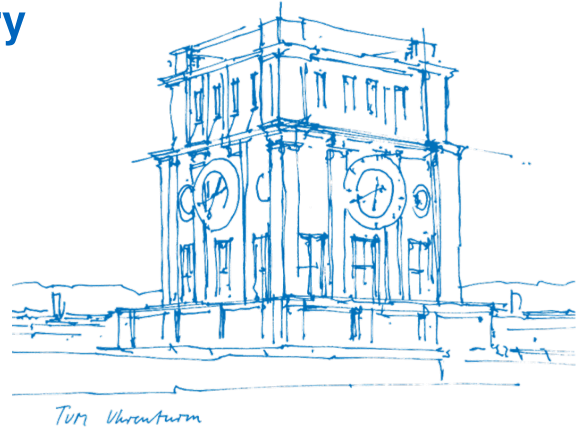
Nonlinear Causal Discovery for Grouped Data

Konstantin Göbler

TUM School of Computation, Information and Technology
Chair of Mathematical Statistics
Technical University of Munich

October 29, 2025



Joint work with:

Mathias Drton

TUM School of CIT

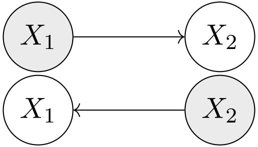
Tobias Windisch

University of Applied Sciences
Kempten

Outline

What are we talking about?

- This talk:



- (Later) with $p > 2$ random variables present.

- (Later) with $\mathbf{X}_i, i \in [p]$ being random vectors (or groups) rather than scalar random variables.

Identifiability

- Without any further assumptions, direction of the edge can't be inferred from observations from X_1 and X_2, even in the infinite data limit [Spirtes et al., 1993].

Identifiability

- Without any further assumptions, direction of the edge can't be inferred from observations from X_1 and X_2, even in the infinite data limit [Spirtes et al., 1993].

- Performing an intervention in this setting will allow us to orient the edge [Pearl, 2009].

Identifiability

- Without any further assumptions, direction of the edge can't be inferred from observations from X_1 and X_2, even in the infinite data limit [Spirtes et al., 1993].

- Performing an intervention in this setting will allow us to orient the edge [Pearl, 2009].

- However, interventions might be costly, unethical, or infeasible in practice.

Identifiability

- Without any further assumptions, direction of the edge can't be inferred from observations from X_1 and X_2, even in the infinite data limit [Spirtes et al., 1993].

- Performing an intervention in this setting will allow us to orient the edge [Pearl, 2009].

- However, interventions might be costly, unethical, or infeasible in practice.

- Certain model classes, it turns out, allow us to orient the causal edge, without interventional data.

Outline

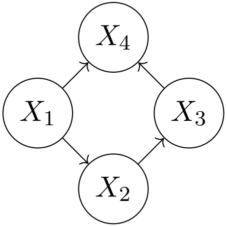
Structural equation models

A structural equation model (SEM) [Bollen, 1989] is a tuple $(\mathcal{S}, P(N))$, where
$\mathcal{S} = (S_1, \ldots, S_p)$ is a collection of p equations

$$S_k : \quad X_k = f_k(X_{pa(k)}, N_k), \quad k \in [p],$$

and $P(N) = P(N_1, \ldots, N_p)$ is the joint product distribution of exogenous noise terms.

$$\begin{aligned} X_1 &:= f_1(N_1) \\ X_2 &:= f_2(X_1, N_2) \\ X_3 &:= f_3(X_2, N_3) \\ X_4 &:= f_4(X_1, X_3, N_4), \end{aligned}$$



$N_1, \ldots, N_4$ jointly independent

Outline

Identifiability

- Let us make the notion of identifiability more precise.

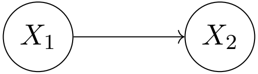
Identifiability

- Let us make the notion of identifiability more precise.

- Consider the following SEM over two (not necessarily scalar) random variables X_1 and X_2,

$$X_1 = N_1, \quad X_2 = f_2(X_1, N_2),$$

with $N_1 \perp\!\!\!\perp N_2$ and the following DAG



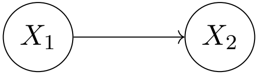
Identifiability

- Let us make the notion of identifiability more precise.

