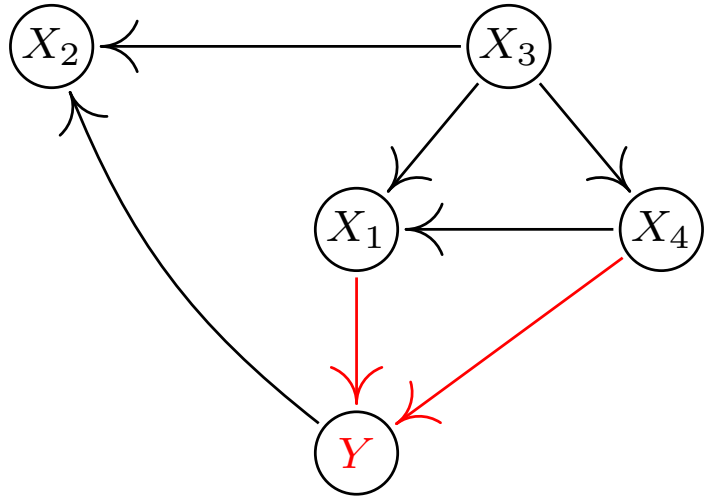
Causality



Jonas Peters and Sorawit (James) Saengkyongam
University of Copenhagen
VISUM, 7.7.2021

CARLSBERGFONDET
VILLUM FONDEN
UNIVERSITY OF COPENHAGEN

Open exercises:

- `https://github.com/CoCaLa/causality-tutorial-exercises`
- Mybinder (R or Python)
- If notebook is inactive, click ('Kernel' $\rightarrow$ 'Restart').

Disclaimer:

- This tutorial presents work by many people. Apologies if the references are not complete. A more complete list can be found in

  Peters et al.: Elements of Causal Inference, MIT Press 2017.

Disclaimer:

- This tutorial presents work by many people. Apologies if the references are not complete. A more complete list can be found in

  Peters et al.: Elements of Causal Inference, MIT Press 2017.

- The presentation is biased. In particular, there is little statistics and nothing about potential outcomes. Good books include

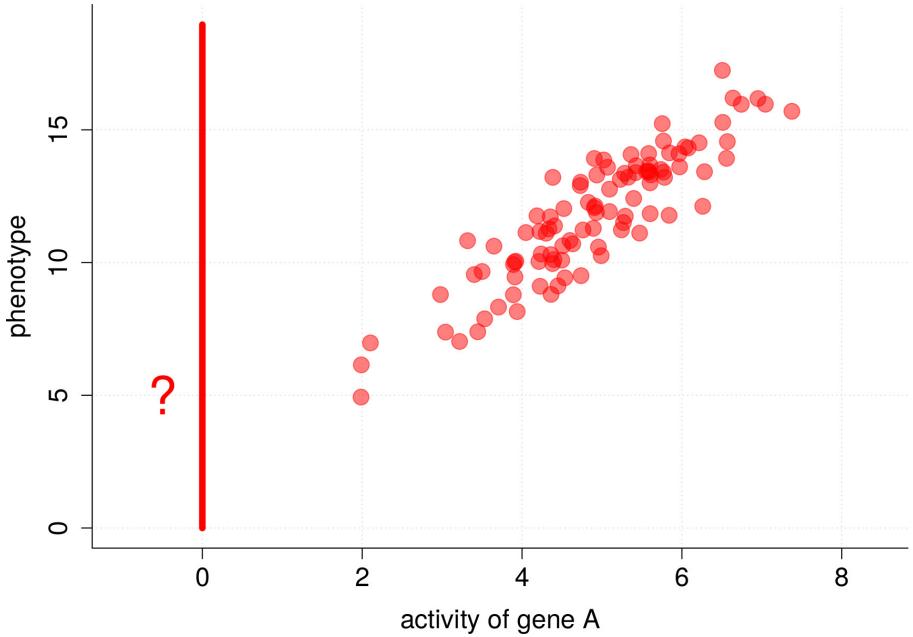
  Hernan & Robins: Causal Inference, Chapman & Hall/CRC 2019,

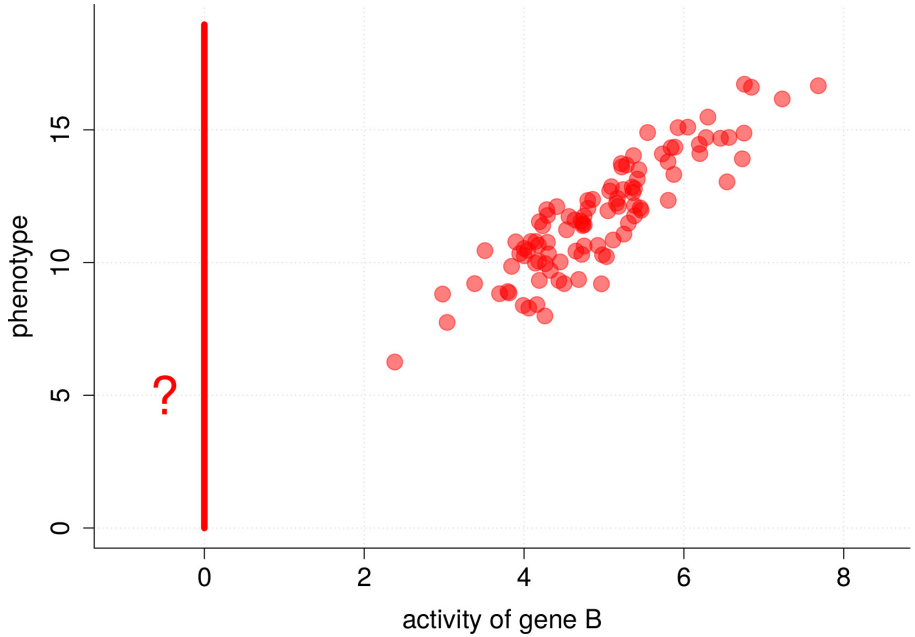
  Imbens & Rubin: Causal Inference for Statistics, Cambridge Univ. Press 2015,

