

Vector Causal Inference between Two Groups of Variables

Jonas Wahl^{*,1,2}, Urmi Ninad^{*,1,2}, Jakob Runge^{1,2}

¹Technische Universität Berlin

²DLR Institut für Datenwissenschaften Jena

wahl@tu-berlin.de, urmi.ninad@tu-berlin.de, runge@tu-berlin.de

Abstract

Methods to identify cause-effect relationships currently mostly assume the variables to be scalar random variables. However, in many fields the objects of interest are vectors or groups of scalar variables. We present a new constraint-based non-parametric approach for inferring the causal relationship between two vector-valued random variables from observational data. Our method employs sparsity estimates of directed and undirected graphs and is based on two new principles for groupwise causal reasoning that we justify theoretically in Pearl’s graphical model-based causality framework. Our theoretical considerations are complemented by two new causal discovery algorithms for causal interactions between two random vectors which find the correct causal direction reliably in simulations even if interactions are nonlinear. We evaluate our methods empirically and compare them to other state-of-the-art techniques.

Introduction

In recent years, many methods have been developed in order to infer cause-effect relationships between random variables from observational data, see e.g. Pearl (2009); Spirtes, Glymour, and Scheines (2000); Shimizu et al. (2006); Peters, Janzing, and Schölkopf (2017); Runge et al. (2015). Most often these random variables are assumed to be scalar; an assumption which covers many but by no means all questions of interest in science. For instance, neuroscience researchers are often interested in the causal interactions between different brain *regions* each of which are represented by a multitude of measurement locations in fMRI data. Similarly, in the Earth sciences, researchers would like to understand causal relationship between variables (such as sea surface temperature, air pressure or wind speed) that have each been measured at a number of grid locations in predefined areas on the planet. To make inferences in such a setup, standard approaches often proceed to drastically reduce the number of measurement variables, for instance by computing average values across each region of interest or by applying statistical dimension reduction techniques such as principal component analysis. However, aggregating data in such a way might lead to faulty causal conclusions: For

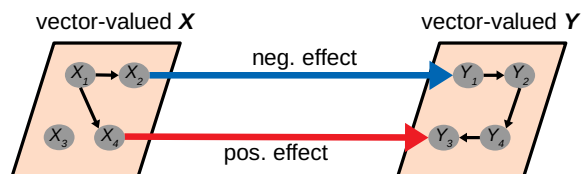


Figure 1: Two components in the vector-valued variable X have effects on Y that are of opposing signs, such that aggregation of X and Y leads to a dilution or cancellation of dependence.

instance, conditional independencies $X \perp\!\!\!\perp Y | Z$ between vectors might not be visible in their mean values (Spirtes, Glymour, and Scheines 2000). Opposing causal effects from different parts of a region might average to zero, and causally relevant information might be diluted or lost when considering aggregate variables, see Figure 1. Furthermore, methods that rely on the non-Gaussianity of noise to make inferences about the causal direction, such as LinGaM (Shimizu et al. 2006), are rendered weak by the averaging due to the central limit theorem driving the average noise closer to a Gaussian.

In this work, we aim to develop new techniques to infer the causal relationship between two groups of variables, represented as random *vectors*, without a dimension reduction step. Assuming that the causal arrows go from one group to the other only, the most straightforward way to do so is to run a standard causal discovery algorithm, e.g. the PC algorithm, on all microvariables (i.e. all entries of both vectors), and then choose the cause group as the one that has the most edges pointing to the other group. When groups become large, this approach, henceforth called *Vanilla-PC*, has disadvantages: since it needs to determine the full causal ‘microstructure’, it has to run many conditional independence tests and due to the sequential error propagation of the PC algorithm becomes unreliable quickly at small sample size (see Subsection 12 in the Supplement for an empirical illustration of this). In essence, *Vanilla-PC* computes more structure than is needed to answer the causal query at hand and therefore uses the data inefficiently. We therefore take a different road and combine the constraint-based approach for causal discovery with sparsity measures of the *internal (causal) structure* of the groups. Our methods are based on

*These authors contributed equally.

the following two principles for causal interaction between groups of random variables:

- (P1) generically, conditioning on the cause group does not *create* new conditional dependencies *within* the effect group;
- (P2) generically, conditioning on the effect group does not *delete* conditional dependencies *within* the cause group.

Here, the term *generically* is to be understood as in other assumptions for causal inference such as the causal Markov property, Faithfulness and the principle of independent cause and mechanism (ICM), see e.g. Peters, Janzing, and Schölkopf (2017). That is to say, (P1) and (P2) can only be violated if the causal mechanism is in some way fine-tuned to the exogenous noise variables of the model.

We justify these principles more thoroughly by considering the setting where the scalar microvariables are modelled by a causal graphical model over a directed acyclic graph (DAG). In this setup, (P1) and (P2) turn out to be *implied* by the causal Markov property and Faithfulness. Based on these principles, we prove that, in this purely causal setup, the correct causal direction is identifiable under weak assumptions (Theorem 1). Moreover, we provide an algorithm for distinguishing cause from effect called **2G-VecCI.PC** (two group vector causal inference, PC method) which is based on density estimation of each group through the PC algorithm and which is sound and complete under said assumptions. To our knowledge, our method is the first non-parametric method to infer causal directionality between two groups of variables (other than Vanilla-PC).

