Causal Discovery from Interventional Data

Bachelor's Thesis

Adam Gorm Hoffmann

Supervisors

Jonas Peters   Niklas Pfister

Department of Mathematical Sciences
University of Copenhagen

September 2021

Agenda

1. Problems addressed by the thesis
2. Proposed methods
3. Simulation studies and results
4. Conclusion

Problem description

Goal: Learn causes of response Y among covariates X.
Setting: Two repetitions of the same set of experiments.

Problem description

Goal: Learn causes of response Y among covariates X.
Setting: Two repetitions of the same set of experiments.

Problems:

(A) X and Y come from separate sets of experiments

(B) We observe (X, Y) in both sets of experiments

(C) We observe (X, Y) in a single set of experiments

Underlying SCM and shift interventions

Underlying SCM

$$
\begin{aligned}
H &:= N_H \\
X &:= A(H, X, Y) + N_X \\
Y &:= \beta^t X + \gamma^t H + N_Y
\end{aligned}
$$

Underlying SCM and shift interventions

Underlying SCM

$$
\begin{aligned}
H &:= N_H \\
X &:= A(H, X, Y) + N_X \\
Y &:= \beta^t X + \gamma^t H + N_Y
\end{aligned}
$$

Shift interventions

$$
\begin{aligned}
H^1 &:= N_{H^1} \\
X^1 &:= A(H^1, X^1, Y^1) + N_{X^1} + {\color{red}W^1} \\
Y^1 &:= \beta^t X^1 + \gamma^t H^1 + N_{Y^1}
\end{aligned}
$$

$$
\begin{aligned}
H^2 &:= N_{H^2} \\
X^2 &:= A(H^2, X^2, Y^2) + N_{X^2} + {\color{green}W^2} \\
Y^2 &:= \beta^t X^2 + \gamma^t H^2 + N_{Y^2}
\end{aligned}
$$

Underlying SCM and shift interventions

Underlying SCM

$$H := N_H$$
$$X := A(H, X, Y) + N_X$$
$$Y := \beta^t X + \gamma^t H + N_Y$$

Shift interventions

$$H^1 := N_{H^1}$$
$$X^1 := A(H^1, X^1, Y^1) + N_{X^1} + {\color{red}W^1}$$
$$Y^1 := \beta^t X^1 + \gamma^t H^1 + N_{Y^1}$$

$$H^2 := N_{H^2}$$
$$X^2 := A(H^2, X^2, Y^2) + N_{X^2} + {\color{green}W^2}$$
$$Y^2 := \beta^t X^2 + \gamma^t H^2 + N_{Y^2}$$

Data

$$(\mathbf{X}^1, \mathbf{Y}^1) = \begin{pmatrix} x^{1,1} & y^{1,1} \\ \vdots & \vdots \\ x^{1,n_1} & y^{1,n_1} \end{pmatrix}$$

$$(\mathbf{X}^2, \mathbf{Y}^2) = \begin{pmatrix} x^{2,1} & y^{2,1} \\ \vdots & \vdots \\ x^{2,n_2} & y^{2,n_2} \end{pmatrix}$$

Same ${\color{red}W^1}$ for all rows

Same ${\color{green}W^2}$ for all rows

Two repetitions of the same experiments

$$H := N_H$$
$$X := A(H, X, Y) + N_X$$
$$Y := \beta^t X + \gamma^t H + N_Y$$

$$\widetilde{H} := \widetilde{N}_{\widetilde{H}}$$
$$\widetilde{X} := A(\widetilde{H}, \widetilde{X}, \widetilde{Y}) + \widetilde{N}_{\widetilde{X}}$$
$$\widetilde{Y} := \beta^t \widetilde{X} + \gamma^t \widetilde{H} + \widetilde{N}_{\widetilde{Y}}$$