  Pearl: Causality, Cambridge Univ. Press 2009,

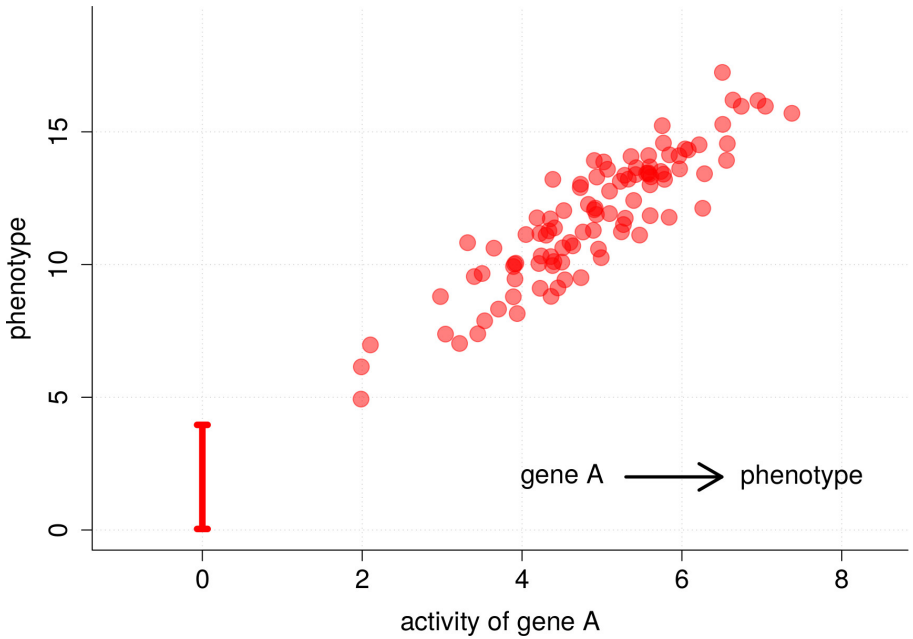
  ... and others.

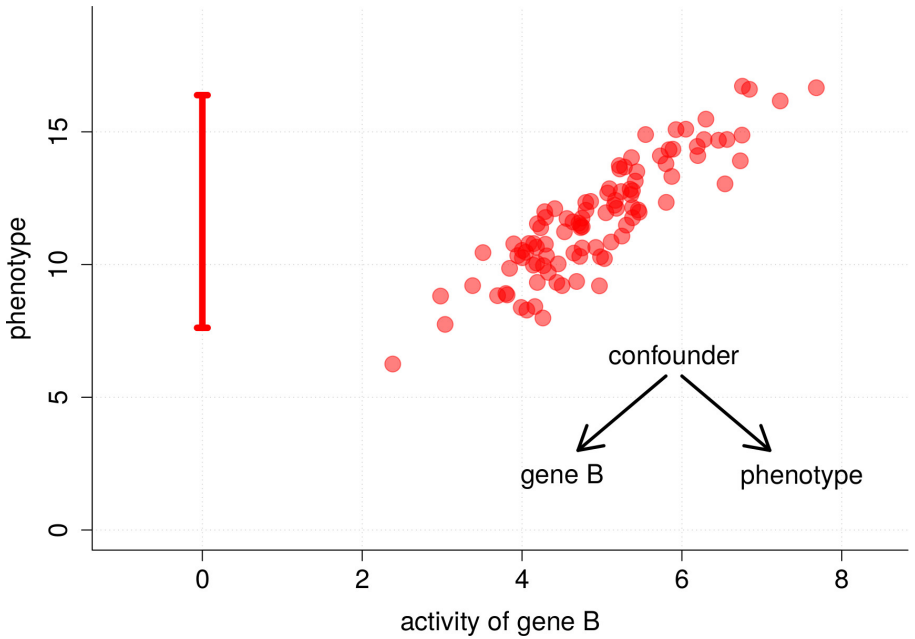
Consider the following problem.