In addition to the purely causal setting, we consider the setup where cause and effect group are related through a structural causal model of the form

$$\begin{aligned} \mathbf{X} &:= \eta_{\mathbf{X}}, \\ \mathbf{Y} &:= \mathbf{f}(\mathbf{X}, \eta_{\mathbf{Y}}), \quad \eta_{\mathbf{X}} \perp\!\!\!\perp \eta_{\mathbf{Y}}. \end{aligned}$$

Importantly, we do not assume that the individual components of the noise vector $\eta_{\mathbf{X}}$ (respectively $\eta_{\mathbf{Y}}$) are pairwise independent. Such a model can be reasonable when the internal interactions *within* a variable group do not admit a straightforward causal interpretation while the interactions *between* groups do. One might therefore consider such a model *semi-causal*. For instance, if \mathbf{X} describes a field of surface temperature measurements on different grid locations, stating that the measurement at location i causes the measurement at location j might be inappropriate. For such a semi-causal SCM, it is much harder to prove theoretical guarantees for the validity of (P1) and (P2), and we only demonstrate what violations of these principles would entail in a toy example. Nevertheless a second version of our inference algorithm dubbed **2G-VecCI.Full** (two group vector causal inference, full conditioning method) is able to find the correct causal direction in many cases in simulated data. In both the causal and the semi-causal setting, our algorithms make inferences by estimating the sparsity of graphs that encode conditional dependence or causal relationships within a variable group before and after conditioning on the other variable group. At present our methods

assume that samples are i.i.d., and that the sample size is larger than the total size of the vector-valued variables. Both methods are based purely on conditional (in)dependence relationships and, with appropriately chosen tests, work even if causal interactions are nonlinear.

We will present our theoretical identifiability results and the necessary assumptions in Section 3. After that, we describe two different versions of our algorithm for causal discovery between variable groups in Section 4. In Section 5, we analyse the empirical performance of these algorithms in experiments with synthetic data and compare it to that of other approaches (Vanilla-PC and the Trace Method (Janzing et al. 2009; Zscheischler, Janzing, and Zhang 2012)). We also consider a real world climate science example of surface temperatures in the El Niño Southern Oscillation (ENSO 3.4) region in the Pacific and in British Columbia to test our algorithms. We conclude with a discussion and outlook in Section 6.

Related Work

Although the majority of causal discovery results focus on scalar variables, the idea to study causal interactions between groups of random variables is not new. On the theoretical side, Rubenstein* et al. (2017), Chalupka, Eberhardt, and Perona (2016), Chalupka et al. (2016) and Chalupka, Eberhardt, and Perona (2017) discuss to which extent micro variables can be aggregated to macro variables without losing causal information. Parviainen and Kaski (2017) discuss theoretical assumptions of multi-group causal discovery in connection to the PC-algorithm (as opposed to the two group identifiability problem discussed here). For two linearly interacting groups, Janzing et al. (2009) introduce a causal discovery algorithm called the Trace method, see also Zscheischler, Janzing, and Zhang (2012). In Entner and Hoyer (2012), scalar causal discovery techniques based on non-Gaussianity assumptions such as LiNGaM (Shimizu et al. 2006) are generalized to the vector-valued setting. We summarize the existing approaches, as well as their assumptions, strengths and weaknesses in Table 7 in the supplement.

Identifiability Results

Theoretical Setup We will consider scalar random variables $X_1, \dots, X_n, Y_1, \dots, Y_m$ that are grouped into two vectors $\mathbf{X} = \{X_1, \dots, X_n\}$, $\mathbf{Y} = \{Y_1, \dots, Y_m\}$. We assume that the data is generated by a causal process $\mathbf{X} \rightarrow \mathbf{Y}$ as outlined below, and our goal is to infer the correct causal direction from the observational distribution $P_{\mathbf{X}, \mathbf{Y}}$. We will operate under different sets of assumptions that relate $P_{\mathbf{X}, \mathbf{Y}}$ to causal representations (see Model 1 and 2 below). We refer to the mathematical appendix for a quick overview on directed and undirected graphs, d-separation, the causal Markov property, Faithfulness and causal sufficiency. We will always assume that a statistical association $\mathbf{X} \not\perp\!\!\!\perp \mathbf{Y}$ is present in the data.

Model 1 (Unidirectional Causal Vector Model) The scalar variables $X_1, \dots, X_n, Y_1, \dots, Y_m$ are represented as

the nodes of a directed acyclic graph (DAG) \mathcal{G} . In addition, we assume that

- (A1) The joint distribution $P_{\mathbf{X}, \mathbf{Y}}$ fulfills the causal Markov property and is faithful to \mathcal{G} . In other words, d-separation in \mathcal{G} completely characterizes the conditional independencies present in $P_{\mathbf{X}, \mathbf{Y}}$. This implies that the model is causally sufficient, i.e. no hidden confounders are present.
- (A2) Arrows between groups only point in one direction, i.e., without loss of generality, the $\mathbf{X} \rightarrow \mathbf{Y}$ -direction. In other words, there can be no arrow $X_k \leftarrow Y_\ell$ for any $X_k \in \mathbf{X}, Y_\ell \in \mathbf{Y}$.

We will now justify principles (P1) and (P2) in this scenario. The main result of this section will be summarized in Theorem 1. Proofs of the following results will be given in the technical appendix.

Lemma 1. *Assume that the assumptions of Model 1 are satisfied. Then principle (P1) holds in the sense that there is no subset $\mathcal{S} \subset \mathbf{Y}$ of the effect group and nodes $Y_k, Y_\ell \in \mathbf{Y}$ such that*

$$Y_k \perp\!\!\!\perp Y_\ell \mid \mathcal{S} \quad \text{and} \quad Y_k \not\perp\!\!\!\perp Y_\ell \mid \mathcal{S}, \mathbf{X}.$$

Next, we say that a causal vector model $\mathbf{X} \rightarrow \mathbf{Y}$ satisfies condition

- (C1) if there is a subset $\mathcal{S} \subset \mathbf{X}$ and scalar variables $X_i, X_j \in \mathbf{X}$ such that

$$X_i \perp\!\!\!\perp X_j \mid \mathcal{S} \quad \text{and} \quad X_i \not\perp\!\!\!\perp X_j \mid \mathcal{S}, \mathbf{Y}.$$

In the appendix, we will characterize (C1) graphically and depict some motivational examples. For instance, Condition (1) is satisfied only if there exists a *cross-regional v-structure* $X_i \rightarrow Y_k \leftarrow X_j$. We can now deduce the following result for cause-effect identification. It states that whenever conditioning on a group creates dependencies within the other group, by (P1) the former must be the effect and the latter must be the cause.

Corollary 1. *Assume that the assumptions of Model 1 as well as (C1) are satisfied. Then, the causal direction $\mathbf{X} \rightarrow \mathbf{Y}$ can be inferred from the observational distribution $P_{\mathbf{X}, \mathbf{Y}}$.*

Next, we justify principle (P2).

