Distribution Generalization and Identifiability in IV Models

Niklas Pfister

October 12, 2022

Graphical Models Workshop

novo nordisk fonden
Benefiting people and society

Copenhagen Causality Lab

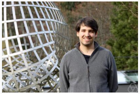
UNIVERSITY OF COPENHAGEN

Joint work with

Rune Christiansen

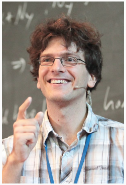
Sebastian Engelke

Nicola Gnecco



Leonard Henckel

Martin Jakobsen



Jonas Peters

Sorawit Saengkyongam

Table of contents

Distribution generalization

Observe: $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{iid}{\sim} \mathbb{P}_{\text{train}}$

Goal: Learn a function $\hat{f}$ that accurately predicts Y from X on shifted distribution $\mathbb{P}_{\text{test}}$, e.g.,

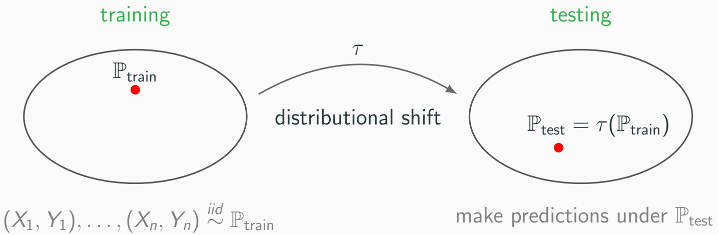
$$\hat{f} = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{\text{test}}\left[(Y - f(X))^2\right]$$

Observe: $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{iid}{\sim} \mathbb{P}_{\text{train}}$

Goal: Learn a function $\hat{f}$ that accurately predicts Y from X on shifted distribution $\mathbb{P}_{\text{test}}$, e.g.,

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{\text{test}}\left[(Y - f(X))^2\right]$$

$\rightarrow$ Requires relation between $\mathbb{P}_{\text{train}}$ and $\mathbb{P}_{\text{test}}$



Worst-case optimal

Let $\mathcal{P}$ be a collection of potential test distributions and consider

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P[(Y - \hat{f}(X))^2] = \inf_{f \in \mathcal{F}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[(Y - f(X))^2].$$

Worst-case optimal

Let $\mathcal{P}$ be a collection of potential test distributions and consider

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P[(Y - \hat{f}(X))^2] = \inf_{f \in \mathcal{F}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[(Y - f(X))^2].$$

Relevant if mistakes have potentially catastrophic consequences! (Self driving cars, medical applications, ...)

Existing (non-causal) approaches

- covariate shift e.g., Shimodaira et al. (2000), Sugiyama et al. (2005), ...
  → *train and test have the same conditional* $Y|X$*, i.e.,*

$$\mathbb{P}^{Y|X}_{\text{train}} = \mathbb{P}^{Y|X}_{\text{test}}$$

Existing (non-causal) approaches

- covariate shift e.g., Shimodaira et al. (2000), Sugiyama et al. (2005), ...
  → *train and test have the same conditional $Y|X$, i.e.,*

$$\mathbb{P}_{\text{train}}^{Y|X} = \mathbb{P}_{\text{test}}^{Y|X}$$

- distributional robustness e.g., Bagnell (2005), Abadeh et al. (2015), ...
  → *given a metric d, test is small perturbation of training*

$$d(\mathbb{P}_{\text{train}}, \mathbb{P}_{\text{test}}) < \epsilon$$

Existing (non-causal) approaches

- covariate shift e.g., Shimodaira et al. (2000), Sugiyama et al. (2005), ...
  → *train and test have the same conditional $Y|X$, i.e.,*

$$\mathbb{P}^{Y|X}_{\text{train}} = \mathbb{P}^{Y|X}_{\text{test}}$$

- distributional robustness e.g., Bagnell (2005), Abadeh et al. (2015), ...
  → *given a metric d, test is small perturbation of training*

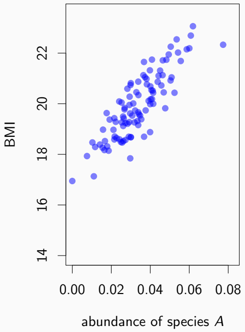
$$d(\mathbb{P}_{\text{train}}, \mathbb{P}_{\text{test}}) < \epsilon$$

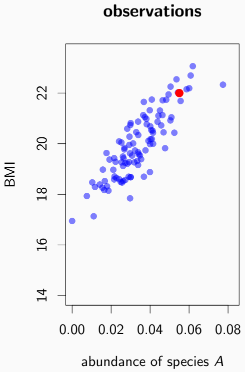
- maximin effects & DRO e.g., Meinshausen and Bühlmann (2015), Sagawa et al. (2019), ...
  → *test lies in convex hull of training distributions*

$$\mathbb{P}_{\text{test}} \in \text{ConvexHull}(\{\mathbb{P}^{1}_{\text{train}}, \dots, \mathbb{P}^{m}_{\text{train}}\})$$

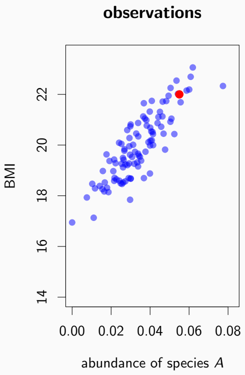
How does causality help?

