Discovering cause-effect relationships in spatial systems with a known direction based on observational data

Konrad Mielke K.MIELKE@SCIENCE.RU.NL
Tom Claassen T.CLAASSEN@SCIENCE.RU.NL
Mark A.J. Huijbregts M.HUIJBREGTS@SCIENCE.RU.NL
Aafke Schipper A.SCHIPPER@SCIENCE.RU.NL
Tom Heskes T.HESKES@SCIENCE.RU.NL
Faculty of Science, Radboud University, Postbus 9010, 6500 GL Nijmegen, The Netherlands

Abstract

Many real-world studies and experiments are characterized by an underlying spatial structure that induces dependencies between observations. Most existing causal discovery methods, however, rely on the IID assumption, meaning that they are ill-equipped to handle, let alone exploit this additional information. In this work, we take a typical example from the field of ecology with an underlying directional flow structure in which samples are collected from rivers and show how to adapt the well-known Fast Causal Inference (FCI) algorithm (Spirtes et al., 2000) to learn cause-effect relationships in such a system efficiently. We first evaluated our adaptation in a simulation study against the original FCI algorithm and found significantly increased performance regardless of the sample size. In a subsequent application to real-world river data from the US state of Ohio, we identified important likely causes of biodiversity measured in the form of the Index of Biotic Integrity (IBI) metric.

Keywords: Causal Discovery; Fast Causal Inference; Spatial Data

1. INTRODUCTION

Discovering cause-effect relationships from finite data with latent confounders is a very challenging task in the field of graphical models and structure discovery. In principle, many features of causal structures can be determined, even from observational data (Pearl (2014)). Constraint-based causal discovery algorithms combine independence constraints from statistical tests to determine common features of the true underlying causal graphs. In real-world applications, however, the assumptions of constraint-based causal discovery algorithms are often violated.

One assumption shared by many causal discovery algorithms, including the PC algorithm and the fast causal inference (FCI) algorithm (Spirtes et al. (2000)), is that observations are independent and identically distributed (IID). When observations are gathered at different locations in space (from here on called spatial data), the IID assumption is typically violated because the observations at locations that are close to each other are likely to be causally related. Irvine and Gitelman (2011) and Ebert-Uphoff and Deng (2014) showed how to perform causal discovery from spatial data: For each location, they introduced a separate set of all variables of interest. Irvine and Gitelman (2011) then compared models with different sets of undirected relationships between locations whereas Ebert-Uphoff and Deng (2014) explicitly learned the cause-effect relationships between locations.

These strategies are useful when the relationships between locations are unknown and the goal is to learn these relationships. In many settings, however, the spatial structure of the system is known a priori and the goal is to learn the *local causal structure* (i.e., a causal structure that applies independently of the location) of a set of variables. As an example of such a setting, we refer to a common application from ecology: one main goal in ecology is to infer cause-effect relationships between environmental variables and the integrity of biotic communities. It is an open research question which methodology is best suited to do so and, to this point, most approaches are based on correlations instead of cause-effect links. For lotic communities (i.e., communities that live in rivers), observations are not IID, as observations at locations that are close to each other will be highly correlated. On the other hand, background knowledge dictates that the river current is always pointing downstream and thus cause-effect relationships between locations in the river should do so, too. How to use this background knowledge to obtain the most accurate local causal structure is an open research question. While this application is very specific, the overarching problem is generic and applies to many spatial systems in which the direction of cause-effect relationships is known, including, among others, blood vessels, sewerage, ocean currents, and air currents. We refer to data originating from this type of spatial structure as current data.

Here, we explore a possible link between current data and temporal data. Temporal data is similar to current data in that the direction of possible cause-effect relationships is known and the quantity of interest is a causal structure that is invariant (i.e., that applies at all times or, in our case, locations). Thus, we build on previous work on causal discovery from temporal data. Chu et al. (2005) and Entner and Hoyer (2010), among others, showed how to apply constraint-based causal discovery algorithms to temporal data: For multiple time steps, they introduced a separate set of all variables of interest and enforced the direction of time in the resulting causal structure. Temporal data is strictly linear, however, whereas current data is not due to possible branching of currents.

In this paper, we show how to adapt the FCI algorithm to account for our background knowledge in the application to current data. We assume that the spatial structure of the system (i.e., the order of the observations) is known and desire to learn a local causal structure that is space invariant (i.e., that applies everywhere). Our method relies on only one observation per location. In our algorithm, we reduce the number of independence tests as much as possible. This does not only save computation time but should also improve accuracy: constraint-based causal discovery algorithms perform conditional independence tests based on the results of previous tests (Spirtes et al. (2000), Richardson (2013)). Consequently, early mistakes in independence testing can lead to further mistakes and thus globally inaccurate solutions (Claassen and Heskes (2012); Hyttinen et al. (2014)).