Lemma 2. *Assume that the assumptions of Model 1 are satisfied. Then principle (P2) holds, in the sense that there is no subset $\mathcal{S} \subset \mathbf{X}$ of the cause group and nodes $X_i, X_j \in \mathbf{X}$ such that*

$$X_i \not\perp\!\!\!\perp X_j \mid \mathcal{S} \quad \text{and} \quad X_i \perp\!\!\!\perp X_j \mid \mathcal{S}, \mathbf{Y}.$$

Again, we need to ensure that conditioning on the cause vector *does not* delete dependencies within the effect vector. We therefore say that a causal vector model $\mathbf{X} \rightarrow \mathbf{Y}$ satisfies

- (C2) if there is a subset $\mathcal{S} \subset \mathbf{Y}$ and scalar variables $Y_k, Y_\ell \in \mathbf{Y}$ such that

$$Y_k \not\perp\!\!\!\perp Y_\ell \mid \mathcal{S} \quad \text{and} \quad Y_k \perp\!\!\!\perp Y_\ell \mid \mathcal{S}, \mathbf{X}.$$

For example, condition (C2) is satisfied if Y_k, Y_ℓ can be d-separated by $\mathcal{S} \subset \mathbf{Y}$ in the subgraph over \mathbf{Y} and there is a common confounder $Y_k \leftarrow X_i \rightarrow Y_\ell$. Again, we will provide a full graphical characterization of (C2) and some examples in the appendix.

Corollary 2. *Assume that the assumptions of Model 1 as well as (C2) are satisfied. Then, the causal direction $\mathbf{X} \rightarrow \mathbf{Y}$ can be inferred from the observational distribution $P_{\mathbf{X}, \mathbf{Y}}$.*

We summarize the results above in the following theorem.

Theorem 1. *Assume that the assumptions of Model 1 are satisfied and that at least one of the conditions (C1) or (C2) holds. Then, the causal direction $\mathbf{X} \rightarrow \mathbf{Y}$ can be inferred from the observational distribution $P_{\mathbf{X}, \mathbf{Y}}$.*

Model 2 (Unidirectional Semi-Causal Vector Model)

Model 2 assumes that the variables $X_1, \dots, X_n, Y_1, \dots, Y_m$ are generated by the *semi-causal* structural causal model

$$\begin{aligned} \mathbf{X} &:= \eta_{\mathbf{X}}, \\ \mathbf{Y} &:= \mathbf{f}(\mathbf{X}, \eta_{\mathbf{Y}}), \quad \eta_{\mathbf{X}} \perp\!\!\!\perp \eta_{\mathbf{Y}}, \end{aligned}$$

where as mentioned before, we do *not* assume that the components within the noise terms $\eta_{\mathbf{X}}, \eta_{\mathbf{Y}}$ are pairwise independent. We can encode the conditional independencies within the \mathbf{X} -group graphically by drawing an undirected edge $X_i - X_j$ if and only if

$$X_i \not\perp\!\!\!\perp X_j \mid \mathbf{X} \setminus \{X_i, X_j\}$$

and similarly within the \mathbf{Y} -group by drawing an undirected edge $Y_k - Y_\ell$ if and only if

$$\eta_k \not\perp\!\!\!\perp \eta_\ell \mid \eta \setminus \{\eta_k, \eta_\ell\}.$$

The undirected graphs obtained in this way will be denoted by $\mathcal{G}'_{\mathbf{X}}, \mathcal{G}'_{\mathbf{Y}}$ respectively.

In this way, we have encoded the distributions of \mathbf{X}, \mathbf{Y} as *Markov random fields* over undirected graphs. Markov random fields of this kind are sometimes employed to model spatial or spatio-temporal measurements on grids (Song, Fuentes, and Ghosh 2008) such as for instance surface temperature measurements (Vaccaro et al. 2021).

In this setup, it is harder to find exact conditions that formally imply principles (P1), (P2). We lack a non-finetuning statement that is as general as faithfulness in the causal setting and quantifying such a statement would probably require additional assumptions on the functional form of the model or the noise distribution.

Let us illustrate why it is nevertheless reasonable to accept (P1) and (P2) with the following toy example:

Example 1. *Consider the model*

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} bX_1 + \eta_1 \\ cX_2 + \eta_2 \end{pmatrix},$$

where (X_1, X_2) are jointly normal with mean $\mathbb{E}[\mathbf{X}] = \mathbf{0}$, $\text{Var}(X_1) = \text{Var}(X_2) = 1$ and $\text{Cov}(X_1, X_2) = a$. The error terms (η_1, η_2) are jointly normal with mean $\mathbb{E}[\eta] = \mathbf{0}$, $\text{Var}(\eta_1) = \text{Var}(\eta_2) = 1$ and $\text{Cov}(\eta_1, \eta_2) = d$. The only way conditioning on \mathbf{X} could create a dependency in \mathbf{Y} would be if $Y_1 \perp\!\!\!\perp Y_2$ in the first place, which is equivalent to $abc = -d$. Thus conditioning on the cause can only create dependencies out of independencies that arose from a finetuning of the coefficients of the mechanism b, c to the coefficients of the noise terms a, d .

Similarly, for (P2) to be violated in this example, conditioning on \mathbf{Y} would have to delete the dependency of X_1

and X_2 . This would entail another (more involved) algebraic equation that the coefficients would have to satisfy. Thus coefficients describing the causal mechanism could not be chosen independently of the noise terms.