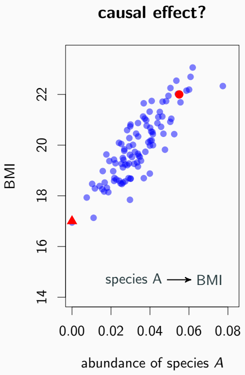
observations




observations
BMI
14
16
18
20
22
0.00
0.02
0.04
0.06
0.08
abundance of species A


https://cdn.the-scientist.com

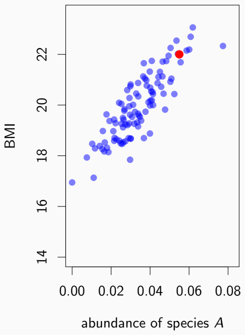

observations
BMI
14
16
18
20
22
0.00
0.02
0.04
0.06
0.08
abundance of species A


causal effect?
BMI
14
16
18
20
22
species A ⟶ BMI
0.00
0.02
0.04
0.06
0.08
abundance of species A

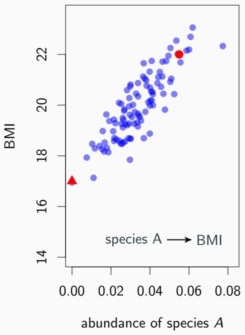

https://cdn.the-scientist.com


https://cdn.the-scientist.com

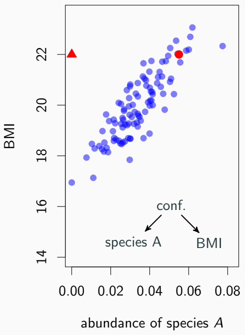
observations

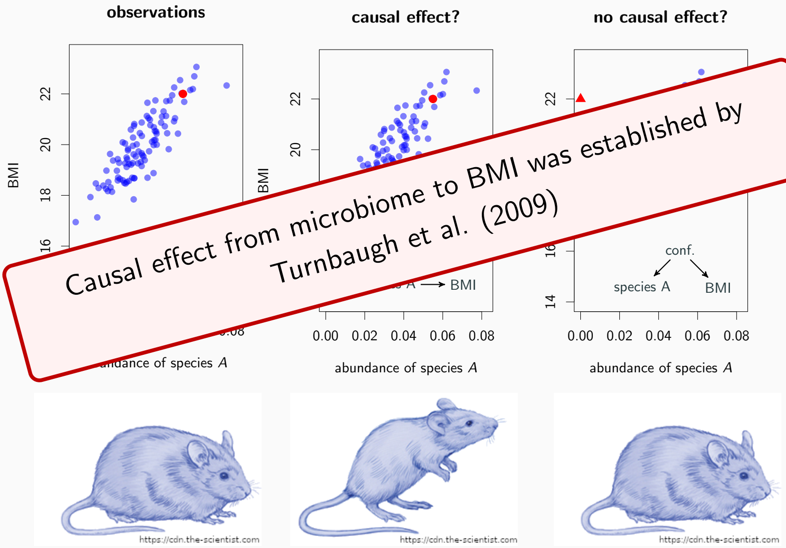


causal effect?

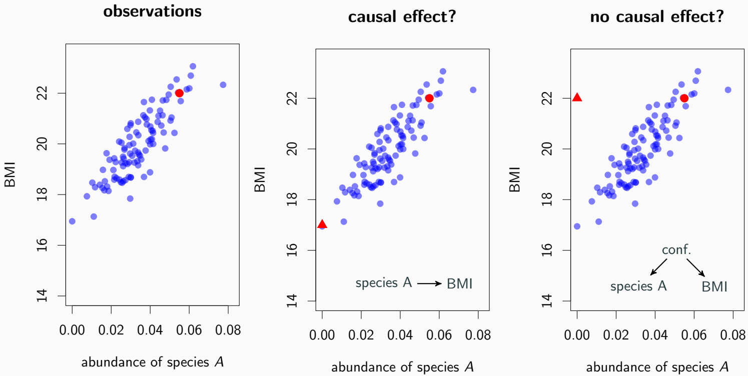


no causal effect?





Causal effect from microbiome to BMI was established by Turnbaugh et al. (2009)



A causal model describes the observational distribution and a set of intervention distributions.

Assume train and test are generated by a causal model $\mathcal{M}$ with

$$\mathbb{P}_{\text{train}} = \mathbb{P}_{\mathcal{M}} \quad \text{(obs. distr.)} \quad \text{and} \quad \mathbb{P}_{\text{test}} = \mathbb{P}_{\mathcal{M}(i)} \quad \text{(int. distr.)}$$

for some intervention $i \in \mathcal{I}$.