- Consider the following SEM over two (not necessarily scalar) random variables X_1 and X_2,

$$X_1 = N_1, \quad X_2 = f_2(X_1, N_2),$$

with $N_1 \perp\!\!\!\perp N_2$ and the following DAG



- We know of course that the DAG with the edge reversed lives in the same Markov equivalence class.

- Consider the following SEM

$$X_2 = f(X_1, N_2), \quad \text{with } X_1 \perp\!\!\!\perp N_2.$$

- Consider the following SEM

$$X_2 = f(X_1, N_2), \quad \text{with } X_1 \perp\!\!\!\perp N_2.$$

- Clearly, if the reversed causal direction were valid, one could write

$$X_1 = \tilde{f}(X_2, N_1), \quad \text{with } X_2 \perp\!\!\!\perp N_1.$$

  with $f, \tilde{f} \in \mathcal{F}$ some general functional class.

- Consider the following SEM

$$X_2 = f(X_1, N_2), \quad \text{with } X_1 \perp\!\!\!\perp N_2.$$

- Clearly, if the reversed causal direction were valid, one could write

$$X_1 = \tilde{f}(X_2, N_1), \quad \text{with } X_2 \perp\!\!\!\perp N_1.$$

with $f, \tilde{f} \in \mathcal{F}$ some general functional class.

- It turns out, without restrictions on the functional class $\mathcal{F}$, Hyvärinen and Pajunen [1999] show that there always exists a suitable function $\tilde{f} \in \mathcal{F}$ ensuring $X_2 \perp\!\!\!\perp N_1$.

- Consider the following SEM

$$X_2 = f(X_1, N_2), \quad \text{with } X_1 \perp\!\!\!\perp N_2.$$

- Clearly, if the reversed causal direction were valid, one could write

$$X_1 = \tilde{f}(X_2, N_1), \quad \text{with } X_2 \perp\!\!\!\perp N_1.$$

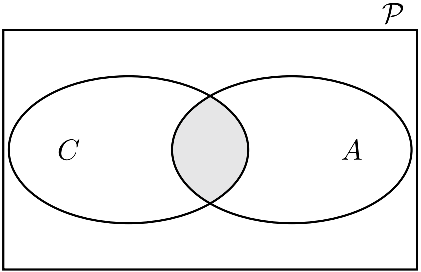
  with $f, \tilde{f} \in \mathcal{F}$ some general functional class.

- It turns out, without restrictions on the functional class $\mathcal{F}$, Hyvärinen and Pajunen [1999] show that there always exists a suitable function $\tilde{f} \in \mathcal{F}$ ensuring $X_2 \perp\!\!\!\perp N_1$.

- The absence of constraints on $\mathcal{F}$ renders the SEM symmetric with respect to variables X_1 and X_2.

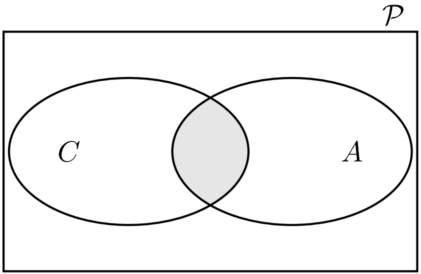
Identifiability

- Let's depict the joint distributions that may be generated from the causal SEM (C) and the anticausal SEM (A) inside the set of all possible joint distributions $\mathcal{P}$.



Identifiability

- Let's depict the joint distributions that may be generated from the causal SEM (C) and the anticausal SEM (A) inside the set of all possible joint distributions $\mathcal{P}$.

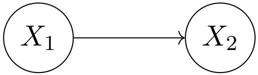


- Identifiability: Size of the intersection $C \cap A$. If C and A were to contain almost the same set of joint distributions, we would regard the model class as non-identifiable.
- Conversely, if the intersection is very small, we would regard the model class as identifiable.

One specific restriction on $\mathcal{F}$

- Let's motivate one model class that enables us to orient the edge.

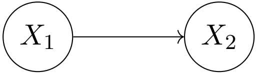
$$X_1 = N_1$$
$$X_2 = f_2(X_1) + N_2$$



One specific restriction on $\mathcal{F}$

- Let's motivate one model class that enables us to orient the edge.

$$\begin{aligned} X_1 &= N_1 \\ X_2 &= f_2(X_1) + N_2 \end{aligned}$$

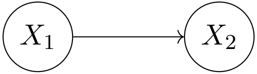


- where f_2 is assumed to be a three-times differentiable nonlinear function.

