t-\tau}\}, \hat{\mathcal{B}}^{CD}_t(X^i_{t-\tau})$$

where $\mathbf{S} \subset \mathcal{A}_t(X^i_t)$ and $\hat{\mathcal{B}}^{CD}_t(X^j_t)$ are the lagged, contextual, and dummy adjacencies found in the previous steps. Finally, all remaining edges (without expert knowledge) are oriented using the PCMCI$^+$ orientation phase while making use of all triples involving one context or dummy variable and two system variables as in the non-time series case.

4.3 THEORETICAL RESULTS

Proofs for the following statements are provided in SM.

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

Note that here the dummy-deleted ground truth graph is the target graph (definition 1) adapted to the non-time series case.

Theorem 2 (Time series consistency result)**.** *Denote the time series graph output of J-PCMCI$^+$ (Algorithm 2) as $\mathcal{G}_{alg}$. Under assumptions 1, 2, 3, and assuming consistent conditional independence tests are used, the dummy deletion of $\mathcal{G}_{alg}$ corresponds to the target graph (definition 1).*

5 NUMERICAL EXPERIMENTS

Data simulation We generate toy data from the SCM 1 where we assume the functions f_i, g_k, and h_l to be linear. In particular, for system variables X and temporal context variables $(C^{\text{time},k})_{k\in\mathcal{K}\text{time}}$ and spatial context variables $(C^{\text{space},l})_{l\in\mathcal{K}\text{space}}$, we consider the following ground truth SCM $X^{i,m}_t = a_i X^{i,m}_{t-1} + \sum_j b_j X^{j,m}_{t-\tau_j} + \sum_j c_j C^{\text{time},j}_{t-\tau_j} + \sum_j d_j C^{j,m}_{\text{space}} + \eta^{i,m}_t$, where $i \in \mathcal{I}$, $t = 1, \dots, T$, and $m = 1, \dots, M$, $C^{\text{time},k} \sim \mathcal{N}(0, 1)$, $C^{\text{space}} \sim \mathcal{N}(0, 1)$. Furthermore, $\eta^i \sim \mathcal{N}(0, 1)$ i.i.d., a_i autocorrelation parameter uniformly drawn from $[0.3, 0.8]$, coefficients b_j, c_j, d_j are uniformly drawn from $[0.5, 0.9]$, 50% of the links are contemporaneous, the remaining lags are drawn uniformly from $[1, 3]$. After the data has been generated, its variance is rescaled to one across all datasets to avoid varsortability Reisach et al. [2021]. In the numerical experiments, we make the restriction that one system node can have at most one contextual parent.After the time series for the ground truth model has been generated a certain fraction (indicated by parameter frac_observed) of the context nodes is selected to be observed, the others are unobserved. Note that in our simulated data all context nodes are exogeneous (they do not even have other context nodes as parents). We decided to set the experiments up in this way to put the focus on the discovery of the system-context links and also the deconfounding property of the dummy nodes.

Setup We evaluate the performance of our method using True (TPR) and False Positive Rate (FPR) for the adjacencies which is calculated only on the system-system

Algorithm 2: J-PCMCI$^+$ (for time series), laggedSkeletonPCMCI$^+$ refers to Algorithm 1 in Runge [2020], partialContempSkeletonPCMCI$^+$ is a small adaption of Algorithm 2 of Runge [2020] which is further described in the SM, colliderPhase and rulePhase refer to Algorithms 3 and 4 in Runge [2020]

Data: Background knowledge on context-system link orientation $\mathcal{E}$, M observational system time series datasets $\mathbf{X} = (\mathbf{X}^i)_{i \in \mathcal{I}}$, observed temporal context variables $\mathbf{C}_{\text{time}}$, observed spatial context variables $\mathbf{C}_{\text{space}}$, temporal and spatial dummy variables $D_{\text{time}}, D_{\text{space}}$, significance level α, maximal time lag $\tau_{\max}$, $CI(X, Y, \mathbf{Z})$

Result: graph $\mathcal{G}$

$\{\hat{\mathcal{B}}_t^-(X) | X \in \mathbf{X} \cup \mathbf{C}_{\text{time}}\} \leftarrow$ laggedSkeletonPCMCI$^+(\mathbf{X} \cup \mathbf{C}_{\text{time}})$