Assume train and test are generated by a causal model $\mathcal{M}$ with

$$\mathbb{P}_{\text{train}} = \mathbb{P}_{\mathcal{M}} \quad \text{(obs. distr.)} \quad \text{and} \quad \mathbb{P}_{\text{test}} = \mathbb{P}_{\mathcal{M}(i)} \quad \text{(int. distr.)}$$

for some intervention $i \in \mathcal{I}$.

$\mathbb{P}_{\text{train}}$ and $\mathbb{P}_{\text{test}}$ are related by constraints on

(1) the underlying causal model $\mathcal{M}$ and

(2) the set of allowed interventions $\mathcal{I}$.

Assume train and test are generated by a causal model $\mathcal{M}$ with

$$\mathbb{P}_{\text{train}} = \mathbb{P}_{\mathcal{M}} \quad \text{(obs. distr.)} \quad \text{and} \quad \mathbb{P}_{\text{test}} = \mathbb{P}_{\mathcal{M}(i)} \quad \text{(int. distr.)}$$

for some intervention $i \in \mathcal{I}$.

Goal: Find $\hat{f} \in \mathcal{F}$ such that

$$\sup_{i \in \mathcal{I}} \mathbb{E}_{\mathcal{M}(i)}[(Y - \hat{f}(X))^2] = \inf_{f \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{\mathcal{M}(i)}[(Y - f(X))^2]$$

Assume train and test are generated by a causal model $\mathcal{M}$ with

$$\mathbb{P}_{\text{train}} = \mathbb{P}_{\mathcal{M}} \quad \text{(obs. distr.)} \quad \text{and} \quad \mathbb{P}_{\text{test}} = \mathbb{P}_{\mathcal{M}(i)} \quad \text{(int. distr.)}$$

for some intervention $i \in \mathcal{I}$.

Goal: Find $\hat{f} \in \mathcal{F}$ such that

$$\sup_{i \in \mathcal{I}} \mathbb{E}_{\mathcal{M}(i)}[(Y - \hat{f}(X))^2] = \inf_{f \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{\mathcal{M}(i)}[(Y - f(X))^2]$$

Invariance assumption:
$\exists f \in \mathcal{F}$ such that

$$\forall i \in \mathcal{I}: \ \mathbb{P}^{Y-f(X)}_{\mathcal{M}(i)} = \mathbb{P}^{Y-f(X)}_{\mathcal{M}}$$

$\rightarrow$ *f is called invariant*

Assume train and test are generated by a causal model $\mathcal{M}$ with

$$\mathbb{P}_{\text{train}} = \mathbb{P}_{\mathcal{M}} \quad \text{(obs. distr.)} \quad \text{and} \quad \mathbb{P}_{\text{test}} = \mathbb{P}_{\mathcal{M}(i)} \quad \text{(int. distr.)}$$

for some intervention $i \in \mathcal{I}$.

Goal: Find $\hat{f} \in \mathcal{F}$ such that

$$\sup_{i \in \mathcal{I}} \mathbb{E}_{\mathcal{M}(i)}[(Y - \hat{f}(X))^2] = \inf_{f \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{\mathcal{M}(i)}[(Y - f(X))^2]$$

Invariance assumption:

$\exists f \in \mathcal{F}$ such that

$$\forall i \in \mathcal{I}: \ \mathbb{P}_{\mathcal{M}(i)}^{Y-f(X)} = \mathbb{P}_{\mathcal{M}}^{Y-f(X)}$$

$\rightarrow$ *f is called invariant*

Strategy:

$$\underset{f \in \mathcal{F} \text{ invariant}}{\arg\min} \ \mathbb{E}_{\mathcal{M}} \left[(Y - f(X))^2\right]$$

Assume train and test are generated by a causal model $\mathcal{M}$ with

$$\mathbb{P}_{\text{train}} = \mathbb{P}_{\mathcal{M}} \quad \text{(obs. distr.)} \quad \text{and} \quad \mathbb{P}_{\text{test}} = \mathbb{P}_{\mathcal{M}(i)} \quad \text{(int. distr.)}$$

for some intervention $i \in \mathcal{I}$.

Goal: Find $\hat{f} \in \mathcal{F}$ such that

$$\sup_{i \in \mathcal{I}} \mathbb{E}_{\mathcal{M}(i)}[(Y - \hat{f}(X))^2] = \inf_{f \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{\mathcal{M}(i)}[(Y - f(X))^2]$$

Invariance assumption:

$\exists f \in \mathcal{F}$ such that

$$\forall i \in \mathcal{I}: \; \mathbb{P}_{\mathcal{M}(i)}^{Y-f(X)} = \mathbb{P}_{\mathcal{M}}^{Y-f(X)}$$

$\rightarrow$ *f is called invariant*

Strategy:

$$\underset{f \in \mathcal{F} \text{ invariant}}{\arg\min} \; \mathbb{E}_{\mathcal{M}}\left[(Y - f(X))^2\right]$$

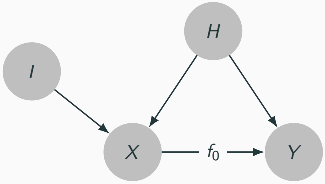
$\rightarrow$ *Can we check invariance?*