One specific restriction on $\mathcal{F}$

- Let's motivate one model class that enables us to orient the edge.

$$X_1 = N_1$$
$$X_2 = f_2(X_1) + N_2$$



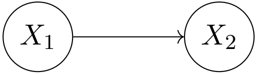
- where f_2 is assumed to be a three-times differentiable nonlinear function.

Why is this model class identiable?

One specific restriction on $\mathcal{F}$

- Let's motivate one model class that enables us to orient the edge.

$$X_1 = N_1$$
$$X_2 = f_2(X_1) + N_2$$



- where f_2 is assumed to be a three-times differentiable nonlinear function.

Why is this model class identiable?

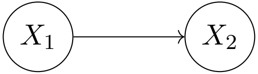
- **Fact:** $\mathbb{E}[X_2 \mid X_1]$ *"best"* predicts X_2 as a function of X_1.
- By construction:

$$\text{Corr}(X_2 - \mathbb{E}[X_2 \mid X_1], X_1) = \text{Corr}(X_1 - \mathbb{E}[X_1 \mid X_2], X_2) = 0$$

Nonlinear Additive noise models (ANMs)

- Let's motivate the model class that enables us to orient the edge below.

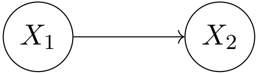
$$\begin{aligned} X_1 &= N_1 \\ X_2 &= f_2(X_1) + N_2 \end{aligned}$$



Nonlinear Additive noise models (ANMs)

- Let's motivate the model class that enables us to orient the edge below.

$$X_1 = N_1$$
$$X_2 = f_2(X_1) + N_2$$



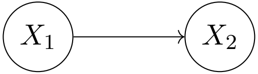
- Regression the right way around: $X_2 - \mathbb{E}[X_2 \mid X_1] = X_2 - f_2(X_1) = N_2$, then

$$N_2 \perp\!\!\!\perp X_1$$

Nonlinear Additive noise models (ANMs)

- Let's motivate the model class that enables us to orient the edge below.

$$\begin{aligned} X_1 &= N_1 \\ X_2 &= f_2(X_1) + N_2 \end{aligned}$$



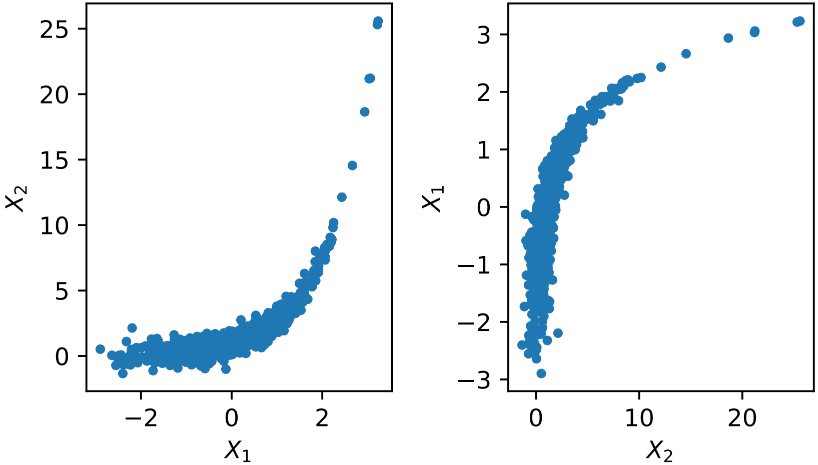
- Regression the right way around: $X_2 - \mathbb{E}[X_2 \mid X_1] = X_2 - f_2(X_1) = N_2$, then