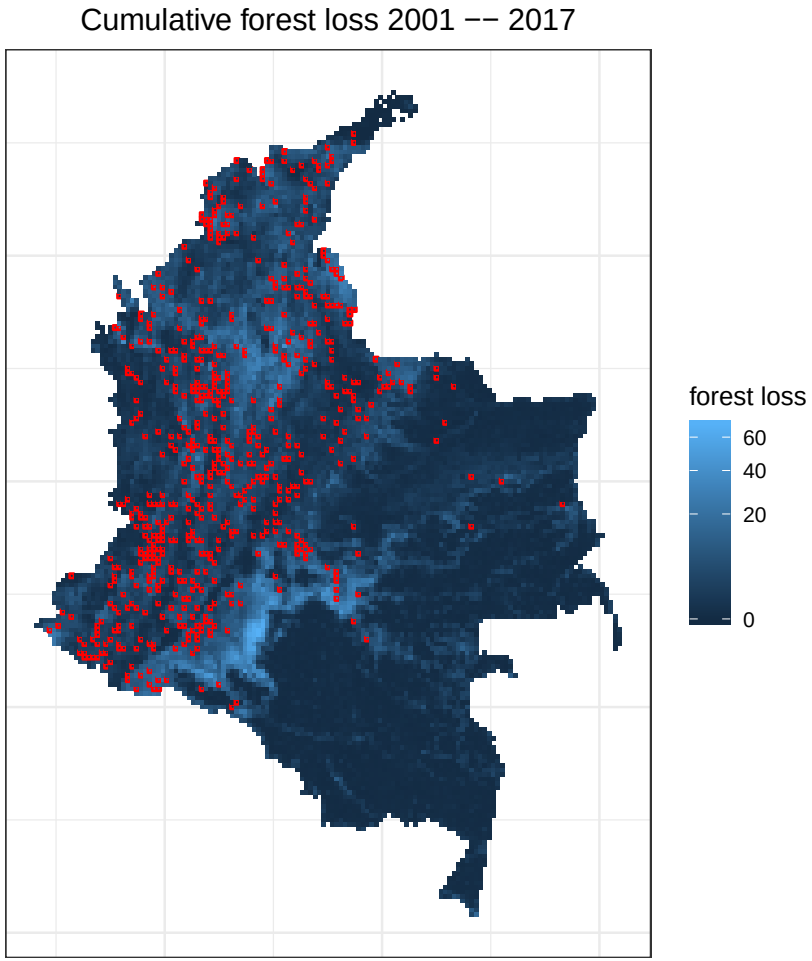


Causality matters!