$\rightarrow$ *Is this solution minimax?*

IV-based models

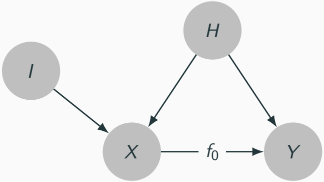
$\mathcal{M}$ is a structural causal model over observed variables (Y, X, I)

$$\begin{aligned} I &\leftarrow \epsilon_I \\ H &\leftarrow \epsilon_H \\ X &\leftarrow g(I, H, \epsilon_X) \\ Y &\leftarrow \underbrace{f_0(X)}_{\text{causal fct}} + h(H, \epsilon_Y) \end{aligned}$$



$\mathcal{M}$ is a structural causal model over observed variables (Y, X, I)

$$
\begin{aligned}
I &\leftarrow \epsilon_I \\
H &\leftarrow \epsilon_H \\
X &\leftarrow g(I, H, \epsilon_X) \\
Y &\leftarrow \underbrace{f_0(X)}_{\text{causal fct}} + h(H, \epsilon_Y)
\end{aligned}
$$

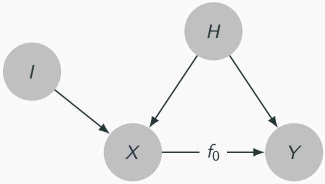


We can now look at two classes of interventions:

- $\mathcal{I}_I$ the set of *all interventions on* I
- $\mathcal{I}_X$ the set of *all interventions on* X

$\mathcal{M}$ is a structural causal model over observed variables (Y, X, I)

$$\begin{aligned} I &\leftarrow \epsilon_I \\ H &\leftarrow \epsilon_H \\ X &\leftarrow g(I, H, \epsilon_X) \\ Y &\leftarrow \underbrace{f_0(X)}_{\text{causal fct}} + h(H, \epsilon_Y) \end{aligned}$$



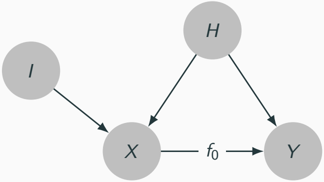
We can now look at two classes of interventions:

- $\mathcal{I}_I$ the set of *all interventions on* I
- $\mathcal{I}_X$ the set of *all interventions on* X

What functions are invariant in each case?

$\mathcal{M}$ is a structural causal model over observed variables (Y, X, I)

$$
\begin{aligned}
I &\leftarrow \epsilon_I \\
H &\leftarrow \epsilon_H \\
X &\leftarrow g(I, H, \epsilon_X) \\
Y &\leftarrow \underbrace{f_0(X)}_{\text{causal fct}} + h(H, \epsilon_Y)
\end{aligned}
$$



We can now look at two classes of interventions:

- $\mathcal{I}_I$ the set of *all interventions on* I
- $\mathcal{I}_X$ the set of *all interventions on* X

What functions are invariant in each case?

Case 1 f is invariant wrt $\mathcal{I}_X$ iff $f = f_0$

$\rightarrow$ generalization wrt $\mathcal{I}_X$ requires identifiability of f_0

Case 2 f is invariant wrt $\mathcal{I}_Z$ iff $Y - f(X) \perp\!\!\!\perp Z$ under $\mathbb{P}_{\mathcal{M}}$

$\rightarrow$ generalization wrt $\mathcal{I}_Z$ does **not** require identifiability of f_0

Identifiability of the causal function

Classical IV: For fixed basis η, f_0 is called identifiable if

$$\big\{f \in \mathcal{F} \mid \mathrm{cov}(\eta(I), Y - f(X)) = 0\big\} = \{f_0\} \quad \text{(moment identif. cond.)}$$

Identifiability of the causal function

Classical IV: For fixed basis η, f_0 is called identifiable if

$$\left\{f \in \mathcal{F} \mid \mathrm{cov}(\eta(I), Y - f(X)) = 0\right\} = \{f_0\} \quad \text{(moment identif. cond.)}$$

E.g., if $\mathcal{F}$ linear functions and η identity

$$\{\beta \mid \mathrm{cov}(I, Y) = \mathrm{cov}(I, X)\beta\} = \{\beta_0\} \Leftrightarrow \mathrm{cov}(I, X) \text{ full col-rank}$$

Identifiability of the causal function

Classical IV: For fixed basis η, f_0 is called identifiable if

$$\left\{f \in \mathcal{F} \mid \operatorname{cov}(\eta(I), Y - f(X)) = 0\right\} = \{f_0\} \quad \text{(moment identif. cond.)}$$

Can this be strengthened?

Identifiability of the causal function

Classical IV: For fixed basis η, f_0 is called identifiable if

$$\{f \in \mathcal{F} \mid \mathrm{cov}(\eta(I), Y - f(X)) = 0\} = \{f_0\} \quad \text{(moment identif. cond.)}$$

Can this be strengthened? Yes!