$$N_2 \perp\!\!\!\perp X_1$$

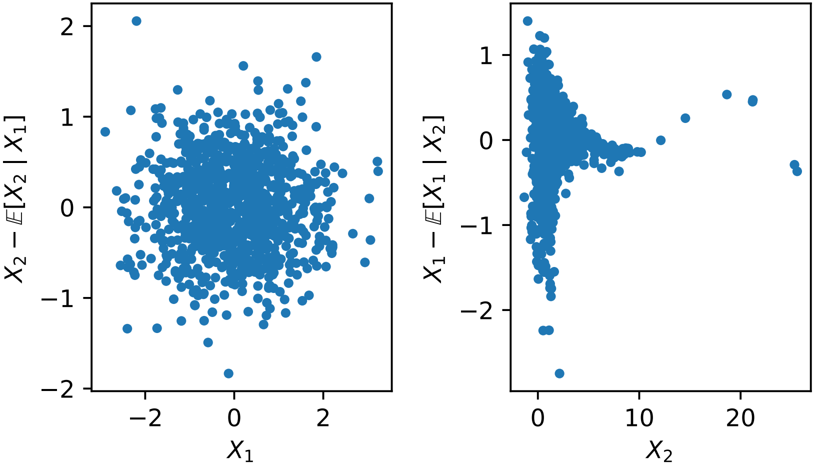
- Regression the wrong way around: for some general nonlinear f_2, it holds that

$$X_1 - \mathbb{E}[X_1 \mid X_2] \not\perp\!\!\!\perp X_2, \quad \text{(but uncorrelated).}$$

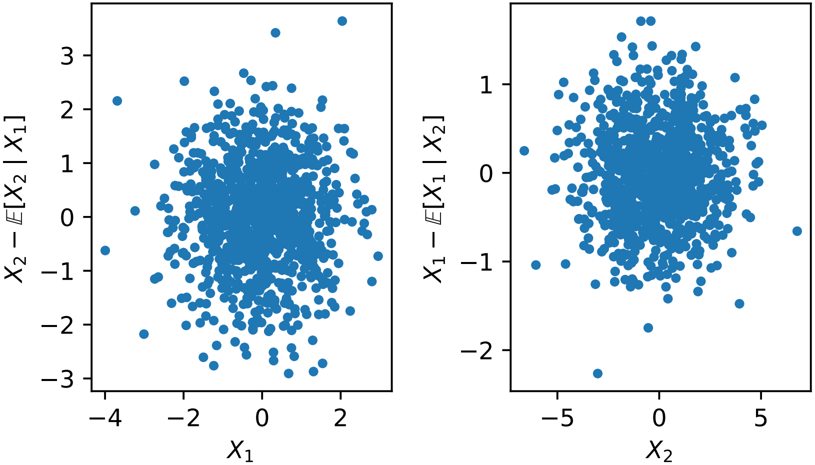
ANM: $X_1 = N_1, X_2 = \exp(X_1) + N_2$ and $N_i \sim N(0,1), i = 1, 2$



Comparing forward and backward model



Non-intentifiable: $f_2(z) = a * z + b$ and $N_i \sim N(0, 1), i = 1, 2$



Exact condition for non-identifiability

- It turns out that we can characterize the intersection $C \cap A$ exactly.

Exact condition for non-identifiability

- It turns out that we can characterize the intersection $C \cap A$ exactly.

- First shown by Hoyer et al. [2008], a bivariate SEM is identifiable if the triple $(f_j, P(X_i), P(N_j))$ for $i, j \in \{1, 2\}$ does not solve the following differential equation

$$\xi''' = \xi'' \left(-\frac{\nu''' f'}{\nu''} + \frac{f''}{f'} \right) - 2\nu'' f'' f' + \nu' f''' + \frac{\nu''' \nu' f'' f'}{\nu''} - \frac{\nu' (f'')^2}{f'},$$

  where $f := f_j, \xi := \log p_{X_i}, \nu := \log p_{N_j}$.

Exact condition for non-identifiability

- It turns out that we can characterize the intersection $C \cap A$ exactly.

- First shown by Hoyer et al. [2008], a bivariate SEM is identifiable if the triple $(f_j, P(X_i), P(N_j))$ for $i, j \in \{1, 2\}$ does not solve the following differential equation

$$\xi''' = \xi'' \left(-\frac{\nu''' f'}{\nu''} + \frac{f''}{f'} \right) - 2\nu'' f'' f' + \nu' f''' + \frac{\nu''' \nu' f'' f'}{\nu''} - \frac{\nu' (f'')^2}{f'},$$

  where $f := f_j$, $\xi := \log p_{X_i}$, $\nu := \log p_{N_j}$.
- The differential equation for ξ has a 3-dimensional space of solutions, while a priori, the space of all possible log-marginals is infinite dimensional.

Exact condition for non-identifiability

- It turns out that we can characterize the intersection $C \cap A$ exactly.