Initialize $\mathcal{C}(X) \leftarrow \emptyset$ for all $X \in \mathbf{X}$

Set $\mathcal{P}_C := \{(C^k_{t-\tau}, X^i_t), (X^i_t, C^k_t) | \tau \geq 0, \forall i, k\}$, $\mathcal{P}_D := \{(X^i_t, D), (D, X^i_t) |$for $D \in \{D_{\text{time}}, D_{\text{space}}\}, \forall i\}$,
$\mathcal{P}_S = \{((X^j_{t-\tau}, X^i_t))_{\tau > 0}, (X^i_t, X^j_t) | i, j\}$

Set $data_C := \mathbf{X} \cup \mathbf{C}_{\text{time}} \cup \mathbf{C}_{\text{space}}$, and $data_D := \mathbf{X} \cup \mathbf{C}_{\text{time}} \cup \mathbf{C}_{\text{space}} \cup \{D_{\text{time}}, D_{\text{space}}\}$

for *index in [C, D]* **do**

  $\mathcal{G} \leftarrow$ partialContempSkeletonPCMCI$^+(data_{\text{index}}, |\mathcal{I}|, \tau_{\max}, \alpha, \hat{\mathcal{B}}_t^-(X), \mathcal{C}(X), \mathcal{P}_{\text{index}})$

  for X *in* $\mathbf{X}$ **do**

    orient context-system edge as in $\mathcal{E}$

    add all context nodes that are adjacent to X in $\mathcal{G}$ to $\mathcal{C}(X)$

  end

end

$\mathcal{G} \leftarrow$ partialContempSkeletonPCMCI$^+(data_D, |\mathcal{I}|, \tau_{\max}, \alpha, \hat{\mathcal{B}}_t^-(X), \mathcal{C}\ \mathcal{P}_S)$

$\mathcal{G}$, sepset, ambigious triples, conflicting links $\leftarrow$ colliderPhase($data_D$), i.e., on all unshielded triples
$X^i_{t-\tau} \rightarrow X^k_t \circ\!\!-\!\!\circ X^j_t$ $(\tau > 0)$ or $X^i_t \circ\!\!-\!\!\circ X^k_t \circ\!\!-\!\!\circ X^j_t$ or $K \rightarrow X^i_t \circ\!\!-\!\!\circ X^j_t$ with $K \in \mathbf{C}_{\text{time}} \cup \mathbf{C}_{\text{space}} \cup \{D_{\text{time}}, D_{\text{space}}\}$

$\mathcal{G}$, conflicting links $\leftarrow$ rulePhase($\mathcal{G}$, ambigious triples, conflicting links)

return $\mathcal{G}$

context links. Separately, we report the TPR and FPR on system-observed context links. All metrics and their standard deviations are computed on the estimated graphs of 50 realizations of the model from time series with length T.

We compare our method to PCMCI$^+$ run on the data of the system nodes only by simply concatenating the data, as well as to PCMCI$^+$ where we have only included the observed (context and system) nodes. In this variant, we took care to include spatial context nodes only once in the time series graph. We build upon the implementation of PCMCI$^+$ algorithm within the Tigramite software package [Runge et al., 2019b] published under the GNU General Public License.

We used the following model parameters in our experiments: Number of system nodes $|\mathcal{I}| = 5$, number of context nodes $|\mathcal{K}_{\text{time}}| + |\mathcal{K}_{\text{space}}| = 3$, maximal time lag $\tau_{\max} = 2$, significance level $\alpha = 0.05$. We use an extension of the (component-wise) partial correlation conditional independence test. We vary the value of the time sample size T, and number of datasets M. Results for other fractions of observed context nodes can be found in the SM.

Benefit of including expressive context nodes and dummy In figure 2, we report the TPR and FPR for observed context-system links, respectively. We see that our method finds the links between all observed context nodes and the system nodes as well as the PCMCI$^+$ variant with observed context nodes included in the graph. In figure 3, we observe that the performance of our method is comparable to only including dummy variables when evaluated only on the system-system links. These two observations illustrate the benefit of using our method: If we would only rely on the dummy context variables, we would achieve deconfounding of the system variables but, naturally, we would not find any of the links between the observed context nodes and the system. In other words, we would not be able to learn which parts of the system are dependent on the properties of the context. On the other hand, if we would only include the observed context nodes, we would not be able to remove the confounding that could be introduced by the latent context variables. The effects of the counfounding latent context variables, are visible in the rise in FPR on system-system links in figure 3. We also see lower edgemark recall and precision when only using system data, see the figures in the SM.

Convergence analysis (time and space dimension) We want to numerically study the finite sample properties of our method. For that, we look at the TPR and FPR on the system links for varying time sample size T while keeping the number of spatial contexts M fixed, and the other way around, see figure 3. In these experiments half of the context nodes are unobserved. In the SM, we have also included 3D-plots for TPR and FPR on all pairs of T and M.