The paper is structured as follows: First, we introduce key concepts of graph theory and causal discovery. We present the data generating process that introduces the assumptions of our algorithm. We then present our adapted FCI algorithm. On simulated data, we compare the performance of the adapted and the standard FCI algorithm to evaluate our main research question: Does taking into account background knowledge of the spatial structure in the proposed way improve our capability to infer causal relationships and if so by how much? Finally, we apply our adaptation to a real-world dataset to discover cause-effect relationships between chemical concentrations, river characteristics, and an index that describes the integrity of a biotic community in rivers of the US state of Ohio.

2. PRELIMINARIES

2.1 Graph theory

A *graph* $\mathcal{G}$ is a pair $(\boldsymbol{V}, \boldsymbol{E})$ where $\boldsymbol{V} = \{1, ..., M\}$ is a finite set of *vertices* and $\boldsymbol{E} \subseteq \boldsymbol{V} \times \boldsymbol{V}$ is a set of ordered pairs called *edges*. Vertices that are connected by an edge are called *adjacent*. A graph that contains only directed edges $i \to j$ and no cycles is called a *directed acyclic graph* (DAG). For directed edges $i \to j$, we say that i is the *parent* of j and for directed paths $i \to ... \to j$, j is the *descendant* of i. Three vertices $i - j - k$ are called (i, j, k) a *triple*. If additionally, i and k are not connected by an edge, then the triple is called *unshielded*. In a triple $i \to j \leftarrow k$, the middle vertex j is called a *collider*. A graph in which all vertices are connected is called *complete*. The graph that results from a DAG by replacing all directed edges with undirected edges is called its *skeleton*.

For a DAG $\mathcal{G}$ to describe a data-generating process, we associate each vertex i with a random variable X_i and a conditional distribution $P(X_i|\mathbf{Pa}_{\mathcal{G}}(X_i))$. The joint distribution $P(\boldsymbol{X})$ is *faithful* to $\mathcal{G}$ if the conditional independence relations of the distribution are entailed in $\mathcal{G}$ through *d-separation*: Vertices i and j are d-separated given a subset $\boldsymbol{M} \subseteq \boldsymbol{V} \setminus \{i, j\}$ if for each path between i and j, a collider on the path or a descendant of the collider is not in $\boldsymbol{M}$ or a non-collider on the path is in $\boldsymbol{M}$. Adjacent vertices i and j cannot be d-separated by any set.

In the presence of *latent confounders*, DAGs are not sufficient to describe a causal system because DAGs are not closed under *marginalization* (Richardson and Spirtes (2002)). For this reason, a new class of graphs, called *maximal ancestral graphs* (MAGs), has been introduced. Each DAG with latent variables can be described by a unique MAG. MAGs contain directed and bidirected edges. A bidirected edge $i \leftrightarrow j$ in a MAG can be interpreted as a latent confounder of i and j.

Conditional independence relations are entailed in MAGs through the criterion of *m-separation* which is an extension of d-separation in DAGs. Different MAGs can share the same conditional independence relations and together they form a *Markov equivalence class*. Each Markov equivalence class is represented by a unique *partial ancestral graph* (PAG) that entails the conditional independence relations shared by all MAGs of the class (Zhang (2007)). PAGs contain four types of edges: $\to$, $\leftrightarrow$, $\circ\!\!-\!\!\circ$ and $\circ\!\!\to$, where a circle represents an unknown edge mark.

2.2 Causal discovery without causal sufficiency

The goal in causal discovery is to infer the Markov equivalence class of the true underlying graph from observational data. In this work, we focus on constraint-based causal discovery without the assumption of *causal sufficieny*. Causal sufficiency states that there are no latent confounders in the system which is rarely satisfied in practice.

Here, we briefly describe the fast causal inference (FCI) algorithm (Spirtes et al. (2000)) which consists of three phases. In the first phase (skeleton phase), we infer the skeleton by testing for conditional independence of ordered pairs of vertices (i, j). We start from a complete graph and test for marginal independence ($n = 0$) first, increasing the size n of the conditioning set after all possible combinations of ordered vertices have been tested. In subsequent iterations of n, we test for conditional independence of i and j given all subsets of variables that are adjacent to i excluding j. When we find an independence, the corresponding edge is marked for removal at the end of the current iteration of n, following an adaptation by Colombo and Maathuis (2014) called *stable FCI*. In the second phase (possible d-separation phase), we perform conditional independence tests with additional conditioning sets to find independence relations that we missed in the skeleton phase

due to causal insufficiency. In the third phase (orientation phase), we orient all remaining edges as $\circ\!-\!\circ$ and subsequently apply a set of orientation rules by Zhang (2008). Given faithfulness, the FCI algorithm is sound (i.e., all edges and directions in the inferred PAG are present in the true underlying MAG) and complete (i.e., all edges and directions in the true underlying PAG are present in the inferred MAG; Spirtes et al. (2000)).