- Independence IV Imbens & Newey (2009), Torgovitsky (2015), Saengkyongam et al. (2022), ...

$$\{f \in \mathcal{F} \mid Y - f(X) \perp\!\!\!\perp I\} = \{f_0\} \quad \text{(independence identif. cond.)}$$

e.g., binary instruments can identify nonlinear effects

Identifiability of the causal function

Classical IV: For fixed basis η, f_0 is called identifiable if

$$\left\{f \in \mathcal{F} \mid \operatorname{cov}(\eta(I), Y - f(X)) = 0\right\} = \{f_0\} \quad \text{(moment identif. cond.)}$$

Can this be strengthened? Yes!

- Independence IV Imbens & Newey (2009), Torgovitsky (2015), Saengkyongam et al. (2022), ...

$$\left\{f \in \mathcal{F} \mid Y - f(X) \perp\!\!\!\perp I\right\} = \{f_0\} \quad \text{(independence identif. cond.)}$$

  e.g., binary instruments can identify nonlinear effects

- Sparse causal effects IV (SpaceIV) NP & Peters (2022)

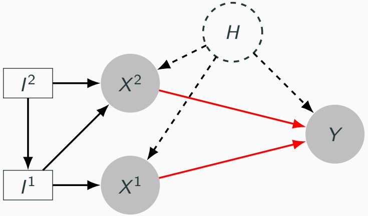
$$\min_{\beta \in \mathcal{B}} \|\beta\|_0 \quad \text{with} \quad \mathcal{B} = \{\beta \mid \operatorname{cov}(I, Y) = \operatorname{cov}(I, X)\beta\}$$

  e.g., settings with many more Xs than Is can be identifiable

SpaceIV

Linear SCM with interventions:

$$X = BX + AI + h(H, \varepsilon_X)$$
$$Y = \beta_1^* X^1 + \beta_2^* X^2 + g(H, \varepsilon_Y)$$



(IV1) If I and Y are d-separated when removing $X^1, X^2 \to Y$, then

$$(\beta_1, \beta_2) = (\beta_1^*, \beta_2^*) \quad \Rightarrow \quad \mathrm{cov}\left(I, Y - \beta_1 X^1 - \beta_2 X^2\right) = 0.$$

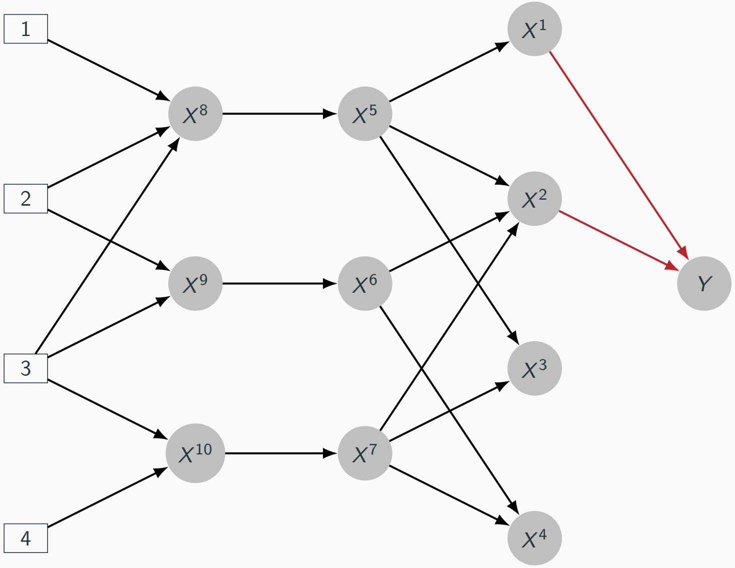
(IV2) If, in addition, $cov(I, X)$ is col-full rank, then

$$(\beta_1, \beta_2) = (\beta_1^*, \beta_2^*) \quad \Leftrightarrow \quad \mathrm{cov}\left(I, Y - \beta_1 X^1 - \beta_2 X^2\right) = 0.$$

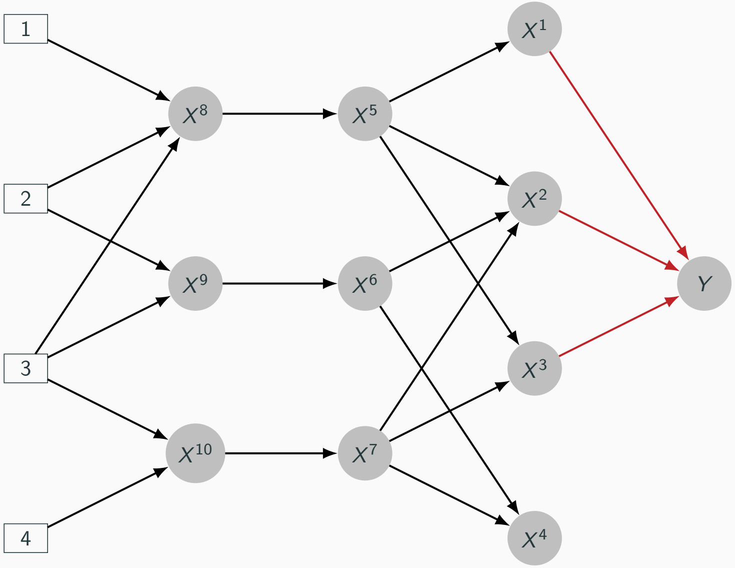
Anderson and Rubin 1949, Theil 1953, Mendelian Randomization...