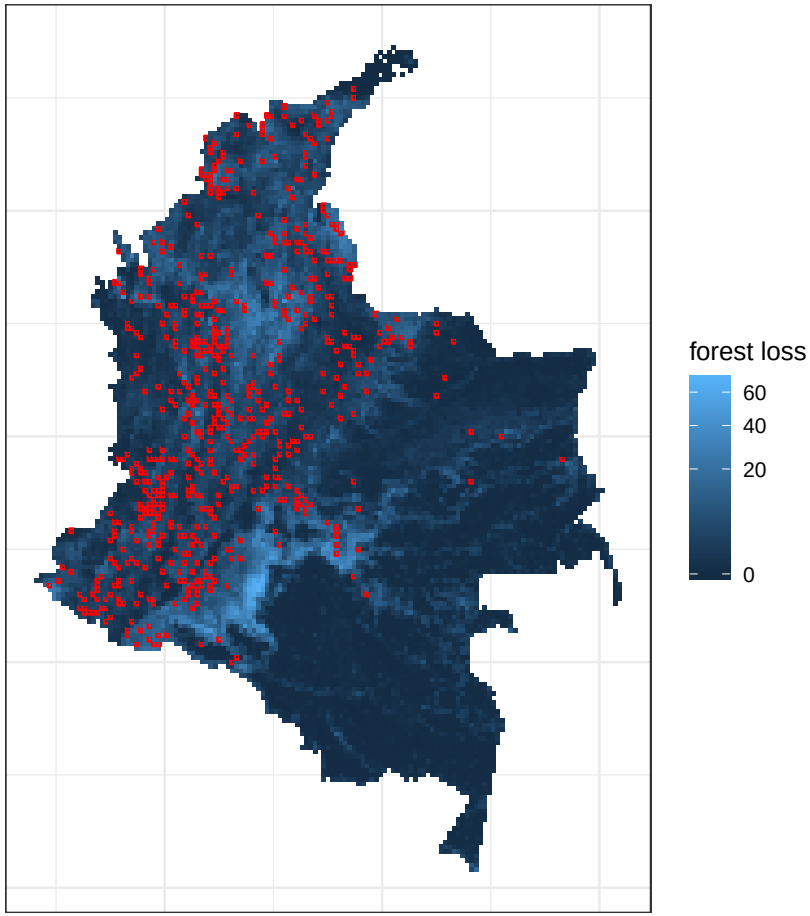
Example: forest loss



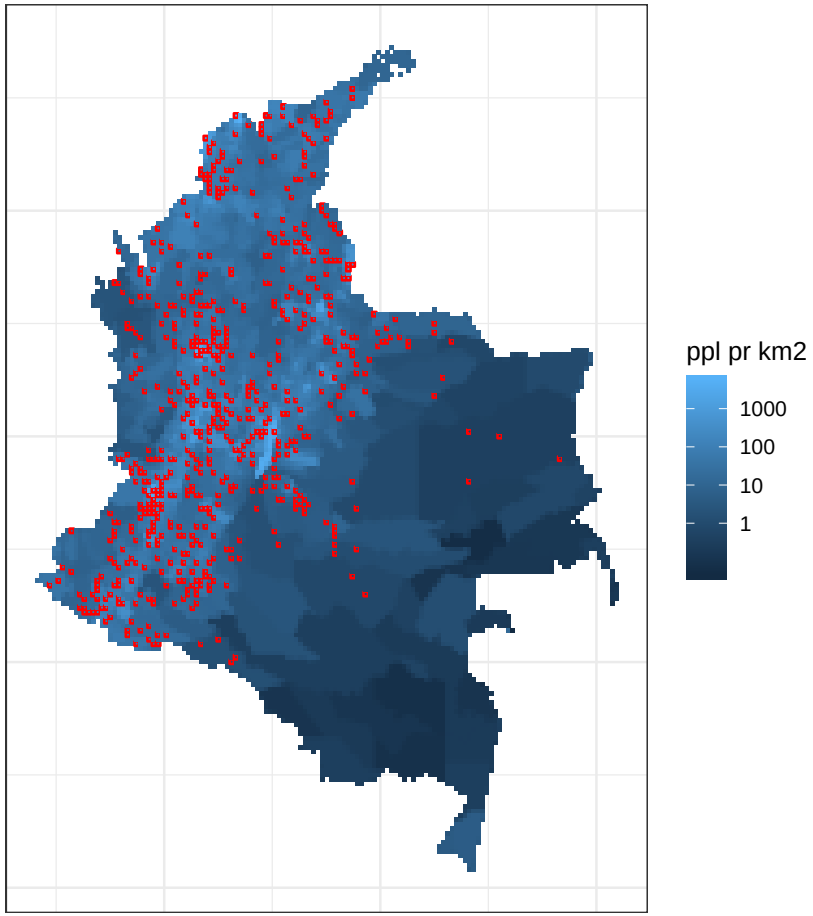
Example: forest loss

Colombia

Cumulative forest loss 2001 −− 2017

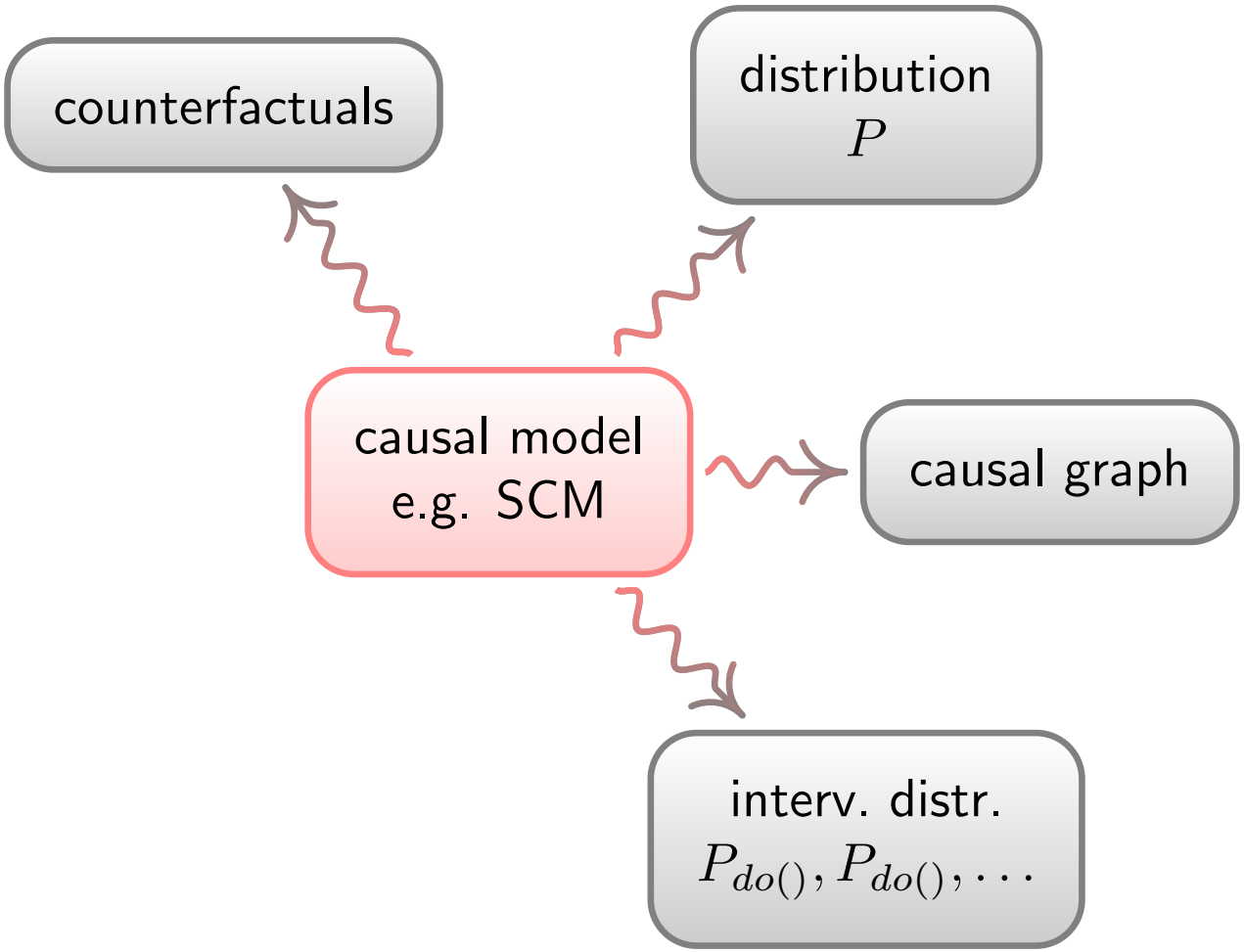


Population density in 2000



Christiansen et al. 2019

What is a causal model?