- First shown by Hoyer et al. [2008], a bivariate SEM is identifiable if the triple $(f_j, P(X_i), P(N_j))$ for $i, j \in \{1, 2\}$ does not solve the following differential equation

$$\xi''' = \xi'' \left(-\frac{\nu''' f'}{\nu''} + \frac{f''}{f'} \right) - 2\nu'' f'' f' + \nu' f''' + \frac{\nu''' \nu' f'' f'}{\nu''} - \frac{\nu' (f'')^2}{f'},$$

  where $f := f_j$, $\xi := \log p_{X_i}$, $\nu := \log p_{N_j}$.

- The differential equation for ξ has a 3-dimensional space of solutions, while a priori, the space of all possible log-marginals is infinite dimensional.

- Thus, in generic cases, a backward model does not exist.

Outline

Grouped additive noise models (GANMs)

- Now, suppose that $\mathbf{X}_1 = (X_1^1, \ldots, X_{d_1}^1)$ and $\mathbf{X}_2 = (X_1^2, \ldots, X_{d_2}^2)$ are random vectors with positive density w.r.t. the Lesbesgue measure, respectively.

Grouped additive noise models (GANMs)

- Now, suppose that $\mathbf{X}_1 = (X_1^1, \ldots, X_{d_1}^1)$ and $\mathbf{X}_2 = (X_1^2, \ldots, X_{d_2}^2)$ are random vectors with positive density w.r.t. the Lesbesgue measure, respectively.

- Consider the GANM:

$$\mathbf{X}_1 = \mathbf{N}_1, \quad \mathbf{X}_2 = f_2(\mathbf{X}_1) + \mathbf{N}_2, \quad \text{with } \mathbf{N}_1 \perp\!\!\!\perp \mathbf{N}_2,$$

  with $f_2 \in \mathcal{F} \subseteq C^3(\mathbb{R}^{d_1}, \mathbb{R}^{d_2})$

Grouped additive noise models (GANMs)

- Now, suppose that $\mathbf{X}_1 = (X_1^1, \ldots, X_{d_1}^1)$ and $\mathbf{X}_2 = (X_1^2, \ldots, X_{d_2}^2)$ are random vectors with positive density w.r.t. the Lesbesgue measure, respectively.

- Consider the GANM:

$$\mathbf{X}_1 = \mathbf{N}_1, \quad \mathbf{X}_2 = f_2(\mathbf{X}_1) + \mathbf{N}_2, \quad \text{with } \mathbf{N}_1 \perp\!\!\!\perp \mathbf{N}_2,$$

  with $f_2 \in \mathcal{F} \subseteq C^3(\mathbb{R}^{d_1}, \mathbb{R}^{d_2})$

- The joint density has the following form

$$p_{\mathbf{X}_1,\mathbf{X}_2}(\mathbf{x}_1, \mathbf{x}_2) = p_{\mathbf{X}_1}(\mathbf{x}_1) p_{\mathbf{N}_2}(\mathbf{x}_2 - f_2(\mathbf{x}_1)).$$

GANMs continued

- Suppose there exists a backward model of the same form

$$p_{\mathbf{X}_1,\mathbf{X}_2}(\mathbf{x}_1, \mathbf{x}_2) = p_{\mathbf{X}_2}(\mathbf{x}_2) p_{\mathbf{N}_1}(\mathbf{x}_1 - f_1(\mathbf{x}_2)).$$

  Define

$$\pi_1(\mathbf{x}_1, \mathbf{x}_2) := \nu(\mathbf{x}_2 - f_2(\mathbf{x}_1)) + \xi(\mathbf{x}_1) \tag{1}$$

  and

$$\pi_2(\mathbf{x}_1, \mathbf{x}_2) := \tilde{\nu}(\mathbf{x}_1 - f_1(\mathbf{x}_2)) + \eta(\mathbf{x}_2), \tag{2}$$

  where $\nu := \log p_{\mathbf{N}_2}$, $\tilde{\nu} := \log p_{\mathbf{N}_1}$, $\xi := \log p_{\mathbf{X}_1}$, and $\eta := \log p_{\mathbf{X}_2}$.