Same underlying SCM

Different noise variables

Two repetitions of the same experiments

$$H := N_H$$
$$X := A(H, X, Y) + N_X$$
$$Y := \beta^t X + \gamma^t H + N_Y$$

$$H^1 := N_{H^1}$$
$$X^1 := A(H^1, X^1, Y^1) + N_{X^1} + W^1$$
$$Y^1 := \beta^t X^1 + \gamma^t H^1 + N_{Y^1}$$

$$H^2 := N_{H^2}$$
$$X^2 := A(H^2, X^2, Y^2) + N_{X^2} + W^2$$
$$Y^2 := \beta^t X^2 + \gamma^t H^2 + N_{Y^2}$$

$$\widetilde{H} := \widetilde{N}_{\widetilde{H}}$$
$$\widetilde{X} := A(\widetilde{H}, \widetilde{X}, \widetilde{Y}) + \widetilde{N}_{\widetilde{X}}$$
$$\widetilde{Y} := \beta^t \widetilde{X} + \gamma^t \widetilde{H} + \widetilde{N}_{\widetilde{Y}}$$

$$\widetilde{H}^1 := \widetilde{N}_{\widetilde{H}^1}$$
$$\widetilde{X}^1 := A(\widetilde{H}^1, \widetilde{X}^1, \widetilde{Y}^1) + N_{\widetilde{X}^1} + W^1$$
$$\widetilde{Y}^1 := \beta^t \widetilde{X}^1 + \gamma^t \widetilde{H}^1 + N_{\widetilde{Y}^1}$$

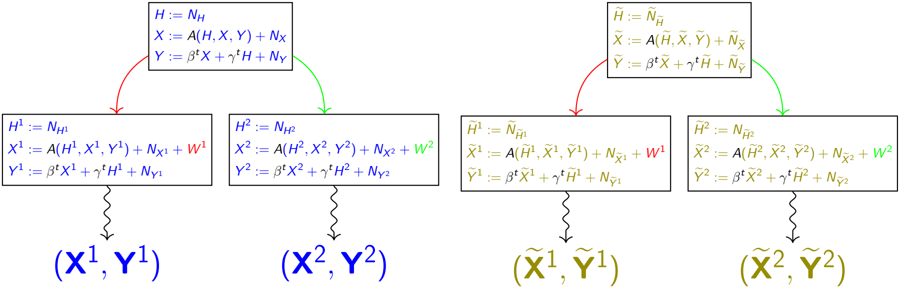
$$\widetilde{H}^2 := N_{\widetilde{H}^2}$$
$$\widetilde{X}^2 := A(\widetilde{H}^2, \widetilde{X}^2, \widetilde{Y}^2) + N_{\widetilde{X}^2} + W^2$$
$$\widetilde{Y}^2 := \beta^t \widetilde{X}^2 + \gamma^t \widetilde{H}^2 + N_{\widetilde{Y}^2}$$

Same underlying SCM

Different noise variables

Same shift interventions

Two repetitions of the same experiments



Same underlying SCM

Different noise variables

Same shift interventions

Two separate data sets
for each intervention

More detailed problem description

$$(\mathbf{X}^1, \mathbf{Y}^1) \quad (\mathbf{X}^2, \mathbf{Y}^2) \qquad (\widetilde{\mathbf{X}}^1, \widetilde{\mathbf{Y}}^1) \quad (\widetilde{\mathbf{X}}^2, \widetilde{\mathbf{Y}}^2)$$

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}^1 \\ \mathbf{X}^2 \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} \mathbf{Y}^1 \\ \mathbf{Y}^2 \end{pmatrix}$$

$$\widetilde{\mathbf{X}} = \begin{pmatrix} \widetilde{\mathbf{X}}^1 \\ \widetilde{\mathbf{X}}^2 \end{pmatrix}, \quad \widetilde{\mathbf{Y}} = \begin{pmatrix} \widetilde{\mathbf{Y}}^1 \\ \widetilde{\mathbf{Y}}^2 \end{pmatrix}$$

Goal: Learn causes of response Y among covariates X.