Graph edge density criterion for identifying causal direction A practical way to make use of Theorem 1 is to compare the edge densities of the internal graphs of one group before and after conditioning on the other group. Say we are interested in these internal graphs of the vector \mathbf{X} in Model 1 where internal graphs are formalized as DAGs. In this case we then compare the number of edges in the (skeleton of the) CPDAG $\mathcal{G}_{\mathbf{X}}$ that is the output of the PC-algorithm run on the scalar variables in \mathbf{X} to the number of edges in the CPDAG $\mathcal{G}_{\mathbf{X}|\mathbf{Y}}$ that is the output of the PC-algorithm over \mathbf{X} in which \mathbf{Y} is added as a conditioning set to every independence test. We normalize these edge counts by the maximal possible number of edges $\text{edgeMax} = n(n-1)/2$ to obtain edge densities

$$\text{edgeDens}(\mathcal{G}_{\mathbf{X}}) = \frac{\text{number of edges of } \mathcal{G}_{\mathbf{X}}}{\text{edgeMax}}$$

$$\text{edgeDens}(\mathcal{G}_{\mathbf{X}|\mathbf{Y}}) = \frac{\text{number of edges of } \mathcal{G}_{\mathbf{X}|\mathbf{Y}}}{\text{edgeMax}}.$$

The edge densities $\text{edgeDens}(\mathcal{G}_{\mathbf{Y}})$, $\text{edgeDens}(\mathcal{G}_{\mathbf{Y}|\mathbf{X}})$ are defined analogously. Presuming that the PC-algorithm is run with perfect oracle independence tests, Conditions (C1) and (C2) then have the following implications.

Theorem 2. *As before we assume the causal direction to be $\mathbf{X} \rightarrow \mathbf{Y}$ and that the assumptions of Model 1 are satisfied. If moreover condition (C1) holds, then*

$$d(\mathbf{X}|\mathbf{Y}) := \text{edgeDens}(\mathcal{G}_{\mathbf{X}|\mathbf{Y}}) - \text{edgeDens}(\mathcal{G}_{\mathbf{X}}) > 0. \quad (1)$$

If condition (C2) holds, then

$$d(\mathbf{Y}|\mathbf{X}) := \text{edgeDens}(\mathcal{G}_{\mathbf{Y}|\mathbf{X}}) - \text{edgeDens}(\mathcal{G}_{\mathbf{Y}}) < 0. \quad (2)$$

In either case, we have $d(\mathbf{X}|\mathbf{Y}) > d(\mathbf{Y}|\mathbf{X})$ and the causal direction can thus be inferred from the sign of $d(\mathbf{X}|\mathbf{Y}) - d(\mathbf{Y}|\mathbf{X})$.

As a consequence, when the causal direction is unknown, we can infer it from the observational distribution by computing $d(\mathbf{X}|\mathbf{Y})$ and $d(\mathbf{Y}|\mathbf{X})$ and choosing \mathbf{X} as the cause if the former is larger and \mathbf{Y} if the latter is larger. Note that this approach also works when \mathbf{X} and \mathbf{Y} have different internal edge densities.

In Model 2 we have to replace $\mathcal{G}_{\mathbf{X}}$ by the undirected conditional independence graph $\mathcal{G}'_{\mathbf{X}}$ and $\mathcal{G}_{\mathbf{X}|\mathbf{Y}}$ by the graph $\mathcal{G}'_{\mathbf{X}|\mathbf{Y}}$ that has edges $X_i - X_j$ iff

$$X_i \not\perp\!\!\!\perp X_j | \mathbf{X} \setminus \{X_i, X_j\}, \mathbf{Y}.$$

If we now replace $d(\mathbf{X}|\mathbf{Y})$, $d(\mathbf{Y}|\mathbf{X})$ by

$$d'(\mathbf{X}|\mathbf{Y}) := \text{edgeDens}(\mathcal{G}'_{\mathbf{X}|\mathbf{Y}}) - \text{edgeDens}(\mathcal{G}'_{\mathbf{X}}), \quad (3)$$

and

$$d'(\mathbf{Y}|\mathbf{X}) := \text{edgeDens}(\mathcal{G}'_{\mathbf{Y}|\mathbf{X}}) - \text{edgeDens}(\mathcal{G}'_{\mathbf{Y}}), \quad (4)$$

we still observe empirically (see below) that the sign of $d'(\mathbf{X}|\mathbf{Y}) - d'(\mathbf{Y}|\mathbf{X})$ is able to read the causal direction from the observational distribution quite efficiently.

Algorithms for Cause-Effect Identification

We now present two algorithms for cause-effect identification that are tailored to the two different models. 2G-VecCI.PC is particularly useful if the user believes the assumption of Model 1 to be valid. Recall that the PC-algorithm is an algorithm for causal discovery on scalar variables that consists of a **skeleton phase** to identify the skeleton of the causal graph and an orientation phase to determine causal directions. For the computation of edge densities, 2G-VecCI.PC will use the PC-algorithm's skeleton phase.

Algorithm 1: 2G-VecCI.PC

Data: two arrays containing samples of \mathbf{X} and \mathbf{Y} , parameter $\alpha \in [0, 1]$.

Result: variable with values ' \mathbf{X} is the cause of \mathbf{Y} ', or ' \mathbf{Y} is the cause of \mathbf{X} ', or 'Causal direction cannot be determined'.

1 Run **skeleton phase** on the components of \mathbf{X} ;

2 Compute

$$\widehat{\text{edgeDens}}(\mathcal{G}_{\mathbf{X}}) = \frac{\text{number of edges found in skeleton phase}}{\text{edgeMax}};$$

3 Run **skeleton phase** on the components of \mathbf{X} with \mathbf{Y} added to the conditioning set of every independence test;

4 Compute $\widehat{\text{edgeDens}}(\mathcal{G}_{\mathbf{X}|\mathbf{Y}})$ and

$$\widehat{d}(\mathbf{X}|\mathbf{Y}) = \widehat{\text{edgeDens}}(\mathcal{G}_{\mathbf{X}|\mathbf{Y}}) - \widehat{\text{edgeDens}}(\mathcal{G}_{\mathbf{X}});$$

5 Repeat (1) to (4) with exchanged roles of \mathbf{X} , \mathbf{Y} to get

$$\widehat{d}(\mathbf{Y}|\mathbf{X});$$

6 Compute $\text{Crit} = \widehat{d}(\mathbf{X}|\mathbf{Y}) - \widehat{d}(\mathbf{Y}|\mathbf{X})$;

7 **if** $|\text{Crit}| < \alpha$ **then return**

8 'Causal direction cannot be determined';

9 **if** $\text{Crit} > \alpha$ **then return** ' \mathbf{X} is the cause of \mathbf{Y} ';

10 **if** $\text{Crit} < -\alpha$ **then return** ' \mathbf{Y} is the cause of \mathbf{X} ';

The parameter $\alpha \in [0, 1]$ is a sensitivity parameter that controls the algorithm's agnosticism. If α is chosen large, then it will return a definite result only in clear cut cases.