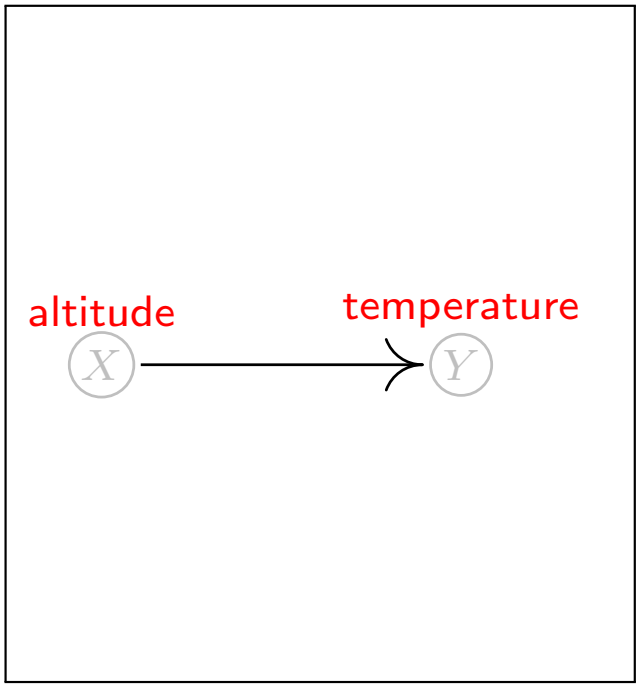
Part I: Causal Models

Example: Two variables

SCMs model observational distributions.

$$X := N_X$$
$$Y := -6X + N_Y$$

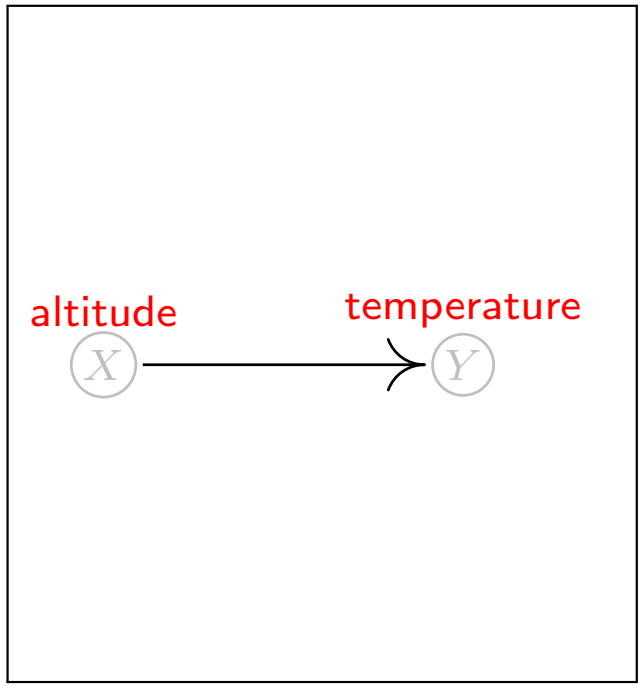
$$N_X, N_Y \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$$



Example: Two variables

SCMs model observational distributions.

$$X := N_X$$
$$Y := -6X + N_Y$$
$$N_X, N_Y \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$$



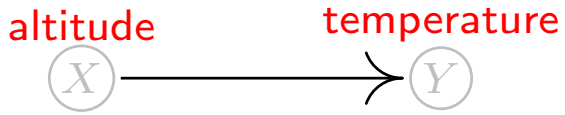
$$P: \qquad (X,Y) \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & -6 \\ -6 & 37 \end{pmatrix}\right)$$

Example: Two variables

SCMs model interventions, too.

$$X := N_X \quad X := 3$$
$$Y := -6X + N_Y$$

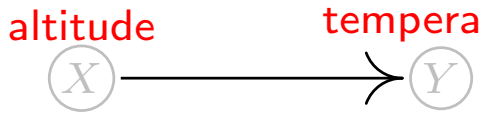
$$N_X, N_Y \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$$



Example: Two variables

SCMs model interventions, too.

$$X := N_X \quad X := 3$$
$$Y := -6X + N_Y$$
$$N_X, N_Y \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$$



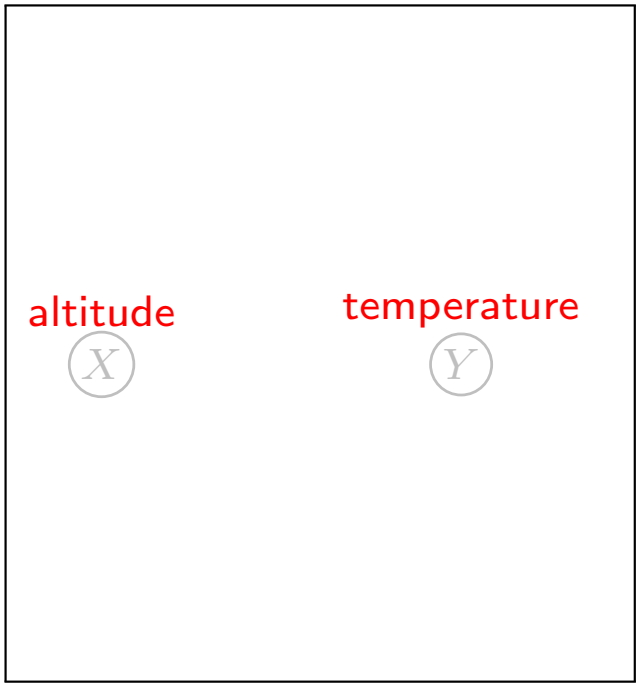
$P_{do(X:=3)}$: $P_{do(X:=3)}(X = 3) = 1$ and $Y \sim \mathcal{N}(-18, 1)$

Example: Two variables

SCMs model interventions, too.