- Clearly, we have that $\pi_1(\mathbf{x}_1, \mathbf{x}_2) = \pi_2(\mathbf{x}_1, \mathbf{x}_2) = \log p_{\mathbf{X}_1,\mathbf{X}_2}(\mathbf{x}_1, \mathbf{x}_2)$.

Differential Equation

$$D_{\mathbf{x}_1}\mathbf{H}_\xi(\mathbf{x}_1)(D_{\mathbf{x}_1\mathbf{x}_1}\pi_1)^{-1}D_{\mathbf{x}_1\mathbf{x}_2}\pi_1 = D_{\mathbf{x}_1}D_{\mathbf{x}_1\mathbf{x}_2}\pi_1\Big[D_{\mathbf{x}_1}\left(\mathbf{H}_{f_2}(\mathbf{x}_1)[\nabla\nu(\mathbf{u})]\right) - D_{\mathbf{x}_1}\left(\mathbf{J}_{f_2}(\mathbf{x}_1)^\top\mathbf{H}_\nu(\mathbf{u})\mathbf{J}_{f_2}(\mathbf{x}_1)\right)\Big] (D_{\mathbf{x}_1\mathbf{x}_1}\pi_1)^{-1}D_{\mathbf{x}_1\mathbf{x}_2}\pi_1$$

where $\mathbf{H}_\xi(\mathbf{x}_1) \in \mathbb{R}^{d_{x_1}\times d_{x_1}}$, $\mathbf{J}_{f_2}(\mathbf{x}_1) \in \mathbb{R}^{d_{x_2}\times d_{x_1}}$, $\mathbf{H}_\nu(\mathbf{u}) \in \mathbb{R}^{d_{x_2}\times d_{x_2}}$, and the Hessian $\mathbf{H}_{f_2} \in \mathbb{R}^{d_{x_2}\times d_{x_1}\times d_{x_1}}$ is a third-order tensor. The remaining second order derivatives of the log marginal ξ are contained in the expression for $D_{\mathbf{x}_1\mathbf{x}_1}\pi_1$.

- Interpretation: directional projection of $D_{\mathbf{x}_1}\mathbf{H}_\xi(\mathbf{x}_1)$ onto the directions defined by the columns of the matrix $(D_{\mathbf{x}_1\mathbf{x}_1}\pi_1)^{-1}D_{\mathbf{x}_1\mathbf{x}_2}\pi_1$. The dimensions d_{x_1} and d_{x_2} determine the range of the resulting tensor contraction.

Outline

More than two variables

- Imposing some mild technical conditions, we can use this bivariate result to recursively hold fix all but two variables and the corresponding conditional distribution to extend these results to more than two variables [Peters et al., 2014].

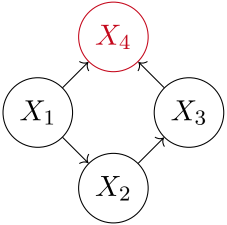
More than two variables

- Imposing some mild technical conditions, we can use this bivariate result to recursively hold fix all but two variables and the corresponding conditional distribution to extend these results to more than two variables [Peters et al., 2014].
- Needed: Each variable's noise term is independent of its non-descendants.

More than two variables

- Imposing some mild technical conditions, we can use this bivariate result to recursively hold fix all but two variables and the corresponding conditional distribution to extend these results to more than two variables [Peters et al., 2014].
- Needed: Each variable's noise term is independent of its non-descendants.

$$
\begin{aligned}
X_1 &:= f_1(N_1)\\
X_2 &:= f_2(X_1, N_2)\\
X_3 &:= f_3(X_2, N_3)\\
X_4 &:= f_4(X_1, X_3, N_4),
\end{aligned}
$$



$N_1, \ldots, N_4$ jointly independent

$$nd(X_4) = \{X_1, X_2, X_3\} = \{f_1(N_1), f_2(X_1, N_2), f_3(X_2, N_3)\}$$

$N_4 \perp\!\!\!\perp X \setminus X_4$

Regression with subsequent independence test (RESIT)

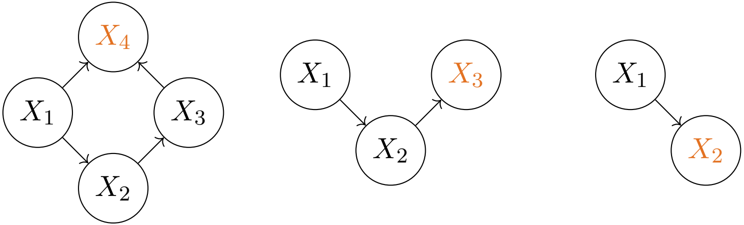
Cycle repeatedly through the following steps [Peters et al., 2014]:

1. Take the current data, and train regression models for each variable onto all other variables i.e. $reg(X_i$ on to $X \setminus X_i)$.
2. Predict and obtain estimates for the residuals (additivity assumption) $\hat{R}_i = X_i - \hat{X}_i$
3. Find the residual that is most independent from all other variables (vector independence test), and remove it from the dataset.
4. Prepend the removed variable to the causal ordering.

Regression with subsequent independence test (RESIT)

Cycle repeatedly through the following steps [Peters et al., 2014]:

1. Take the current data, and train regression models for each variable onto all other variables i.e. $reg(X_i$ on to $X \setminus X_i)$.
2. Predict and obtain estimates for the residuals (additivity assumption) $\hat{R}_i = X_i - \hat{X}_i$
3. Find the residual that is most independent from all other variables (vector independence test), and remove it from the dataset.
4. Prepend the removed variable to the causal ordering.



$$\pi = \{X_1, X_2, X_3, X_4\}$$

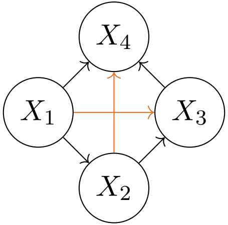
Pruning edges

Given a valid causal ordering, finding the true DAG boils down to a model/feature selection problem:

Pruning edges

Given a valid causal ordering, finding the true DAG boils down to a model/feature selection problem:

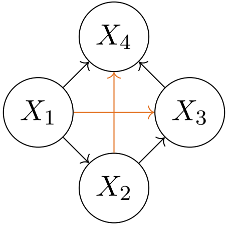
1. Draw the DAG inserting all possible edges that conform to the causal ordering.



Pruning edges

Given a valid causal ordering, finding the true DAG boils down to a model/feature selection problem:

1. Draw the DAG inserting all possible edges that conform to the causal ordering.
2. For each node in this order perform feature selection to obtain the "active" edges.



Nonlinear causal discovery for grouped data

FIRST PHASE:

- Multiresponse/Multitask learning problem $\rightarrow$ Deep NN.
- Vector-vector nonparametric marginal independence test $\rightarrow$ Hilbert Schmidt Independence Criterion (HSIC) [Gretton et al., 2005].

Nonlinear causal discovery for grouped data

FIRST PHASE:

- Multiresponse/Multitask learning problem $\rightarrow$ Deep NN.
- Vector-vector nonparametric marginal independence test $\rightarrow$ Hilbert Schmidt Independence Criterion (HSIC) [Gretton et al., 2005].

SECOND PHASE:

- Multiresponse group sparse additive models (MURGS).

MURGS

- MURGS can be cast as a penalized $\mathbf{M}$-estimator through the following optimization problem

$$\hat{\mathbf{f}} = \min_{\mathbf{f}:f_{g,h}^{(k)} \in \mathcal{H}_{g,h}^{(k)}} \left\{ \frac{1}{2n} \sum_{k \in [d_j], i \in [n]} \mathcal{L}_{f^{(k)}}(\mathbf{x}_i, y_i^{(k)}) + \lambda \Phi^j(f) \right\}$$

- with $\lambda > 0$ a regularization parameter and

$$\Phi^j(f) = \sum_{g \in pa_j} \sqrt{d_g} \max_{k \in [d_j]} \|\mathbf{f}_g^{(k)}\|,$$

MURGS

- MURGS can be cast as a penalized $\mathbf{M}$-estimator through the following optimization problem

$$\hat{\mathbf{f}} = \min_{\mathbf{f}: f_{g,h}^{(k)} \in \mathcal{H}_{g,h}^{(k)}} \left\{ \frac{1}{2n} \sum_{k \in [d_j], i \in [n]} \mathcal{L}_{f^{(k)}}(\mathbf{x}_i, y_i^{(k)}) + \lambda \Phi^j(f) \right\}$$

- with $\lambda > 0$ a regularization parameter and

$$\Phi^j(f) = \sum_{g \in pa_j} \sqrt{d_g} \max_{k \in [d_j]} \|\mathbf{f}_g^{(k)}\|,$$

- combining the sum of sup-norms regularization with the functional version of the $\ell 1/\ell 2$ norms.