If 2G-VecCI.PC is run with a consistent independence test, i.e. with one that recovers conditional independence relations perfectly in the infinite sample limit, it consistently estimates $d(\mathbf{X}|\mathbf{Y})$ and $d(\mathbf{Y}|\mathbf{X})$. Note that this also applies to nonlinear relations, our approach does not rely on functional assumptions. Therefore, the algorithms consistency follows directly from Theorem 2.

Corollary 3. *If the assumptions of Model 1 are satisfied and if at least one of (C1) or (C2) holds, then 2G-VecCI.PC run with a consistent CI test returns the correct causal direction in the infinite sample limit.*

The above result is fully non-parametric, but in practice, under appropriate assumptions, one way to include \mathbf{Y} as a conditioning set for independence tests within \mathbf{X} for computing $d(\mathbf{X}|\mathbf{Y})$, is to regress \mathbf{Y} on \mathbf{X} and run skeleton phase on the residuals of this regression, and vice-versa for $d(\mathbf{Y}|\mathbf{X})$. In particular, if the relationship between groups is assumed linear, this is our method of choice. As an

alternative to 2G-VecCI.PC, it is also possible to run a 'one sided' version that only computes $d(\widehat{\mathbf{X}|\mathbf{Y}})$ (or $d(\widehat{\mathbf{Y}|\mathbf{X}})$) and decides solely based on its sign. This version of the algorithm is computationally less costly but we have found it to be more frail to statistical errors than the full version in practice.

Our second algorithm 2G-VecCI.Full differs from 2G-VecCI.PC in that skeleton phase of the PC algorithm is replaced by independence tests

$$\begin{aligned} X_i \perp\!\!\!\perp X_j \mid \mathbf{X} \setminus \{X_i, X_j\}, \\ X_i \perp\!\!\!\perp X_j \mid \mathbf{X} \setminus \{X_i, X_j\}, \mathbf{Y}, \quad \forall i, j, \end{aligned}$$

to compute edge densities for $d'(\mathbf{X}|\mathbf{Y})$ since we are interested in the (conditional) dependence graph. Again, it is reasonable to regress \mathbf{X} on \mathbf{Y} in many practical settings and to perform independence tests on the residuals. We proceed analogously with \mathbf{X} and \mathbf{Y} exchanged to compute $d'(\mathbf{Y}|\mathbf{X})$. Another remark of practical importance is that 2G-VecCI.Full is also applicable in the context of Model 1, although it measures the edge densities of the *moralized graph* of the causal DAGs, rather than the edge densities of the DAGs itself (see the appendix for the definition of the moralized graph). In this sense, 2G-VecCI.Full is more general than 2G-VecCI.PC. However, if the causal DAG over a group differs dramatically from its moralized graph, 2G-VecCI.PC should be preferred. As an extreme example, consider the case where there is $X_i \in \mathbf{X}$ that is a common child of all other elements of \mathbf{X} and similarly $Y_j \in \mathbf{Y}$ that is a common child of all other elements of \mathbf{Y} . In this situation, the moralized graph over each group would be fully connected (with or without conditioning) and 2G-VecCI.Full would not be suitable.

Experimental Results

Simulated Data

We depict several results for 2G-VecCI.PC and 2G-VecCI.Full for linear and nonlinear (quadratic) models in the main text. For linear models we generate the empirical distributions of \mathbf{X} and $\eta_{\mathbf{Y}}$ by linear SCMs with randomly chosen coefficients and Gaussian noises with randomized variances in the range $[0.5, 2.0]$. We then generate a random $n \times m$ interaction matrix A and set $\mathbf{Y} = A\mathbf{X} + \eta_{\mathbf{Y}}$. Models vary along the following parameters:

- sample size (between 50 and 500);
- group sizes n and m (between 3 and 100);
- edge densities within \mathbf{X} and $\eta_{\mathbf{Y}}$ (between 1% and 90% of all possible edges);
- density of the interaction matrix A (between 1% and 90% of all possible entries non-zero);
- effect size, i.e. size of the entries in A (uniformly randomly drawn from different intervals).

For each parameter choice, 100 random models are generated. In both algorithms, we test for conditional independencies using the partial correlation test at significance level $\tilde{\alpha} = 0.01$. We choose the sensitivity parameter to be

Algorithm 2: 2G-VecCI.Full

Data: two arrays containing samples of \mathbf{X} and \mathbf{Y} , parameter $\alpha \in [0, 1]$.

Result: variable with values ' \mathbf{X} is the cause of \mathbf{Y} ', or ' \mathbf{Y} is the cause of \mathbf{X} ', or 'Causal direction cannot be determined'.

```

1 for  $X_i \neq X_j \in \mathbf{X}$  do
2   test  $X_i \perp\!\!\!\perp X_j \mid \mathbf{X} \setminus \{X_i, X_j\}$ ;
3   if dependent then edgeCount( $\mathcal{G}'_{\mathbf{X}}$ ) + = 1;
4   test  $X_i \perp\!\!\!\perp X_j \mid \mathbf{X} \setminus \{X_i, X_j\}, \mathbf{Y}$ ;
5   if dependent then edgeCount( $\mathcal{G}'_{\mathbf{X}|\mathbf{Y}}$ ) + = 1;
6 end
7 Compute edgeDens( $\widehat{\mathcal{G}'_{\mathbf{X}}}$ ) =  $\frac{\text{edgeCount}(\mathcal{G}'_{\mathbf{X}})}{\text{edgeMax}}$ ;
8 Compute edgeDens( $\widehat{\mathcal{G}'_{\mathbf{X}|\mathbf{Y}}}$ ) =  $\frac{\text{edgeCount}(\mathcal{G}'_{\mathbf{X}|\mathbf{Y}})}{\text{edgeMax}}$ ;
9 Compute
    $d'(\widehat{\mathbf{X}|\mathbf{Y}}) = \text{edgeDens}(\widehat{\mathcal{G}'_{\mathbf{X}|\mathbf{Y}}}) - \text{edgeDens}(\widehat{\mathcal{G}'_{\mathbf{X}}})$ ;
10 Repeat 1-9 with  $\mathbf{X}, \mathbf{Y}$  exchanged to get  $d'(\widehat{\mathbf{Y}|\mathbf{X}})$ ;
11 Compute Crit =  $d'(\widehat{\mathbf{X}|\mathbf{Y}}) - d'(\widehat{\mathbf{Y}|\mathbf{X}})$ ;
12 Repeat steps 7-10 of 2G-VecCI.PC.
```