$$X := N_X$$
$$Y := -6X + N_Y \quad Y := \mathcal{N}(2, 2)$$

$$N_X, N_Y \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$$

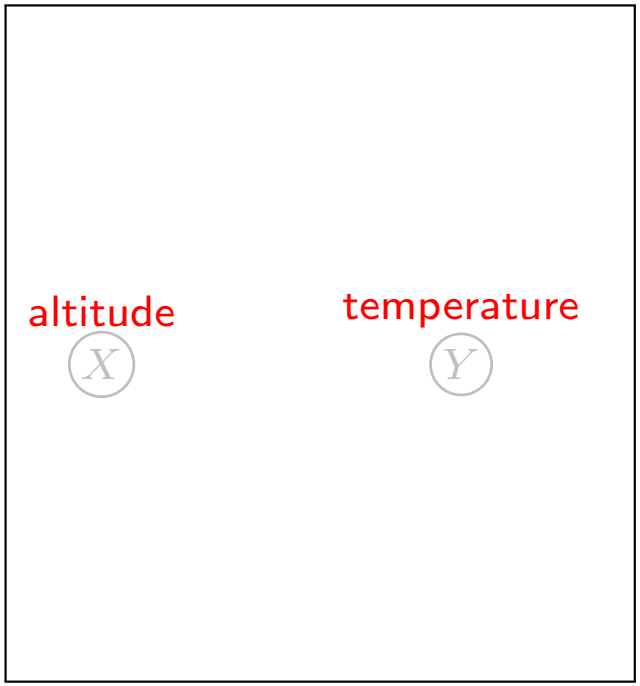


Example: Two variables

SCMs model interventions, too.

$$X := N_X$$
$$Y := -6X + N_Y \quad Y := \mathcal{N}(2,2)$$

$$N_X, N_Y \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$$

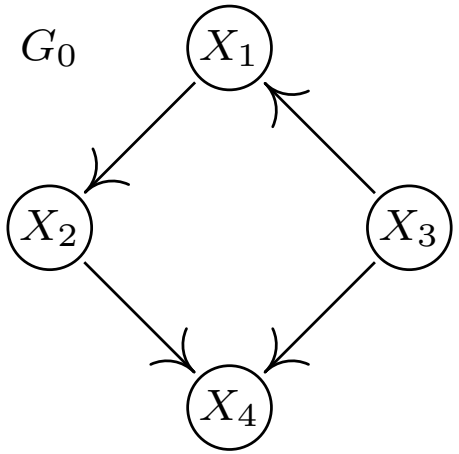


$$P_{do(Y:=\mathcal{N}(2,2))}: \qquad (X,Y) \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}\right)$$

SCMs model observational distributions over $X_1, \dots, X_d$. Call it: P.

$$
\begin{aligned}
X_1 &:= X_3 + N_1 \\
X_2 &:= 2X_1 + N_2 \\
X_3 &:= N_3 \\
X_4 &:= -X_2 - X_3 + N_4
\end{aligned}
$$

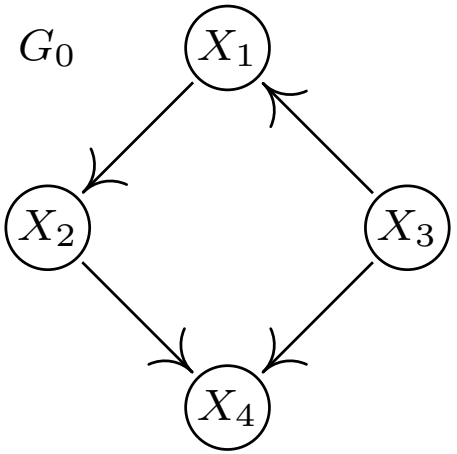
- N_i jointly independent $\mathcal{N}(0,1)$
- G_0 has no cycles



SCMs model observational distributions over $X_1, \dots, X_d$. Call it: P.

$$
\begin{aligned}
X_1 &:= X_3 + N_1 \\
X_2 &:= 2X_1 + N_2 \\
X_3 &:= N_3 \\
X_4 &:= -X_2 - X_3 + N_4
\end{aligned}
$$

- N_i jointly independent $\mathcal{N}(0,1)$
- G_0 has no cycles

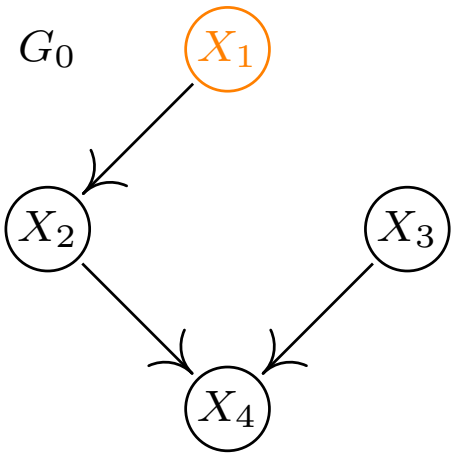


$$
\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix} = \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & 4 & 1 & -5 \\ 4 & 9 & 2 & -11 \\ 1 & 2 & 1 & -3 \\ -5 & -11 & -3 & 15 \end{pmatrix} \right)
$$

SCMs model interventions, too. Call it: $P_{do(X_1:=0)}$.