More detailed problem description

$$(\mathbf{X}^1, \mathbf{Y}^1) \quad (\mathbf{X}^2, \mathbf{Y}^2) \qquad (\widetilde{\mathbf{X}}^1, \widetilde{\mathbf{Y}}^1) \quad (\widetilde{\mathbf{X}}^2, \widetilde{\mathbf{Y}}^2)$$

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}^1 \\ \mathbf{X}^2 \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} \mathbf{Y}^1 \\ \mathbf{Y}^2 \end{pmatrix} \qquad \widetilde{\mathbf{X}} = \begin{pmatrix} \widetilde{\mathbf{X}}^1 \\ \widetilde{\mathbf{X}}^2 \end{pmatrix}, \quad \widetilde{\mathbf{Y}} = \begin{pmatrix} \widetilde{\mathbf{Y}}^1 \\ \widetilde{\mathbf{Y}}^2 \end{pmatrix}$$

Goal: Learn causes of response Y among covariates X.

Problems:

(A) X and Y come from separate sets of experiments $(\mathbf{X}, \widetilde{\mathbf{Y}})$

(B) We observe (X, Y) in both sets of experiments $(\mathbf{X}, \mathbf{Y}, \widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}})$

(C) We observe (X, Y) in a single set of experiments $(\mathbf{X}, \mathbf{Y})$

Strategy for single experiment problem

Permute rows to turn one data set into two:

With permutation matrices P^1 and P^2, let

$$(\mathbf{X}, \mathbf{Y}) = \begin{pmatrix} \mathbf{X}^1 & \mathbf{Y}^1 \\ \mathbf{X}^2 & \mathbf{Y}^2 \end{pmatrix}, \quad (\check{\mathbf{X}}, \check{\mathbf{Y}}) = \begin{pmatrix} P^1\mathbf{X}^1 & P^1\mathbf{Y}^1 \\ P^2\mathbf{X}^2 & P^2\mathbf{Y}^2 \end{pmatrix}.$$

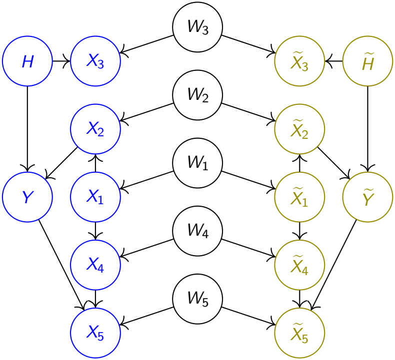
Use $(\check{\mathbf{X}}, \check{\mathbf{Y}})$ as a substitute for $(\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}})$.

OLS: Baseline

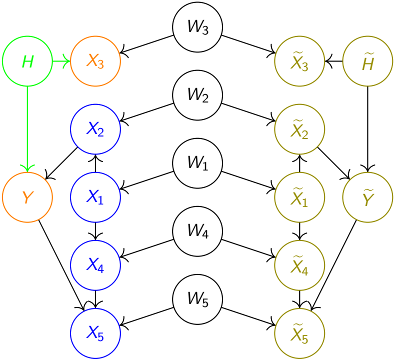
$$\beta^{\mathrm{OLS}} := \left(\sum_{k=1}^{K} \mathrm{cov}(X^k)\right)^{-1} \sum_{k=1}^{K} \mathrm{cov}(X^k, Y^k)$$
$$\hat{\beta}^{\mathrm{OLS}} := (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}$$

Method: Variable with largest $\hat{\beta}^{\mathrm{OLS}}$ value is taken as most likely parent or ancestor.