2.3 Causal discovery from temporal data

In later sections, we will explore the connection between current data and temporal data. A dynamic, probabilistic graphical model that represents a set of variables and their conditional dependencies via a DAG is called a Dynamic Bayesian Network (Dagum et al. (1992)). These can be inferred with the same algorithms as non-dynamic graphical models. Entner and Hoyer (2010) presented an adaptation for time series data called tsFCI. This adaptation introduces the concept of *homologous* edges: two edges are homologous if they connect the same pair of variables at the same temporal distance. The assumption of time invariance then allows to remove all homologous edges if one independence is found. In the orientation phase, all edges are oriented forward in time. Additionally, when one edge is oriented, all homologous edges are oriented in the same direction.

3. CAUSAL DISCOVERY FROM CURRENT DATA

In this chapter, we present our approach for causal discovery from data with a directional flow. First, we introduce the assumed data generating process. Second, we show how how to manipulate the data to subsequently apply an adapted version of the FCI algorithm.

3.1 Data generating process

We assume that the data is generated by the following process: Let $\boldsymbol{X}(k) = (X_1(k), ..., X_N(k))$ be a set of N continuous variables measured at locations $k \in \boldsymbol{K}$. Locations are connected by currents which together form a spatial network (fig. 1a).

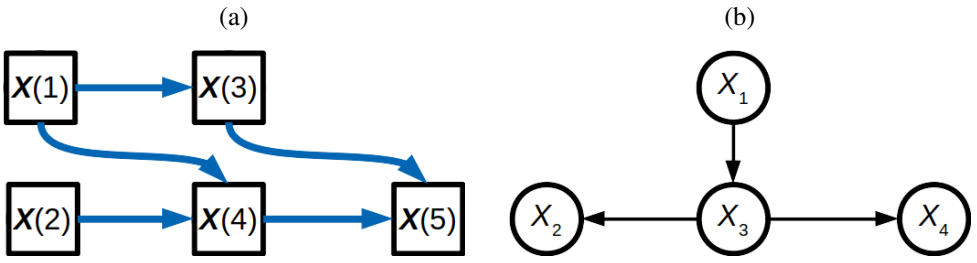

Figure 1: An example of the type of system that we consider. It is described by two separate graphs: (a) a graph describing the spatial network, indicating relations between locations and (b) a graph describing the causal network, indicating relations between variables at the same location. We intentionally use different symbols in the two figures to stress their different function.

This network is similar to a Markov chain as each location depends only on itself and the directly preceding locations in the direction of the current (Breiman (1992)). Importantly, locations are not influenced by locations further down the current. In contrast to a chain, however, the current arms of

the network can split and unite. The entire set of locations directly preceding location k is denoted by $\mathbf{Pre}(k)$. In addition to the spatial network, the system is described by a causal network (fig. 1b). We assume spatial invariance, meaning that the causal network is shared among all locations k. The following definition describes the different types of variables in the system that we consider.

Definition 1 *Let $\boldsymbol{X}$ be a set of variables measured in a system with a directional current. Then there is a partitioning $(\boldsymbol{I}, \boldsymbol{O}, \boldsymbol{R})$ with:*

1. *The subset $\boldsymbol{I}(k) = (I_1(k), ..., I_{n_I}(k))$ of variables that are an effect of previous locations and a cause of subsequent locations (i.e., the variables that are affected by the current, e.g., chemical concentrations).*

2. *The subset $\boldsymbol{O}(k) = (O_1(k), ..., O_{n_O}(k))$ of variables that are exogenous to the system: These may be a cause of the other variables, but not an effect (e.g., riverside activities). Exogenous variables of different locations are assumed to be uncorrelated.*

3. *The subset $\boldsymbol{R}(k) = (R_1(k), ..., R_{n_R}(k))$ of variables that do not fit either of these categories (e.g., the substrate of a river).*

The spatial structure of the system is captured by an additional set of variables, constructed from potentially multiple sets of variables $\boldsymbol{I}$ (one for each preceding location), $\boldsymbol{U}(k) = (U_1(k), ..., U_{n_I}(k))$ $= (f(I_1(\boldsymbol{Pre}(k))), ..., f(I_{n_I}(\boldsymbol{Pre}(k))))$, with f being a function used to merge multiple locations.