If there are more covariates than instruments, the causal function is not identifiable. Can we exploit sparsity of the effect?

Is the causal function identifiable?



Is the causal function identifiable?



Include sparsity

Consider the solution space

$$\mathcal{B} := \{\beta \in \mathbb{R}^d \mid \mathrm{cov}(I, Y) = \mathrm{cov}(I, X)\beta\}$$

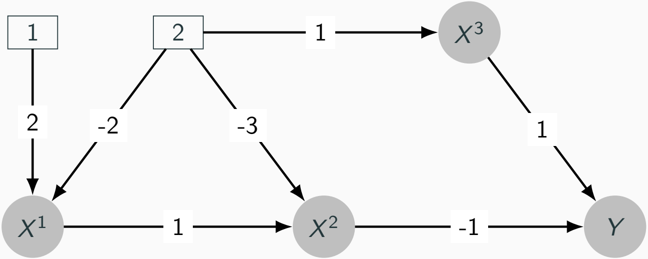
and

$$\underset{\beta \in \mathcal{B}}{\arg\min} \; \|\beta\|_0.$$

When is this equal to β^*?

An important quantity is

$$C_{ij} := \text{ total causal effect from } I^i \text{ to } X^j.$$

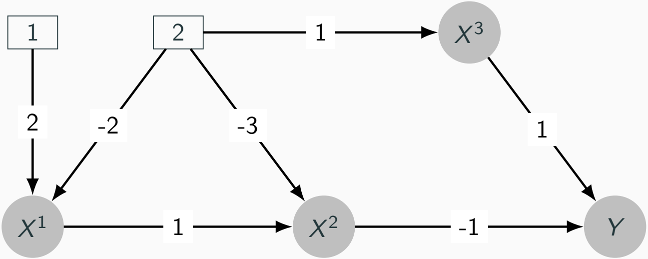


Then

$$C = \begin{pmatrix} 2 & 2 & 0 \\ -2 & -5 & 1 \end{pmatrix}.$$

An important quantity is

$$C_{ij} := \text{ total causal effect from } I^i \text{ to } X^j.$$



Then

$$C = \begin{pmatrix} 2 & 2 & 0 \\ -2 & -5 & 1 \end{pmatrix}.$$

For Lasso "restricted nullspace property of X", here the *intervention subspace* needs to behave nicely...

(A1) **Non-degenerate:** It holds that rank $C_{\mathrm{PA}(Y)} = |\,\mathrm{PA}(Y)|$.

(A2) **No cancellation:** For all $S \subseteq \{1, \ldots, d\}$ it holds that

$$\left.\begin{array}{r} \mathrm{rank}(C_S) \leq \mathrm{rank}(C_{\mathrm{PA}(Y)}) \\ \text{and } \ \mathrm{im}(C_S) \neq \mathrm{im}(C_{\mathrm{PA}(Y)}) \end{array}\right\} \Rightarrow \left\{ \forall w \in \mathbb{R}^{|S|} : \ C_S w \neq C_{\mathrm{PA}(Y)} \beta^*_{\mathrm{PA}(Y)} \right. .$$

(This is implied by random coefficients.)

(A3) **Uniqueness:** For all $S \subseteq \{1, \ldots, d\}$ with $|S| = |\,\mathrm{PA}(Y)|$ and $S \neq \mathrm{PA}(Y)$ we have $\mathrm{im}(C_S) \neq \mathrm{im}(C_{\mathrm{PA}(Y)})$.

(A1) **Non-degenerate:** It holds that $\operatorname{rank} C_{\mathrm{PA}(Y)} = |\mathrm{PA}(Y)|$.

(A2) **No cancellation:** For all $S \subseteq \{1, \ldots, d\}$ it holds that

$$\left.\begin{array}{r}\operatorname{rank}(C_S) \leq \operatorname{rank}(C_{\mathrm{PA}(Y)}) \\ \text{and } \operatorname{im}(C_S) \neq \operatorname{im}(C_{\mathrm{PA}(Y)})\end{array}\right\} \Rightarrow \left\{\forall w \in \mathbb{R}^{|S|} : C_S w \neq C_{\mathrm{PA}(Y)} \beta^*_{\mathrm{PA}(Y)}\right. .$$

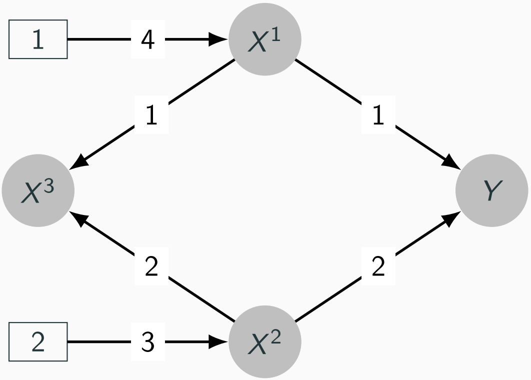
(This is implied by random coefficients.)