$\alpha = 0.01$ if not specified differently. We also compare our algorithms to existing approaches, see below. Computations were done on BullSequana XH2000 with AMD 7763 CPUs. We observe the following:

- Generally speaking, 2G-VecCI.Full outperforms 2G-VecCI.PC on our simulated data even though the data is generated by a causal model.
- Performance increases with increasing effect size and increasing interaction density and decreases with increasing edge densities within variable groups.
- More precisely, both algorithms perform best when dependencies within both variable groups are sparse to medium sparse ($< 10\%$ of all possible edges). Nevertheless, for high sample sizes (e.g. 500), the correct causal direction is still inferred reliably for large groups (100 variables per group) even if variable groups are quite dense (30% of all possible edges).
- For the algorithms to perform well, the interaction matrix should not be too sparse (i.e. $< 10\%$ of non-zero entries) and effect sizes should not be too small < 0.1 . This type of 'weak mechanism problem' is typical for constraint-based causal inference algorithms in general as one operates close to the non-faithful regime.

All code is available at <https://github.com/JonasChoice/2GVecCI>.

Complexity of 2G-VecCI.PC and 2G-VecCI.Full As in other PC-based methods, the number of CI tests run by 2G-VecCI.PC is data dependent and may increase exponentially in the worst case, i.e. for high group and interaction densities where separating sets need to be large. 2G-VecCI.Full on the other hand runs $2(n^2 + m^2)$ CI tests where n, m are

the group sizes, independent of the specifics of the data (but with large conditioning sets). Therefore, if groups and interactions are assumed very sparse, 2G-VecCI.PC may be less costly than 2G-VecCI.Full while 2G-VecCI.Full should be preferred over 2G-VecCI.PC when groups are assumed to be reasonably dense and in practice, we have typically found 2G-VecCI.Full to perform significantly faster. For an in-depth discussion on computational cost of the PC-algorithm, we refer the reader to Kalisch and Bühlmann (2007) and Le et al. (2019).

To test our methods for nonlinear interactions, ground truth data is generated as in the linear case except that we use the model $\mathbf{Y} = \mathbf{A}\mathbf{X}^2 + \eta_{\mathbf{Y}}$, where \mathbf{X}^2 is shorthand for the vector of squared entries of \mathbf{X} .

Here, our methods are affected more strongly by increasing group sizes as nonlinear CI-tests tend to be much slower than tests for partial correlation. Nevertheless, 2G-VecCI.Full run with the Gaussian Process distance correlation independence test still finds the correct causal direction significantly better than a random choice would for groups of size 15 (with 100 samples) and 25 (with 200 samples), see Figure 4. At present, we did not implement non-linear interactions for 2G-VecCI.PC. For large groups, performance could potentially be sped up by reducing dimensions locally. For instance, if the data is structured spatially, small subregions could be averaged to scalar variables to obtain a coarser variable group. The approach of Chalupka, Eberhardt, and Perona (2017) might be helpful here and combining it with our work might be an avenue for future research.

Comparison to Other Methods

Two established methods for inferring causal relations between variable groups are multivariate LiNGaM (Shimizu et al. 2006) (Entner and Hoyer 2012) and the Trace Method (Janzing et al. 2009) (Zscheischler, Janzing, and Zhang 2012). In contrast to our methods, both of these techniques assume interactions to be linear and LiNGaM additionally requires non-Gaussian noise to be applicable. In simulations with linearly interacting groups and Gaussian noise, the trace method and both versions of 2G-VecCI perform comparably well, see Figures 2, 3 and 10, although the trace method is significantly faster. We also analysed the performance of our method against a baseline PC algorithm (Vanilla-PC) by treating each component in the vector-valued variables as a separate node and counting the arrow directions from (nodes belonging to) one group to the other. In general our methods outperform Vanilla PC except when groups are small and the interaction matrix is sparse, see Figures 2 and 3 as well as Section 12 and Figures 8, 9 in the supplement available at <https://arxiv.org/abs/2209.14283>.

A Real-World Example

In order to test our algorithms in a typical causal discovery setting in Earth sciences, we consider surface temperatures over the ENSO 3.4 region and over British Columbia (denoted by BCT) from 1948-2021. We consider this example because a causal effect of temperatures in the tropical pacific on those in North America is established in climate

science. Additionally, Runge et al. (2019) used this example to test the PCMCI algorithm and found a causal link $\text{Nino}_{t-2} \rightarrow \text{BCT}_t$, i.e. at a time-lag of two months.