Dependence between the two sets of experiments

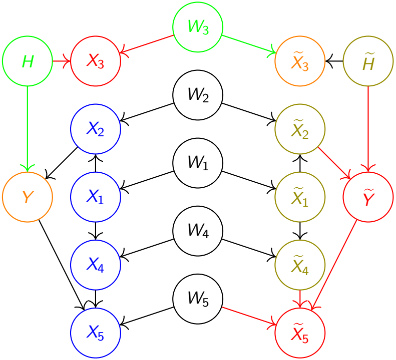


Dependence between the two sets of experiments



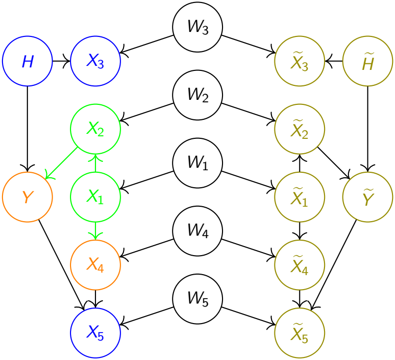
Y and X_3 confounded by H

Dependence between the two sets of experiments



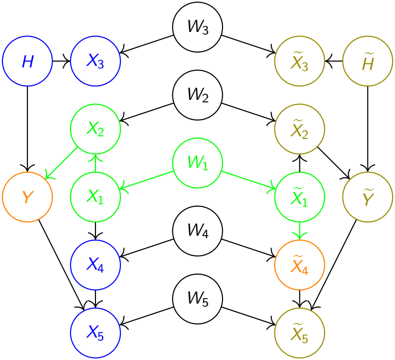
Y and X_3 confounded by H but
$Y \perp\!\!\!\perp \widetilde{X}_3$ by global Markov

Dependence between the two sets of experiments



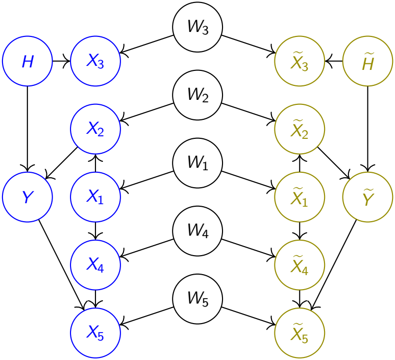
Y and X_4 confounded by X_1

Dependence between the two sets of experiments



Y and X_4 confounded by X_1
and $Y \not\perp\!\!\!\perp_{\mathcal{G}} \widetilde{X}_4$

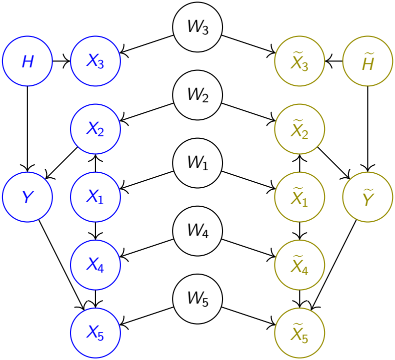
Strong Reichenbach's Common Cause Principle



$Y \not\perp\!\!\!\perp_{\mathcal{G}} \widetilde{X}_i$ if and only if

- there is a **non-hidden** confounder X_ℓ of Y and X_i, or
- X_i is an ancestor of Y

Strong Reichenbach's Common Cause Principle



$X_j \not\perp\!\!\!\perp_{\mathcal{G}} \widetilde{X}_i$ if and only if

- there is a **non-hidden** confounder X_ℓ of X_j and X_i, or
- X_i is an ancestor of X_j, or
- X_j is an ancestor of X_i

Novel methods

Idea: Break hidden confounding by using
$\widetilde{Y}^k$ instead of Y^k, or
$\widetilde{X}^k$ instead of X^k.

POLS: Learning from unpaired data

$$\beta^{\mathrm{POLS}} := \left(\sum_{k=1}^{K} \mathrm{cov}(X^k)\right)^{-1} \sum_{k=1}^{K} \mathrm{cov}(X^k, \widetilde{Y}^k)$$

$$\hat{\beta}^{\mathrm{POLS}} := (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\widetilde{\mathbf{Y}}$$

Method: Variable with largest $\hat{\beta}^{\mathrm{POLS}}$ value is taken as most likely parent or ancestor.

DPOLS: Learning from paired data

$$\beta^{\mathrm{DPOLS}} := \left(\sum_{k=1}^{K} \operatorname{cov}(X^k, \widetilde{X}^k)\right)^{-1} \sum_{k=1}^{K} \operatorname{cov}(X^k, \widetilde{Y}^k)$$
$$\hat{\beta}^{\mathrm{DPOLS}} := (\mathbf{X}^t \widetilde{\mathbf{X}})^{-1} \mathbf{X}^t \widetilde{\mathbf{Y}}$$

Method: Variable with largest $\hat{\beta}^{\mathrm{DPOLS}}$ value is taken as most likely parent or ancestor.