It is a well established result in the econometrics literature, that, when considering fixed effects models, a bias is introduced in the OLS estimator of the slope parameter Nickell

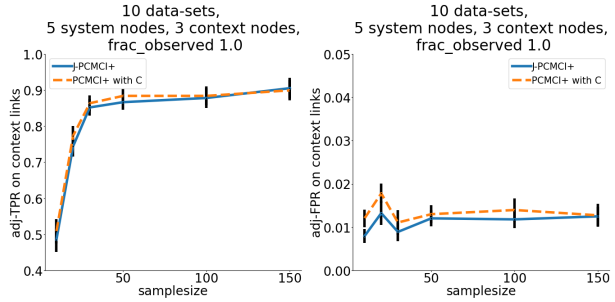

Figure 2: Discovery results of context-system links for varying sample sizes T, $M = 10$. All other setup parameters are set as the defaults described in the main text. In this setting all context nodes are observed. Here, we compare our method (J-PCMCI$^+$) to PCMCI$^+$ using all data of observed nodes. Note that the maximal TPR that can be reached is equal to frac_observed.

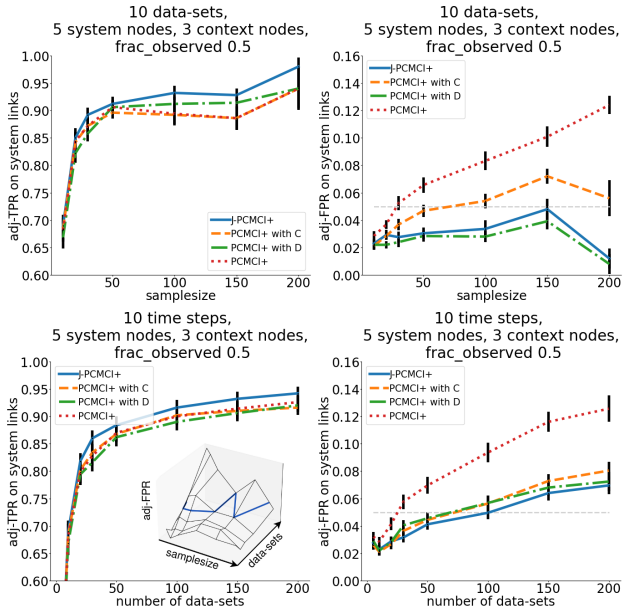

Figure 3: Discovery results of system-system links for varying sample sizes T, and fixed $M = 10$ (top row), and varying number of contexts M, and fixed $T = 10$ (bottom row). All other setup parameters are set as the defaults described in the main text. In this setting half of the context nodes are 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$^+$). The inset shows the adj-FPR-surface with the contour of the α-level in a simplified experimental setup to visualize the convergence of the method. Refer to the SM for details.

[1981]. A similar inconsistency problem can be observed in our method whenever T is kept fixed and is small compared to M: Even when the number of spatial contexts M goes to infinity, the links are not discovered correctly, see figure 3 (in particular the FPR plots) and the SM.

6 DISCUSSION AND CONCLUSIONS

We presented an algorithm (J-PCMCI$^+$) for causal discovery from a collection of multivariate time series datasets that is able to deal with observed and unobserved context variables underlying the datasets. We established its asymptotic consistency and studied its convergence properties numerically.

The **main strengths** of J-PCMCI$^+$ are that it combines the efficient algorithm from Runge [2020] (handling highly autocorrelated time series) with ideas in Mooij et al. [2020], Huang et al. [2020] to model observed as well as unobserved contexts. Pooling data from multiple datasets and adding observed contexts as well as dummies has several important benefits: (1) pooling increases sample size, (2) adding observed contexts and dummies allows to learn context-system relations and can help to orient system-system links, (3) dummy variables allow to remove confounding at least from latent context variables. J-PCMCI$^+$ inherits the benefits of PCMCI$^+$ for high-recall and accounting for autocorrelation in the conditional independence tests. We find numerically that J-PCMCI$^+$ has good performance for sufficiently large sample sizes and moderate numbers of datasets.

The **main weaknesses** are that some assumptions might be strong and unrealistic. In particular, JCI assumption 2 and the assumption that prohibits latent context mediation of the observed context to system link can be hard to justify depending on the setup, see Assumption 2. In the SM, we discuss ramifications of partially relaxing this assumption. Our numerical experiments indicate that for too small sample sizes, we get inflated false positives due to missing latent confounders. We also cannot overcome the fundamental finite-sample bias in the OLS estimator [Nickell, 1981] for small sample sizes and dummy variables carry the disavantage of increasing dimensionality.