(A3) **Uniqueness:** For all $S \subseteq \{1, \ldots, d\}$ with $|S| = |\mathrm{PA}(Y)|$ and $S \neq \mathrm{PA}(Y)$ we have $\operatorname{im}(C_S) \neq \operatorname{im}(C_{\mathrm{PA}(Y)})$.

Theorem (Identifiability of sparse causal parameters)

- If (A1) and (A2) hold, then $\beta^* \in \arg\min_{\beta \in \mathcal{B}} \|\beta\|_0$.
- If additionally (A3) holds, then β^* is unique solution.

An example violating (A2):



What does this mean graphically?

(B1) There are at least $|\operatorname{PA}(Y)|$ disjoint directed paths (not sharing any node) from I to $\operatorname{PA}(Y)$.

(B2) Random coefficients.

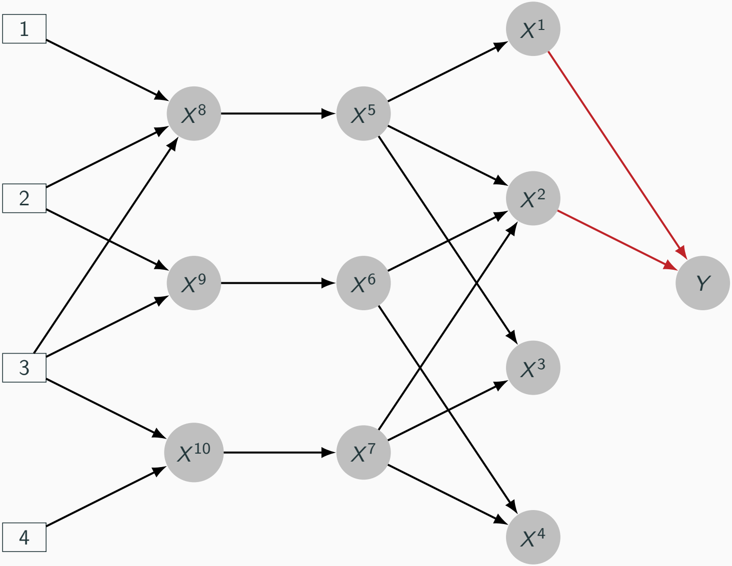
(B3) For all $S \subseteq \{1, \ldots, d\}$ with $|S| = |\operatorname{PA}(Y)|$ and $S \neq \operatorname{PA}(Y)$ at least one of the following conditions is satisfied

- (i) $\operatorname{AN}_I[S] \neq \operatorname{AN}_I[\operatorname{PA}(Y)]$.
- (ii) The smallest set T of nodes such that all directed paths from I to $\operatorname{PA}(Y)$ and from I to S go through T is of size at least $|\operatorname{PA}(Y)| + 1$.

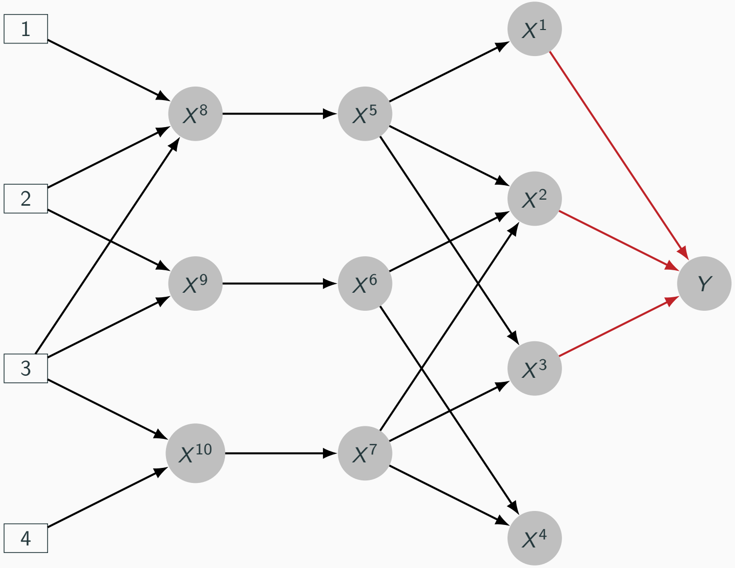
Theorem:

(B1)–(B3) imply (A1)–(A3).

Is the causal function identifiable?



Is the causal function identifiable?



Conclusions

- Causal models can be used to formalize distributional shifts.
- IV-type models offer a rich class of practically relevant models on which distribution generalization is possible.
- Two types of generalizations:
  (1) Interventions on X: requires identifiability
  (2) Interventions on Z: possible even in the non-identifiable case
- Sparse causal effects may lead to identifiability and hence generalization to interventions on X.