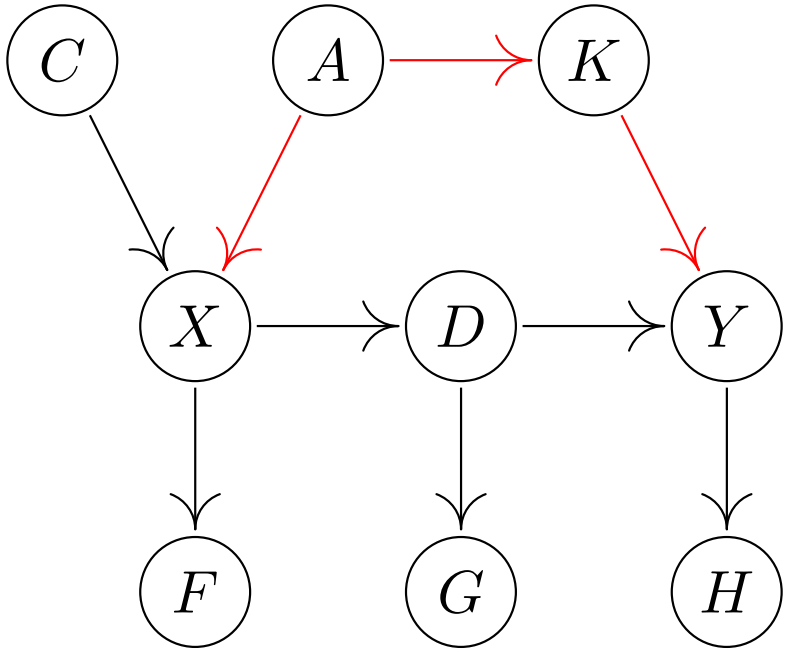
$$X_1 := 0$$
$$X_2 := f_2(X_1, N_2)$$
$$X_3 := f_3(N_3)$$
$$X_4 := f_4(X_2, X_3, N_4)$$

- N_i jointly independent
- G_0 has no cycles

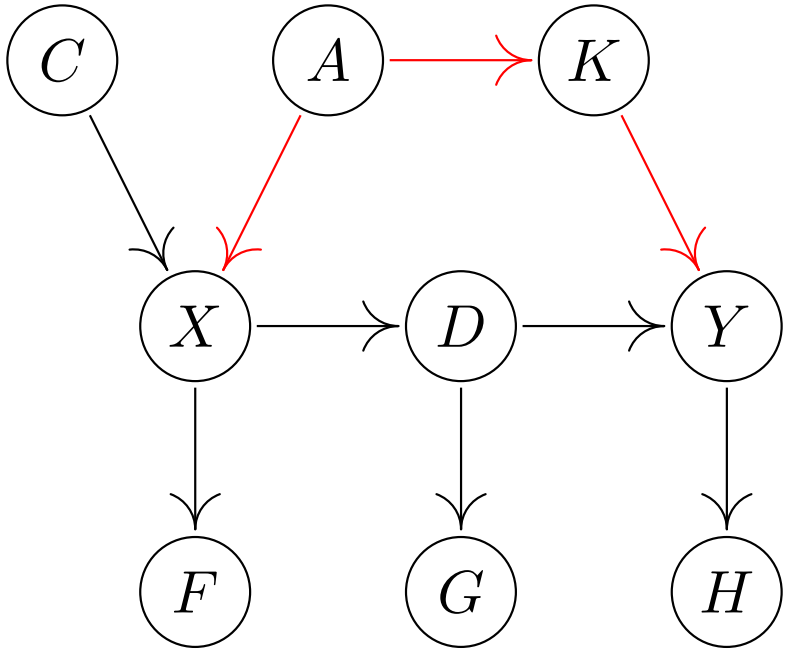


Exercise–SCM.R

Exercise–SCM.R

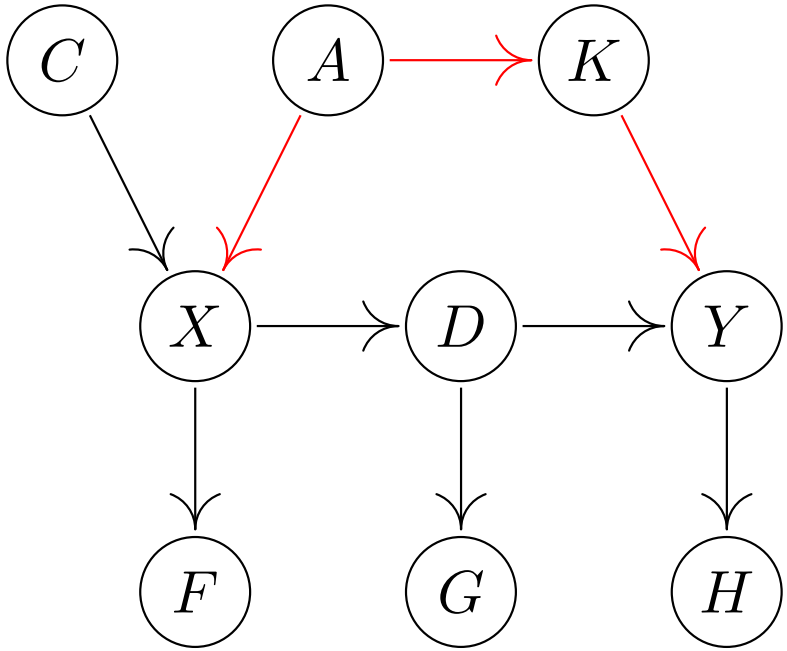


Exercise–SCM.R



Given: graph and P (or data). We want to compute the 'causal effect' from X to Y:

Exercise–SCM.R



Given: graph and P (or data). We want to compute the 'causal effect' from X to Y:

$$\frac{\partial}{\partial x} E_{do(X:=x)} Y.$$

If you intervene only on X_j, you intervene only on X_j (MUTE).

If you intervene only on X_j, you intervene only on X_j (MUTE).

MUTE: Most useful tautology ever.