With the constructed variables $\boldsymbol{U}$, the system can be represented by a single graph $\mathcal{G}$ with vertices $\boldsymbol{V} = \{\boldsymbol{U}, \boldsymbol{O}, \boldsymbol{R}, \boldsymbol{I}\}$ and edges $\boldsymbol{E} \subseteq \boldsymbol{V} \times \boldsymbol{V}$ (fig. 2). For each i and location k, $X_i(k)$ is

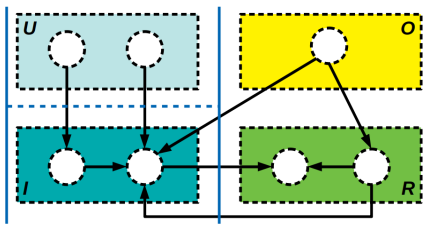

Figure 2: Schematic graph describing the data generating process system. Variables $\boldsymbol{I}$ are affected by the current and are thus potentially caused by locations further up the stream ($\boldsymbol{U}$). Variables $\boldsymbol{O}$ are exogenous to the system, while variables $\boldsymbol{R}$ may interact with variables $\boldsymbol{I}$, but only locally. An edge from one variable set into another indicates that a variable in the first set can potentially be a cause of a variable in the second set.

drawn from a distribution $P(X_i|\mathbf{Pa}_{\mathcal{G}}(X_i))$ with the system at equilibrium. Note that P does not depend on the location k which means that the IID assumption is satisfied. Our main assumption then is that causal relationships against the direction of the current are impossible. We further assume that variables in $\boldsymbol{I}$ are always caused by their counterpart in $\boldsymbol{U}$. The subgraph over vertices $\{\boldsymbol{I}, \boldsymbol{R}\}$ is called the local causal graph. Latent variables are allowed, but cycles are not.

3.2 Preprocessing

We start from a set of measured variables $\boldsymbol{X}$ at measurement locations $\boldsymbol{K}$ to learn a local causal graph $\mathcal{G}$ that is valid for all locations. Trivially, we could run the FCI algorithm. This, however, is suboptimal due to several reasons. First, the measurements are not IID, violating one of the basic assumptions of the FCI algorithm. Second, by neglecting the spatial structure, we would be neglecting a lot of information that could potentially be useful (Spirtes et al. (2000)).

To avoid this loss of information, we construct upstream variables $\boldsymbol{U}$ as outlined in def. 1, using the mean as function f. We stress that this choice depends on the application. For example, for a system with currents of largely differing discharge volume, the weighted average might be a better choice. Further, we note that not all locations have a preceding location which results in locations with an incomplete set of variables. In this work, we wanted to evaluate our general approach, which is why we decided to exclude missing data imputation as a potential influence on modeling performance. In principle, however, the missing data problem could be tackled by any appropriate strategy, including regression imputation and Bayesian estimation (Enders (2010)). The spatial structure of the system could offer additional information here that could allow for better imputation of missing data. Note that our strategy of excluding locations with an incomplete set of variables slightly reduces the size of the data set. Our preprocessing strategy is outlined in alg. 1.

Algorithm 1: Current data preprocessing

Input : Set of measured variables $\boldsymbol{X}$ at locations $\boldsymbol{K}$ in a system with directional currents, with $\boldsymbol{K'}$ being the set of locations $k \in \boldsymbol{K}$ for which $\mathbf{Pre}(k) \neq \emptyset$

Output: Set of variables $\boldsymbol{X}' = \{\boldsymbol{U}, \boldsymbol{O}, \boldsymbol{R}, \boldsymbol{I}\}$ at locations $\boldsymbol{K}'$

1 Partition variables into subsets $\boldsymbol{I}$, $\boldsymbol{O}$ and $\boldsymbol{R}$ following def. 1;
2 **repeat**
3 Select an unvisited measurement location $k \in \boldsymbol{K'}$ and calculate $\boldsymbol{U}(k)$ following def. 1, using the average as the function f;
4 **until** *All measurement points $k \in \boldsymbol{K'}$ have been visited*;
5 Remove entries of locations $k \in \boldsymbol{K}$ for which $\mathbf{Pre}(k) = \emptyset$;

3.3 Connection to time series

After preprocessing, the data is very similar to time series data: Whereas for time series, temporal invariance says that processes do not change over time, spatial invariance says that processes are invariant of the location. Time series are most commonly analyzed with the concept of Granger causality (Granger (1969)). which is based on the assumption that a cause precedes an effect in time. If a variable x holds information on another variable y at a later time point, it is said that x Granger-causes y.

This principle does not align well with our setup. First, it is not obvious how to incorporate variables that do not adhere to the current ($\boldsymbol{R}$ and $\boldsymbol{O}$). Second, Granger causality needs extensions (see, e.g., Eichler (2007)) to allow for contemporaneous causal relationships between variables. Third, Granger causality does not allow for latent variables which is often unrealistic (but see Chu and Glymour (2008)). Fourth, our data is not a single series of observations. Due to branching, it is not possible to order the observations which means that the basic idea of Granger causality which is to predict one variable using the time series of another variable cannot reach its full potential.

3.4 Adaptation of FCI: currentFCI

We build our approach on the FCI algorithm which allows for latent confounders. Trivially, we could apply the standard FCI algorithm to our extended set of measurements. We do, however, have additional information readily available that enables us to do better.