Conclusions

- Causal models can be used to formalize distributional shifts.
- IV-type models offer a rich class of practically relevant models on which distribution generalization is possible.
- Two types of generalizations:
  (1) Interventions on X: requires identifiability
  (2) Interventions on Z: possible even in the non-identifiable case
- Sparse causal effects may lead to identifiability and hence generalization to interventions on X.

NP, J. Peters: *Identifiability of Sparse Causal Effects using Instrumental Variables.* In Proceedings of the 38th Annual Conference on Uncertainty in Artificial Intelligence (UAI).

S. Saengkyongam, L. Henckel, NP, J. Peters: *Exploiting Independent Instruments: Identification and Distribution Generalization.* In Proceedings of the 39th International Conference on Machine Learning (ICML).

R. Christiansen, NP, M. Jakobsen, N. Gnecco, J. Peters: *A Causal Framework for Distribution Generalization.* IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI).

Thank you!

Additional slides...

Simulations

Simulation setup:

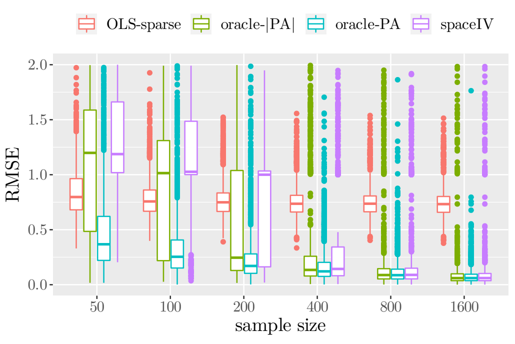
- Generate 2000 random linear SCMs with $d = 20$ predictors and $m = 10$ interventions.
- For each model generate a data set of $n = 1600$ iid observations of (X, Y, I).
- For each model check which assumptions A1 and A3 are satisfied (A2 is true by construction).
- Compute prediction error (root mean squared error) and estimated probability that the correct sparsity level was selected.

Comparison methods:

- *OLS-sparse:* Goes over all subsets of size 3, fits linear OLS and selects model using AIC.
- *oracle-PA:* Uses the correct parent set and fits an IV estimator.
- *oracle-$|PA|$:* Goes over all subsets of size 2, fits IV estimator and selects model with smallest squared moment condition loss.

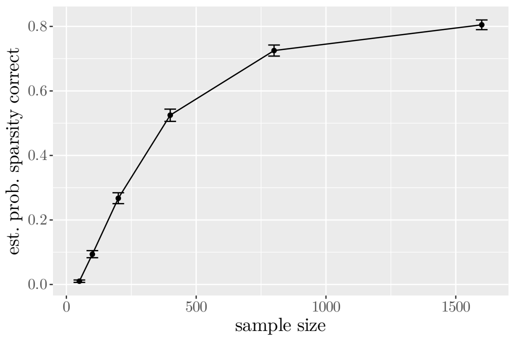
Prediction error

- Only includes random models satisfying (A1)-(A3)
- Varying sample size



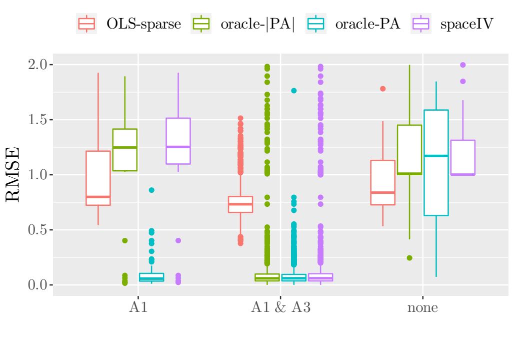
Estimation of sparsity

- Only includes random models satisfying (A1)-(A3)
- Varying sample size



Validating assumptions

- Fixed sample size $n = 1600$
- Prediction error depending on which assumptions are satisfied



References

H. Shimodaira. *Improving predictive inference under covariate shift by weighting the log-likelihood function.* Journal of Statistical Planning and Inference, 90(2):227 – 244, 2000.

M. Sugiyama and K. Müller. *Generalization error estimation under covariate shift.* In Workshop on Information-Based Induction Sciences, 2005.

J. A. Bagnell. *Robust supervised learning. In Proceedings of the 20th National Conference on Artificial Intelligence*, pages 714–719, 2005.

S. Abadeh, P. Esfahani, and D. Kuhn. *Distributionally robust logistic regression.* In Advances in Neural Information Processing Systems (NeurIPS), pages 1576–1584, 2015.

N. Meinshausen and P. Bühlmann. *Maximin effects in inhomogeneous large-scale data.* The Annals of Statistics 43(4):1801-1830, 2015.

S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. *Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization.* International Conference on Learning Representations, 2019.

G. Imbens, and W. Newey. *Identification and estimation of triangular simultaneous equations models without additivity.* Econometrica 77, no. 5 (2009): 1481-1512.

A. Torgovitsky. *Identification of nonseparable models using instruments with small support.* Econometrica 83, no. 3 (2015): 1185-1197.