DPOLS finds correct parents given distribution

$$\mathrm{cov}(X^k, \widetilde{Y}^k) = \mathrm{cov}(X^k, \beta^t \widetilde{X}^k + \gamma^t \widetilde{H}^k + \widetilde{N}_{\widetilde{Y}^k}) = \mathrm{cov}(X^k, \widetilde{X}^k)\beta$$

so

$$\begin{aligned}
\beta^{\mathrm{DPOLS}} &= \left(\sum_{k=1}^{K} \mathrm{cov}(X^k, \widetilde{X}^k)\right)^{-1} \sum_{k=1}^{K} \mathrm{cov}(X^k, \widetilde{Y}^k) \\
&= \left(\sum_{k=1}^{K} \mathrm{cov}(X^k, \widetilde{X}^k)\right)^{-1} \sum_{k=1}^{K} \mathrm{cov}(X^k, \widetilde{X}^k)\beta \\
&= \beta
\end{aligned}$$

(argument from unpublished notes by Niklas Pfister)

Simulating data

1. Simulate 1000 random DAGs and coefficient matrices
2. Choose data parameters (number of observations, etc.)
3. Simulate data sets from the 1000 DAGs using parameters

Fixed parameters in this presentation

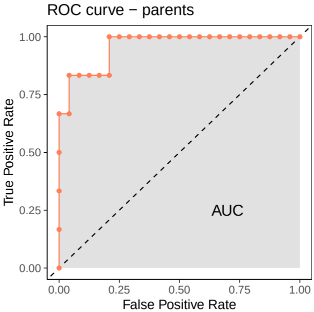
30 X's and 30 H's

$$N_{Y_j^i}, N_{X_j^i}, \widetilde{N}_{\widetilde{Y}_j^i}, \widetilde{N}_{\widetilde{X}_j^i} \overset{\text{iid.}}{\sim} \mathcal{N}(0,1)$$

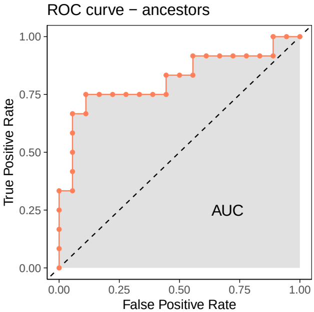
$$N_{H_j^i}, \widetilde{N}_{\widetilde{H}_j^i} \overset{\text{iid.}}{\sim} \mathcal{N}(0,5^2)$$

$$W_j^i \overset{\text{iid.}}{\sim} \mathcal{N}(0,7^2)$$

Evaluating the methods



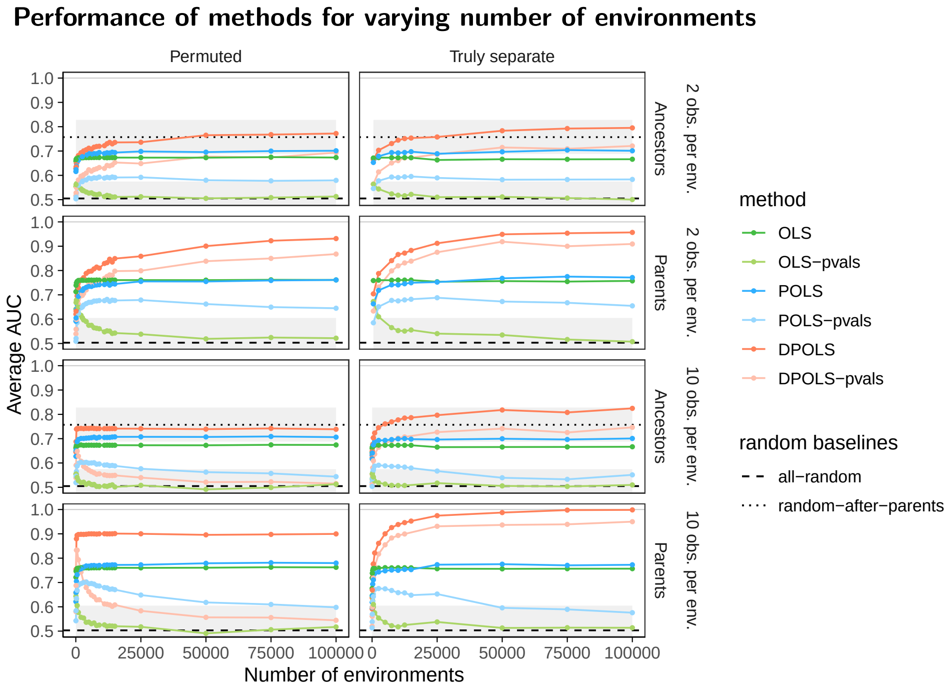
1. For all $n \in \{0, \ldots, \#X\}$
   a. Select n highest ranked variables.
   b. Calculate true positive- and false positive rates.
2. Draw ROC curve
3. Calculate AUC
4. Average 1000 AUCs