1. There cannot be a cause-effect relationship of any variable in $\boldsymbol{X} \setminus \boldsymbol{U}$ on any variable in $\boldsymbol{U}$, or of any variable in $\boldsymbol{X} \setminus \boldsymbol{O}$ on any variable in $\boldsymbol{O}$.
2. Counterpart variables always have a cause-effect relationship in the direction of the current.
3. Spatial invariance dictates that the causal structure is independent of the location. Therefore, the graph over $\boldsymbol{U}$ has to be the same as the graph over $\boldsymbol{I}$.

Parts of our adjustments of the FCI algorithm (alg. 2) contribute directly to a more accurate causal graph. As we know that variables in $\boldsymbol{U}$ and $\boldsymbol{O}$ cannot be caused by variables in $\boldsymbol{X} \setminus \boldsymbol{U}$ and $\boldsymbol{X} \setminus \boldsymbol{O}$, respectively, we can orient all edge marks of $\boldsymbol{U}$ and $\boldsymbol{O}$ into another vertex as arrowheads. Edges between a variable in $\boldsymbol{U}$ and a variable in $\boldsymbol{O}$ are oriented as confounders solely based on this strategy. For counterpart variables, we orient the edges into the variables in $\boldsymbol{U}$ as tails, because there has to be a causal relationship due to our assumptions on the data generating process (line 6).

Likewise, parts of our adjustments improve the inferred causal graph indirectly by reducing the number of conditional independence tests that we have to perform. We know that counterpart variables will be connected by an edge. Therefore, we do not have to test for these edges, reducing the number of potential mistakes (line 4). Lastly, as the structure of $\boldsymbol{U}$ is equivalent to the structure of $\boldsymbol{I}$, we would like to avoid testing for conditional independence of edges within $\boldsymbol{U}$. We have to make sure, though, that not doing the tests does not compromise our ability to learn the rest of the causal graph. We have two trivial options: To cut all edges or to keep all edges within $\boldsymbol{U}$. Both options, however, negatively impact our ability to test for conditional independence between variables in $\boldsymbol{U}$ and variable in $\boldsymbol{X} \setminus \boldsymbol{U}$ (see fig. SI 1 in Mielke (2020)).

We propose to, again, use spatial invariance to our advantage. Spatial invariance dictates that the causal graph over vertices $\boldsymbol{U}$ and the causal graph over vertices $\boldsymbol{I}$ are the same with variables replaced by their counterparts. This allows us to infer the graph over $\boldsymbol{I}$ first (line 1), which we then use as the graph over $\boldsymbol{U}$ (line 3). Theorem 2 states the guarantees of our algorithm. A sketch of the proof is in the supplementary information (available via Mielke (2020)).

Theorem 2 *Given the data generating process outlined in section 3.1, the adapted FCI algorithm outlined in alg. 1 and alg. 2 gives sound results for the local part of the PAG.*

4. EXPERIMENTS

The code of our experiments is available at Mielke (2020). We used R, version 3.6.3. (R Core Team (2020)). Our implementation of currentFCI is based on the R package pcalg, version 2.6.6. (Kalisch et al. (2012)) and we used a partial correlation test to test for conditional independence with the assumption of linearity and a maximum size of conditioning sets of 3.

4.1 Simulations

We first compared the performance of currentFCI on simulated data to that of standard FCI. For that, we generated 100 random spatial current structures (fig. 1a) with pre-set numbers of measurement

Algorithm 2: currentFCI.

Input : Set of variables $\boldsymbol{X}' = \{\boldsymbol{U}, \boldsymbol{O}, \boldsymbol{R}, \boldsymbol{I}\}$ at measurement locations $\boldsymbol{K}'$
Output: PAG over vertices $\boldsymbol{X}'$

1 Infer the undirected graph $\mathcal{M}$ on the vertex set $\mathbf{I}$ following the stable FCI skeleton phase;
2 Form the complete undirected graph $\mathcal{G}$ on the vertex set $\mathbf{X}'$;
3 Remove all edges from $\mathcal{G}$ that were removed in $\mathcal{M}$ and translate $\mathcal{M}$ to counterpart variables to remove edges within $\boldsymbol{U}$;
4 Standard skeleton search while avoiding tests for edges between two variables in $\boldsymbol{U}$ and edges between counterpart variables;
5 Conservatively orient colliders (Ramsey et al. (2012)) in $\mathcal{M}$ and translate to counterpart variables to orient edges in $\boldsymbol{U}$;
6 Orient edges involving $\boldsymbol{U}$ as $U \circ\!\!\rightarrow$, edges involving $\boldsymbol{O}$ as $O \circ\!\!\rightarrow$ and edges between counterpart variables as $U \rightarrow I$;
7 Possible d-separation phase while not overruling background knowledge;
8 Remove edges within $\boldsymbol{U}$, orient all remaining edges as $\circ\!\!-\!\!\circ$ and repeat step 6;
9 Standard orientation phase Zhang (2008) while not overruling background knowledge;

locations of $n_{\text{meas}} \in \{250, 1000, 4000\}$. For each current structure, we generated a random causal structure. We adapted a method described in Kalisch and Bühlmann (2007) to the data generating process described in section 3.1. We imposed the causal link between counterpart variables and excluded links that were ruled out by the data generating process. All other links were present with probability $s = 0.2$ and each pair of variables that were not linked was confounded with a probability $l = 0.01$. The noise was sampled from a normal distribution and links between variables were linear with a random strength of $\pm[0.1, 1]$. We simulated $I = 10$, $O = 5$ and $R = 5$ variables and build models for different confidence levels α of the conditional independence tests.