The data is first de-seasonalized and any long-term trend is removed from the raw time series with a Gaussian kernel smoothing mean with a bandwidth of $\sigma = 120$ months as in Runge et al. (2019). Furthermore, in order to mitigate auto-correlation in time, we consider means of ENSO temperature anomalies from October to December and means of BCT anomalies from January to March, leading to 73 samples each for the two groups \mathbf{X} and \mathbf{Y} . To get more robust results, we consider different coarse grainings of the two regions, for instance every second grid-box for ENSO and every third grid-box for BCT anomalies. This has the additional effect of reducing group sizes in comparison to the sample size. We moreover reject coarse grainings that correspond to a difference of more than 10 grid-boxes between the two groups, in order to avoid a bias due to region sizes. Algorithm 2G-VecCI.Full with the partial correlation CI-test then computes $\text{Crit} = d'(\mathbf{X}|\mathbf{Y}) - d'(\mathbf{Y}|\mathbf{X})$, see (3) and (4). We find the mean and the standard deviation of the Crit values to be $\mu = 0.031$ and $\sigma = 0.063$, respectively, indicating a causal effect of ENSO on BCT. If the sensitivity parameter α is chosen to be 0.01, then the fraction of correct inferences is 0.59 and the fraction of wrong inferences is 0.27. Thus 2G-VecCI.Full deduces the correct casual direction $\text{Nino} \rightarrow \text{BCT}$ with high probability. Algorithm 2G-VecCI.PC with partial correlation and $\alpha = 0.01$ on the other hand yields $\mu = 0.001$ and $\sigma = 0.019$, respectively and is thus indecisive. We attribute the low detection power of 2G-VecCI.PC to the reduced effect size arising from unobserved confounders and insufficiently mitigated auto-correlation in this simplified study. NCEP-NCAR Reanalysis 1 data was provided by NOAA PSL, Boulder, Colorado, USA, from their website at <https://psl.noaa.gov>, see Kalnay et al. (1996).

Discussion and Outlook

We have introduced two new algorithms to infer causal direction between two potentially high-dimensional groups of variables and provided a theoretical analysis of groupwise causal inference in the DAG-based causality framework.

The **main strengths** of our work are that it contains a novel identifiability result for the unidirectional causal vector model and practical implementations using density estimates of the vector-valued variables. It is comparable to the trace method in the high sample regime when interactions are linear and better when the group sizes are large and interactions are sparse. Moreover our methods are also able to deal with non-linear interactions. Currently, the **main weaknesses** are that we work with an i.i.d. assumption on the data samples and that unobserved confounding variables are not addressed. Furthermore, our algorithms have slower runtime than the trace method or dimension reduction techniques.

In **future work**, we plan to extend this work to the setting of multiple variable groups similar to Parviainen and Kaski (2017) as well as to use partial dimension reduction techniques. We also plan to relax the i.i.d. assumption to better deal with autocorrelations following Runge et al. (2019).

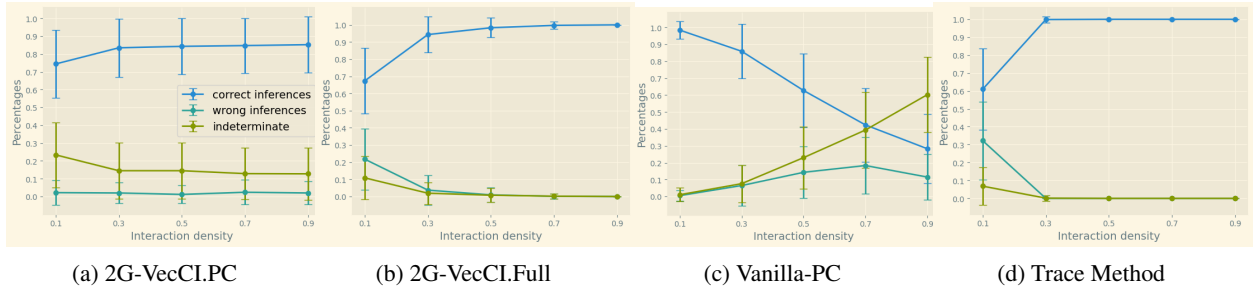


Figure 2: Performance of 2G-VecCI.PC, 2G-VecCI.Full, Vanilla-PC and the trace method for groups of size 30 with 100 available samples and linear interactions. Performance is shown along increasing density of the interaction matrix A , and percentages are averaged across different parameters for internal group densities (1%, 5%, 10%, 30% of all possible edges present). Non-zero entries of A are drawn uniformly randomly from $[-0.7, 0.7]$ and 100 random models are run per parameter combination. All approaches except Vanilla PC recover the correct causal direction well. Vanilla-PC is challenged by dense interaction matrices as this increases the overall density of the causal graph over all microvariables. We set the sensitivity parameter of Vanilla-PC to 10^{-4} to ensure that the lack of performance is not due to an overly conservative choice.

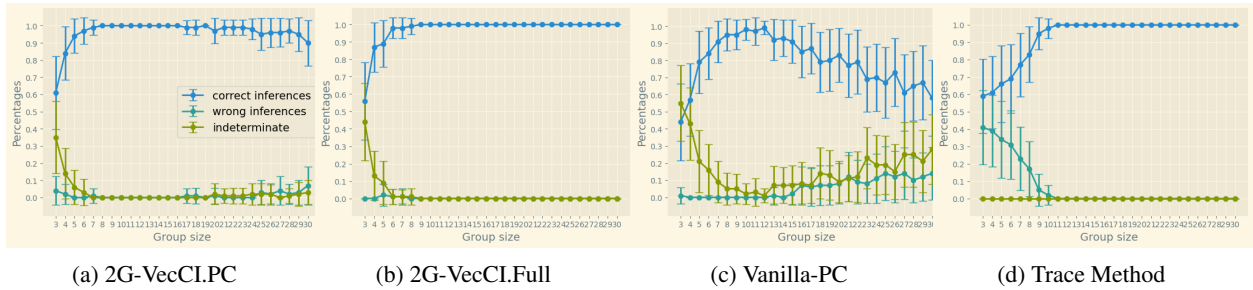


Figure 3: Performance of 2G-VecCI.PC, 2G-VecCI.Full, Vanilla-PC and the Trace Method for groups different sizes and low densities (10%). The density of the interaction matrix A are set to 50%. Non-zero entries of A are drawn uniformly randomly from $[-0.7, 0.7]$ and 100 random models are run per parameter combination with 100 samples each. We set the sensitivity parameter of both 2G-VecCI.PC and Vanilla-PC to 10^{-4} . Vanilla PC performs well for small groups but decreases in performance as group sizes grow.