In **future work**, we plan to extend the method to weaken the partially strong assumptions and to account for latent system variable confounding Gerhardus and Runge [2020]. Furthermore, while our method only uses high-dimensional dummy variables where it finds dependence, we haven't looked yet into adapting CI tests specifically for dummy variables. In this work, we also only considered acyclic SCMs. However, an extension to include the cyclic case is possible based on the consistency of the PC algorithm in the cyclic setting Mooij and Claassen [2020]. Moreover, an in-depth analysis of finite sample properties of the presented method is needed.

References

Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2009.

Elias Bareinboim and Judea Pearl. A general algorithm for deciding transportability of experimental results. *Journal of causal Inference*, 1(1):107–134, 2013.

Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

Stephan Bongers, Patrick Forré, Jonas Peters, and Joris M. Mooij. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49 (5):2885 – 2915, 2021. doi: 10.1214/21-AOS2064. URL `https://doi.org/10.1214/21-AOS2064`.

David Maxwell Chickering. Learning equivalence classes of bayesian-network structures. *The Journal of Machine Learning Research*, 2:445–498, 2002.

A Philip Dawid. Influence diagrams for causal modelling and inference. *International Statistical Review*, 70(2): 161–189, 2002.

Doris Entner and Patrik O. Hoyer. On causal discovery from time series data using fci. In P. Myllymäki, T. Roos, and T. Jaakkola, editors, *Proceedings of the 5th European Workshop on Probabilistic Graphical Models*, pages 121–128, Helsinki, FI, 2010a. Helsinki Institute for Information Technology HIIT.

Doris Entner and Patrik O Hoyer. On causal discovery from time series data using fci. *Probabilistic graphical models*, pages 121–128, 2010b.

Andreas Gerhardus and Jakob Runge. High-recall causal discovery for autocorrelated time series with latent confounders. *Advances in Neural Information Processing Systems*, 33:12615–12625, 2020.

Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438, 1969.

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.

Daniel Malinsky and Peter Spirtes. Causal structure learning from multivariate time series in settings with unmeasured confounding. In T D Le, K Zhang, E Kıcıman, A Hyvärinen, and Lin Liu, editors, *Proceedings of 2018 ACM SIGKDD Workshop on Causal Disocvery*, volume 92 of *Proceedings of Machine Learning Research*, pages 23–47, London, UK, 20 Aug 2018. PMLR. URL `http://proceedings.mlr.press/v92/malinsky18a.html`.

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

Joris M Mooij and Tom Claassen. Constraint-based causal discovery using partial ancestral graphs in the presence of cycles. In *Conference on Uncertainty in Artificial Intelligence*, pages 1159–1168. PMLR, 2020.

Joris M Mooij, Sara Magliacane, and Tom Claassen. Joint causal inference from multiple contexts. *The Journal of Machine Learning Research*, 21(1):3919–4026, 2020.

Stephen Nickell. Biases in dynamic models with fixed effects. *Econometrica: Journal of the econometric society*, pages 1417–1426, 1981.

Judea Pearl. [bayesian analysis in expert systems]: Comment: Graphical models, causality and intervention. *Statistical Science*, 8(3):266–269, 1993. ISSN 08834237. URL `http://www.jstor.org/stable/2245965`.

Judea Pearl and Elias Bareinboim. Transportability of causal and statistical relations: A formal approach. In *Twenty-fifth AAAI conference on artificial intelligence*, 2011.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 947–1012, 2016.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Alexander Reisach, Christof Seiler, and Sebastian Weichwald. Beware of the simulated dag! causal discovery benchmarks may be easy to game. *Advances in Neural Information Processing Systems*, 34:27772–27784, 2021.

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, Sebastian Bathiany, Erik Bollt, Gustau Camps-Valls, Dim Coumou, Ethan Deyle, Clark Glymour, Marlene Kretschmer, Miguel D Mahecha, Jordi Muñoz-Marí, et al. Inferring causation from time series in earth system sciences. *Nature communications*, 10(1):1–13, 2019a.

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

Peter Spirtes and Kun Zhang. Causal discovery and inference: concepts and recent methodological advances. In *Applied informatics*, volume 3, pages 1–28. SpringerOpen, 2016.

Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

Thorsten Wagener, Murugesu Sivapalan, Peter Troch, and Ross Woods. Catchment classification and hydrologic similarity. *Geography compass*, 1(4):901–931, 2007.