To evaluate the algorithms, we calculated evaluation metrics for both edges and edge marks. We compared the inferred models of both algorithms to the true underlying models and counted true positives (TPs), true negatives (TNs), false positives (FPs) and false negatives (FNs) for edges and edge marks across the 100 random realizations of each combination of n_{meas} and α. We summed over all realizations and calculated the precision $=$ TP/(TP + FP) and the recall $=$ TP/(TP + FN) for edges, arrowhead marks and tail marks separately. To make the comparison fair, we evaluated the local causal structure only, excluding edges and edge marks that we imposed in our algorithm.

The results are shown in fig. 3 and table 1. Both the precision and the recall of currentFCI were higher than those of standard FCI, independent of the number of locations. As currentFCI uses more information than standard FCI, it is better able to assess the relevancy of a correlation and to identify latent confounders. In table 1, we show the confusion matrices for one specific configuration (1000 observations and $\alpha = 0.01$). For the edges, we found that the number of false positives of currentFCI is very low, meaning that edges that are present in the inferred model are almost always part of the true underlying graph. For the edge marks, currentFCI has much higher recall ($> +100\%$) and accuracy ($> +20\%$).

To test the stability of currentFCI in situations in which the true function to merge multiple locations is unknown, we performed an additional analysis. In this analysis, we imitated a distorted aggregation function by adding additional normally distributed noise to the upstream variables $\boldsymbol{U}$. The full results of the analysis are available in the supplementary information (Fig. SI 2 - 4 in

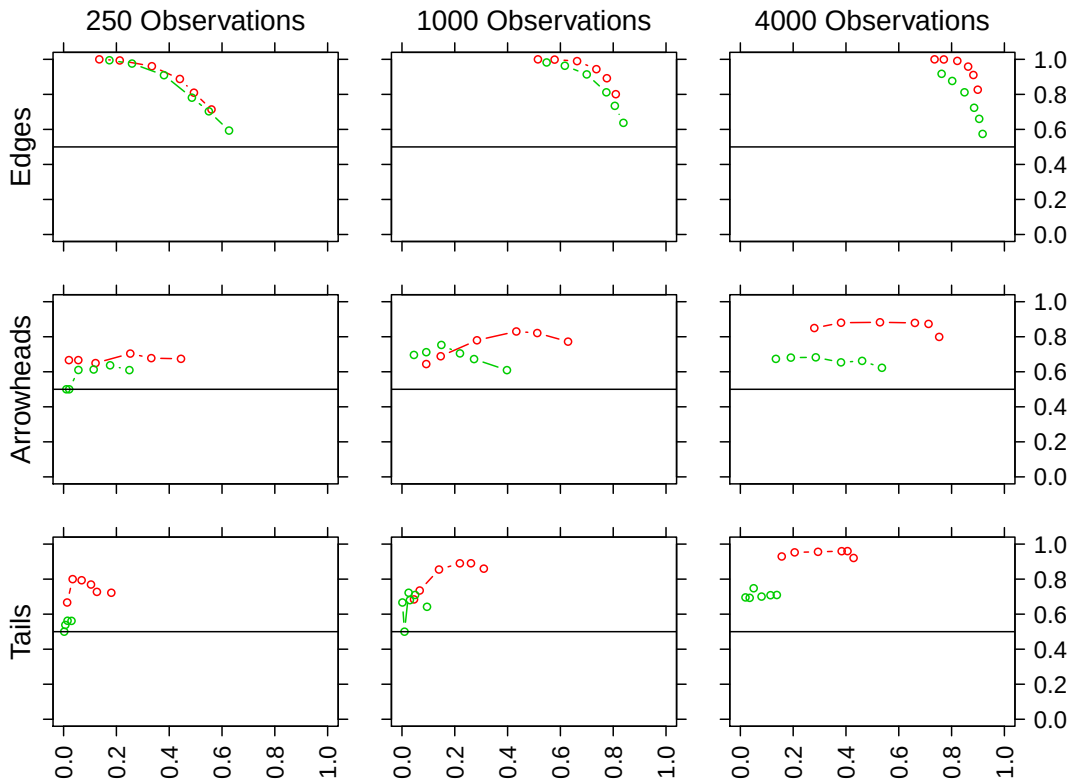

Figure 3: Results of the simulation study. Recall (x-axis) against precision (y-axis) obtained for the standard FCI (green) and the currentFCI (red) for different numbers of observations (columns) and edges (top) and edge marks (center and bottom), respectively.