Closed-form backfitting update

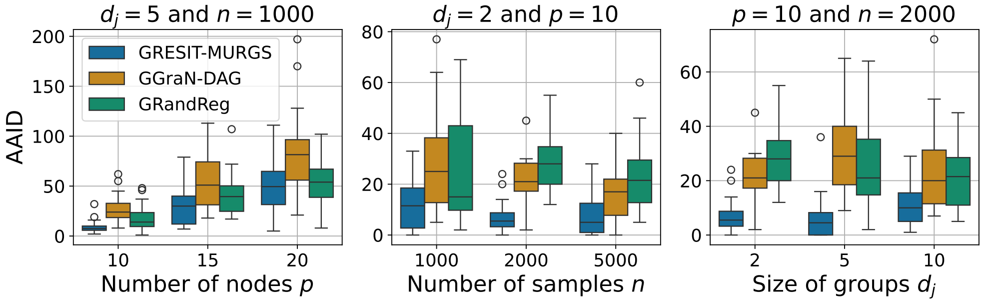
Denote $P_h = \mathbb{E}[\,\cdot \mid X_h^{(g)}]$ the conditional expectation operator, $\mathbf{Q} = (P_h)_{h\in[d_g]}$ and $s_g^{(k)} = \|\mathbf{Q}R_g^{(k)}\|$. Assume that $\mathbb{E}[f_{g,h'}^{(k)} \mid X_h^{(g)}] = 0$ for all $h' \neq h$, i.e., the covariance among the component functions within groups is zero. Order the indices according to $s_g^{(k_1)} \geq s_g^{(k_2)} \geq \cdots \geq s_g^{(k_{d_j})}$. Then the backfitting solution is given by

$$f_{g,h}^{(k_i)} = \begin{cases} P_h^{(k_i)} R_g^{(k_i)} & \text{for } i > m^* \\ \frac{1}{m^*}\left[\sum_{l=1}^{m^*} s_g^{(k_l)} - \sqrt{d_g}\lambda\right]_+ \frac{P_h^{(k_i)} R_g^{(k_i)}}{s_g^{(k_i)}} & \text{for } i \leq m^*, \end{cases}$$

for all $h \in [d_g]$ and

$$m^* = \operatorname*{arg\,max}_{m\in[d_j]} \frac{1}{m}\left(\sum_{l=1}^{m} s_g^{(k_l)} - \sqrt{d_g}\lambda\right).$$

Simulation results



Thank you all for your interest

Questions?

References I

K. A. Bollen. *Structural equations with latent variables*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Inc., New York, 1989. ISBN 0-471-01171-1. doi: 10.1002/9781118619179. URL https://doi.org/10.1002/9781118619179. A Wiley-Interscience Publication.

A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic learning theory*, volume 3734 of *Lecture Notes in Comput. Sci.*, pages 63–77. Springer, Berlin, 2005. ISBN 978-3-540-29242-5; 3-540-29242-X. doi: 10.1007/11564089_7. URL https://doi.org/10.1007/11564089_7.

P. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008. URL https://proceedings.neurips.cc/paper_files/paper/2008/

References II

file/f7664060cc52bc6f3d620bcedc94a4b6-Paper.pdf.

A. Hyvärinen and P. Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999. ISSN 0893-6080. doi: https://doi.org/10.1016/S0893-6080(98)00140-3. URL https://www.sciencedirect.com/science/article/pii/S0893608098001403.

J. Pearl. *Causality*. Cambridge University Press, Cambridge, second edition, 2009. ISBN 978-0-521-89560-6; 0-521-77362-8. doi: 10.1017/CBO9780511803161. URL https://doi.org/10.1017/CBO9780511803161. Models, reasoning, and inference.

J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. Causal discovery with continuous additive noise models. *J. Mach. Learn. Res.*, 15:2009–2053, 2014. ISSN 1532-4435,1533-7928.

References III

P. Spirtes, C. Glymour, and R. Scheines. *Causation, prediction, and search*, volume 81 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1993. ISBN 0-387-97979-4. doi: 10.1007/978-1-4612-2748-9. URL https://doi.org/10.1007/978-1-4612-2748-9.