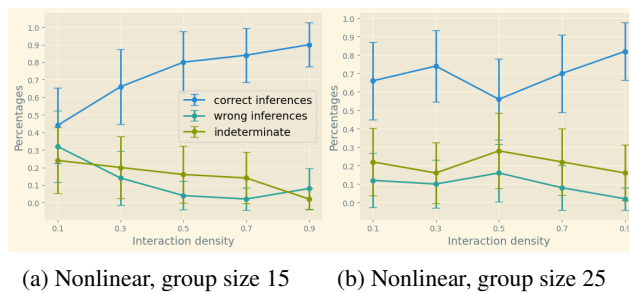


Figure 4: Performance of 2G-VecCI.Full for quadratic interactions of size 15 (left) and 25 (right) with 200 samples. Performance is shown along increasing density of the interaction matrix A . Non-zero entries of A are drawn uniformly randomly from $[-0.7, 0.7]$, groups are assumed medium dense on the left (≈ 5 connections per variable) and sparse on the right (≈ 3 connections per variable). 50 random models are run per choice of parameters. In both cases 2G-VecCI.Full finds the correct causal direction better than chance would expect when A is extremely sparse.

Acknowledgements

This research has been supported by grant no. 948112 *Causal Earth* of the European Research Council (ERC). This work used resources of the Deutsches Klimarechenzentrum (DKRZ) granted by its Scientific Steering Committee (WLA) under project ID bd1083. We thank the reviewers for their insightful remarks and suggestions.

References

- Chalupka, K.; Bischoff, T.; Perona, P.; and Eberhardt, F. 2016. Unsupervised Discovery of El Nino Using Causal Feature Learning on Microlevel Climate Data. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence, UAI'16*, 72–81. Arlington, Virginia, USA: AUAI Press. ISBN 9780996643115.
- Chalupka, K.; Eberhardt, F.; and Perona, P. 2016. Multi-Level Cause-Effect Systems. In Gretton, A.; and Robert, C. C., eds., *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, 361–369. Cadiz, Spain: PMLR.
- Chalupka, K.; Eberhardt, F.; and Perona, P. 2017. Causal feature learning: an overview. *Behaviormetrika*, 44(1): 137–164.
- Entner, D.; and Hoyer, P. O. 2012. Estimating a Causal Order among Groups of Variables in Linear Models. In Villa, A. E. P.; Duch, W.; Érdi, P.; Masulli, F.; and Palm, G., eds., *Artificial Neural Networks and Machine Learning – ICANN 2012*, 84–91. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-642-33266-1.
- Janzing, D.; Hoyer, P.; Schölkopf, B.; Fürnkranz, J.; and Joachims, T. 2009. Telling cause from effect based on high-dimensional observations. *Proceedings of the 27th International Conference on Machine Learning (ICML 2010)*, 479–486 (2010).
- Kalisch, M.; and Bühlmann, P. 2007. Estimating High-Dimensional Directed Acyclic Graphs with the PC-Algorithm. *Journal of Machine Learning Research*, 8(22): 613–636.
- Kalnay, E.; Kanamitsu, M.; Kistler, R.; Collins, W.; Deaven, D.; Gandin, L.; Iredell, M.; Saha, S.; White, G.; Woollen, J.; Zhu, Y.; Chelliah, M.; Ebisuzaki, W.; Higgins, W.; Janowiak, J.; Mo, K. C.; Ropelewski, C.; Wang, J.; Leetmaa, A.; Reynolds, R.; Jenne, R.; and Joseph, D. 1996. The NCEP/NCAR 40-Year Reanalysis Project. *Bulletin of the American Meteorological Society*, 77(3): 437 – 472.
- Le, T. D.; Hoang, T.; Li, J.; Liu, L.; Liu, H.; and Hu, S. 2019. A Fast PC Algorithm for High Dimensional Causal Discovery with Multi-Core PCs. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 16(5): 1483–1495.
- Parviainen, P.; and Kaski, S. 2017. Learning structures of Bayesian networks for variable groups. *International Journal of Approximate Reasoning*, 88: 110–127.
- Pearl, J. 2009. *Causality: Models, Reasoning and Inference*. USA: Cambridge University Press, 2nd edition. ISBN 052189560X.
- Peters, J.; Janzing, D.; and Schölkopf, B. 2017. *Elements of Causal Inference - Foundations and Learning Algorithms*. Adaptive Computation and Machine Learning Series. Cambridge, MA, USA: The MIT Press.
- Rubenstein*, P. K.; Weichwald*, S.; Bongers, S.; Mooij, J. M.; Janzing, D.; Grosse-Wentrup, M.; and Schölkopf, B. 2017. Causal Consistency of Structural Equation Models. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI)*, ID 11. *equal contribution.
- Runge, J.; Nowack, P.; Kretschmer, M.; Flaxman, S.; and Sejdinovic, D. 2019. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, 5(11): eaau4996.
- Runge, J.; Petoukhov, V.; Donges, J. F.; Hlinka, J.; Jajcay, N.; Vejmelka, M.; Hartman, D.; Marwan, N.; Paluš, M.; and Kurths, J. 2015. Identifying causal gateways and mediators in complex spatio-temporal systems. *Nature communications*, 6(1): 1–10.
- Shimizu, S.; Hoyer, P. O.; Hyvärinen, A.; and Kerminen, A. 2006. A Linear Non-Gaussian Acyclic Model for Causal Discovery. *J. Mach. Learn. Res.*, 7: 2003–2030.
- Song, H.-R.; Fuentes, M.; and Ghosh, S. 2008. A comparative study of Gaussian geostatistical models and Gaussian Markov random field models. *Journal of Multivariate Analysis*, 99(8): 1681–1697.
- Spirites, P.; Glymour, C.; and Scheines, R. 2000. *Causation, Prediction, and Search*. MIT press, 2nd edition.
- Vaccaro, A.; Emile-Geay, J.; Guillot, D.; Verna, R.; Morice, C.; Kennedy, J.; and Rajaratnam, B. 2021. Climate Field Completion via Markov Random Fields: Application to the HadCRUT4.6 Temperature Dataset. *Journal of Climate*, 34(10): 4169 – 4188.
- Zscheischler, J.; Janzing, D.; and Zhang, K. 2012. Testing whether linear equations are causal: A free probability theory approach. *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence, UAI 2011*.