Mielke (2020)). We found that the performance of currentFCI declined slightly with increasing levels of additional noise, but was still superior to that of standard FCI in all tested scenarios.

	currentFCI			
	$\rightarrow$	$-\circ$	$-$	$\times$
$\rightarrow$	709	884	44	585
$-$	145	1127	357	
$\times$		97		8185

	standard FCI			
	$\rightarrow$	$-\circ$	$-$	$\times$
$\rightarrow$	380	1326	24	501
$-$	159	1494	51	
$\times$		398		7884

Table 1: Confusion matrices for $n_{\text{meas}} = 1000$ and $\alpha = 0.05$ for currentFCI (left) and standard FCI (right) with inferred edge marks in columns and true edge marks in rows.

4.2 Case Study

Finally, we applied currentFCI to find causes of fish community integrity in rivers in the US state of Ohio (fig. 4a). Using hydrosheds (Lehner et al. (2008)), we linked each observation to the closest river. Fish community integrity is measured by the multimetric Index of Biotic Integrity (IBI). Potential causes include chemical concentrations (which together formed the set $\boldsymbol{I}$) and stream characteristics ($\boldsymbol{R}$), as well as riverside activities and geographic location ($\boldsymbol{O}$). Knowing the potential causes of fish community integrity is vital to perform cost-effective conservation measures. We assumed that fish integrity cannot be a cause of any other variable and consequently oriented all resulting edges into the IBI as arrowheads.

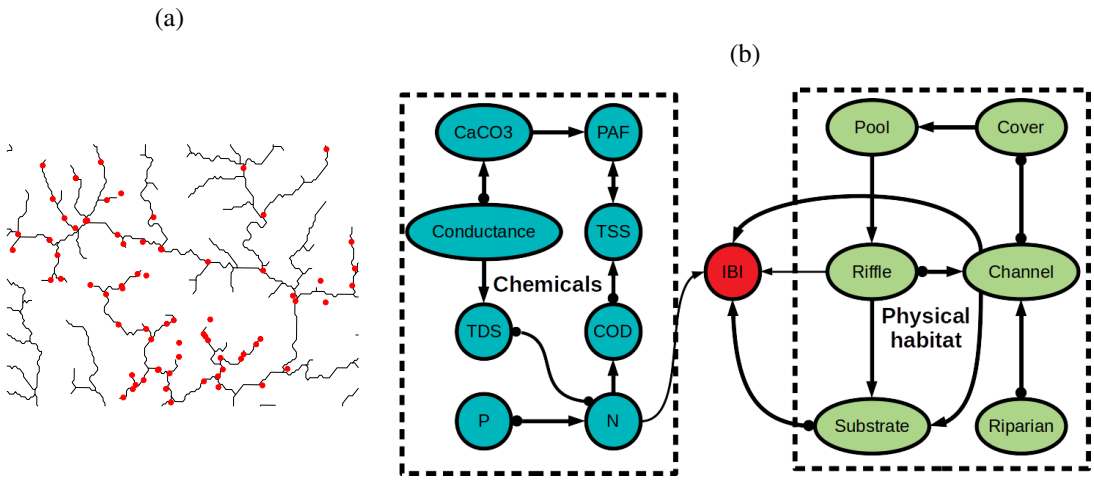

Figure 4: Case study: (a) a small extract of rivers (black) and measurement locations (red) in the US state of Ohio and (b) inferred PAG over the Index of Fish Integrity (red, IBI), chemicals (blue), and stream characteristics (green). The width of an edge is proportional to the number of graphs that it appeared in. For clarity, we do not show vertices of $\boldsymbol{U}$ and $\boldsymbol{O}$ variables here.

As some of our variables were highly non-normally distributed, we calculated ranks for all variables and used Spearman's rank correlations as the input of the conditional independence tests. After preprocessing, our data set consisted of 1167 observations without missing values. $\boldsymbol{I}$ consisted of 8 variables, $\boldsymbol{R}$ consisted of 6 variables and $\boldsymbol{O}$ consisted of 4 variables. To get a robust final result, we performed bootstrapping with 1000 samples with each sample being drawn with replacement and of the size of the original data set. In the modeling, we used a confidence level for the conditional independence tests of $\alpha = 0.01$. Both edges and edge marks had to be found in at least 50% of the graphs to be reported in the combined graph. The final PAG is shown in fig. 4b.

We found three causes of the IBI. The quality of the river channel (Channel), the quality of fast-flowing, shallow river parts (Riffle), and the concentration of nitrogen in the water (N). Of these three causes, Channel was by far the most reliable one (being found in more than 97% of all samples) whereas the connections to Riffle and N appeared only in a relatively small subset of models (less than 60%). A previous study with boosted regression trees (Pilière et al. (2014)) concluded that Channel and Riffle have the strongest association with the IBI. Our analysis suggests that these

associations are likely causal. Pilière et al. (2014), however, also reported a strong association of the IBI and phosphorus concentration (P), which in our model are not adjacent. In general, we found the water chemistry to have less of an impact on the IBI than physical habitat quality. This finding indicates that physical habitat should be the main focus of conservation efforts in lotic ecosystems.