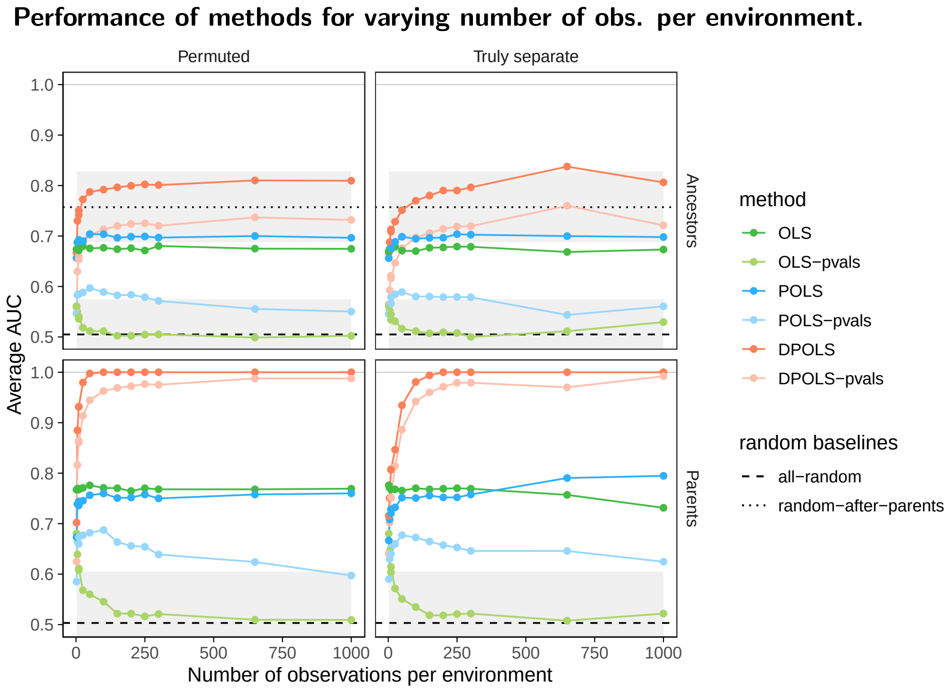


Random baseline methods

all-random
Random ranking of variables.

random-after-parents
Ranks correct parents highest; ranks remaining variables in random order.


Performance of methods for varying number of environments
Permuted
Truly separate
Ancestors
Parents
2 obs. per env.
10 obs. per env.
Average AUC
Number of environments
0
25000
50000
75000
100000
0.5
0.6
0.7
0.8
0.9
1.0
method
OLS
OLS−pvals
POLS
POLS−pvals
DPOLS
DPOLS−pvals
random baselines
all−random
random−after−parents



Performance of methods for varying number of obs. per environment.
Permuted
Truly separate
Ancestors
Parents
Average AUC
Number of observations per environment
1.0
0.9
0.8
0.7
0.6
0.5
0
250
500
750
1000
method
OLS
OLS−pvals
POLS
POLS−pvals
DPOLS
DPOLS−pvals
random baselines
all−random
random−after−parents

Conclusions

- DPOLS
  - selects correct parents asymptotically on truly separate data
  - performs well on permuted data
  - is able to select some extra ancestors after selecting all parents
- POLS
  - is viable for causal discovery from unpaired data
  - is not as good as DPOLS on paired data

Conclusions

- DPOLS
  - selects correct parents asymptotically on truly separate data
  - performs well on permuted data
  - is able to select some extra ancestors after selecting all parents
- POLS
  - is viable for causal discovery from unpaired data
  - is not as good as DPOLS on paired data

Future work:
How many variables to select?