5. DISCUSSION

We found that currentFCI outperformed the standard FCI algorithm regardless of the size of the dataset. We were consistently able to identify edges with better precision, while for edge marks both precision and recall were higher. The differences were largest for medium and large datasets (1000 or 4000 observations). Specifically here, the standard FCI algorithm struggled with latent confounders in the form of upstream and out of stream variables. As we included those in current-FCI, we were able to identify the missing confounders which improved the overall performance.

Finally, we hint to possible extensions that we did not incorporate in our work. We assumed that out of stream variables are not correlated between different locations. This assumption might not be justified in some situations and might require an extension. As these correlations are likely undirected, we think that chain graphs (Lauritzen and Richardson (2002)) would be a natural choice. Further, we encountered infrequent conflicts ($< 1\%$) between background knowledge and orientation rules. In this work, we let background knowledge overrule orientation rules. While the low conflict rate strengthens our approach in general, further investigation into the best strategy to handle conflicts is required.

References

L. Breiman. *Probability*. Society for Industrial and Applied Mathematics, 1992.

T. Chu and C. Glymour. Search for additive nonlinear time series causal models. *Journal of Machine Learning Research*, 9:967–991, 2008.

T. Chu, D. Danks, and C. Glymour. Data driven methods for nonlinear Granger causality: Climate teleconnection mechanisms. 2005.

T. Claassen and T. Heskes. A Bayesian approach to constraint-based causal inference. In *Conference on Uncertainty in Artificial Intelligence*, pages 207–216, 2012.

D. Colombo and M. H. Maathuis. Order-independent constraint-based causal structure learning. *The Journal of Machine Learning Research*, 15(116):3921–3962, 2014.

P. Dagum, A. Galper, and E. Horvitz. Dynamic network models for forecasting. In *Conference on Uncertainty in Artificial Intelligence*, pages 41–48, 1992.

I. Ebert-Uphoff and Y. Deng. Causal discovery from spatio-temporal data with applications to climate science. In *International Conference on Machine Learning and Applications*, pages 606–613, 2014.

M. Eichler. Granger causality and path diagrams for multivariate time series. *Journal of Econometrics*, 137(2):334–353, 2007.

C. K. Enders. *Applied missing data analysis*. Guilford press, 2010.

D. Entner and P. O. Hoyer. On causal discovery from time series data using FCI. In *Conference on Probabilistic Graphical Models*, pages 121–128, 2010.

C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, 37(3):424–438, 1969.

A. Hyttinen, F. Eberhardt, and M. Järvisalo. Constraint-based causal discovery: Conflict resolution with answer set programming. In *Conference on Uncertainty in Artifical Intelligence*, pages 340–349, 2014.

K. M. Irvine and A. I. Gitelman. Graphical spatial models: a new view on interpreting spatial pattern. *Environmental and ecological statistics*, 18:447–469, 2011.

M. Kalisch and P. Bühlmann. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8:613–636, 2007.

M. Kalisch, M. Mächler, D. Colombo, M. H. Maathuis, and P. Bühlmann. Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, 47(11):1–26, 2012.

S. L. Lauritzen and T. S. Richardson. Chain graph models and their causal interpretations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):321–348, 2002.

B. Lehner, K. L. Verdin, and A. Jarvis. New global hydrography derived from spaceborne elevation data. *Eos, Transactions, American Geophysical Union*, 89(10):93–94, 2008.

K. P. Mielke. Github repository: currentFCI, 2020. doi: 10.5281/zenodo.4010939.

J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.

A. Pilière, A. M. Schipper, A. M. Breure, L. Posthuma, D. de Zwart, S. D. Dyer, and M. A. J. Huijbregts. Comparing responses of freshwater fish and invertebrate community integrity along multiple environmental gradients. *Ecological indicators*, 43:215–226, 2014.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL `https://www.R-project.org/`.

J. Ramsey, J. Zhang, and P. L. Spirtes. Adjacency-faithfulness and conservative causal inference. *preprint arXiv:1206.6843*, 2012.

T. Richardson. A discovery algorithm for directed cyclic graphs. *preprint arXiv:1302.3599*, 2013.

T. Richardson and P. Spirtes. Ancestral graph Markov models. *The Annals of Statistics*, 30(4): 962–1030, 2002.

P. Spirtes, C. N. Glymour, and R. Scheines. *Causation, prediction, and search*. MIT press, 2000.

J. Zhang. A characterization of Markov equivalence classes for directed acyclic graphs with latent variables. In *Conference on Uncertainty in Artificial Intelligence*, pages 450–457, 2007.

J. Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873–1896, 2008.