Adjusting: compute int. distribution from obs. distribution and graph.
Adjusting in Linear Gaussian Models:

Adjusting: compute int. distribution from obs. distribution and graph.
Adjusting in Linear Gaussian Models:

Z is a valid adj. set (not defined here)

$\Rightarrow$ $\frac{\partial}{\partial x} E_{do(X:=x)} Y$ can be obtained by regressing Y on X, Z and taking coefficient of X.

Adjusting: compute int. distribution from obs. distribution and graph.
Adjusting in Linear Gaussian Models:

Z is a valid adj. set (not defined here)

$\Rightarrow$ $\frac{\partial}{\partial x} E_{do(X:=x)} Y$ can be obtained by regressing Y on X, Z and taking coefficient of X.

- Assume $Y \notin PA(X)$. Then $PA(X)$ is a valid adjustment set for (X, Y).

Adjusting: compute int. distribution from obs. distribution and graph.
Adjusting in Linear Gaussian Models:

Z is a valid adj. set (not defined here)

$\Rightarrow$ $\frac{\partial}{\partial x} E_{do(X:=x)} Y$ can be obtained by regressing Y on X, Z and taking coefficient of X.

- Assume $Y \notin PA(X)$. Then $PA(X)$ is a valid adjustment set for (X, Y).
- Assume Z blocks all backdoor paths from X to Y (a backdoor path starts with $X \leftarrow ...$). Then Z is a valid adjustment set for (X, Y).

Adjusting: compute int. distribution from obs. distribution and graph.
Adjusting in Linear Gaussian Models:

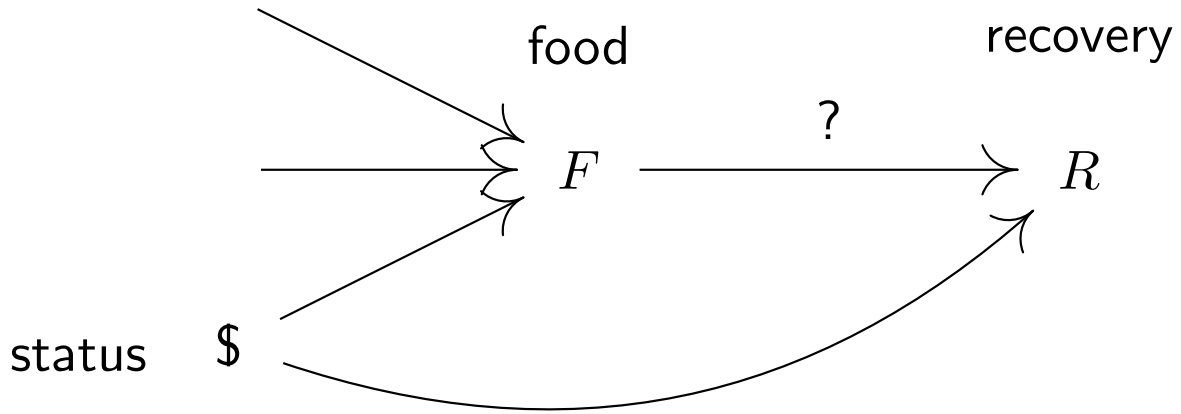
$$\begin{aligned} & Z \text{ is a valid adj. set (not defined here)} \\ \Rightarrow \quad & \frac{\partial}{\partial x} E_{do(X:=x)} Y \text{ can be obtained by regressing } Y \text{ on } X, Z \\ & \qquad\qquad \text{and taking coefficient of } X. \end{aligned}$$

- Assume $Y \notin PA(X)$. Then $PA(X)$ is a valid adjustment set for (X, Y).
- Assume Z blocks all backdoor paths from X to Y (a backdoor path starts with $X \leftarrow ...$). Then Z is a valid adjustment set for (X, Y).
- Alternative: on each directed path multiply coefficients; sum over directed paths.

Exercise-Adjusting

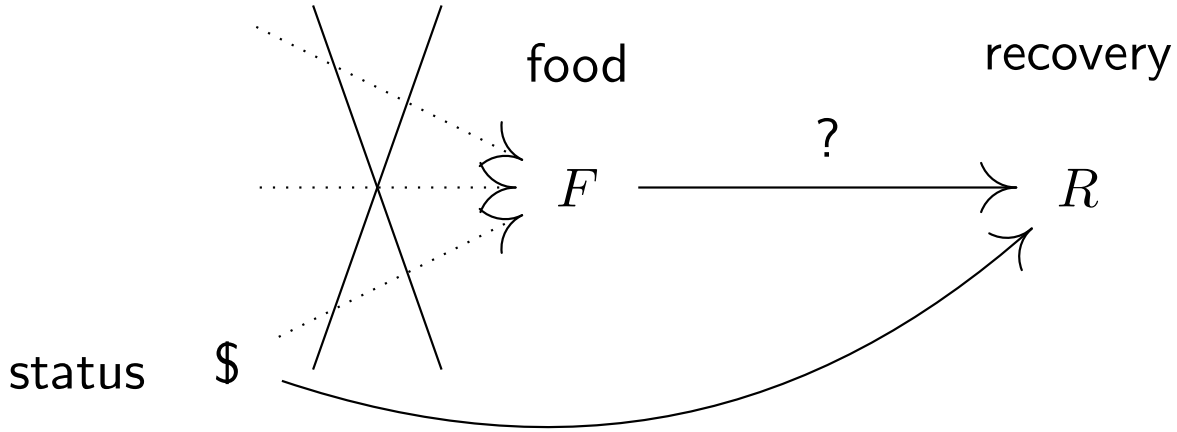
James Lind (1716–94):

James Lind (1716–94):
Causal relationship unclear.



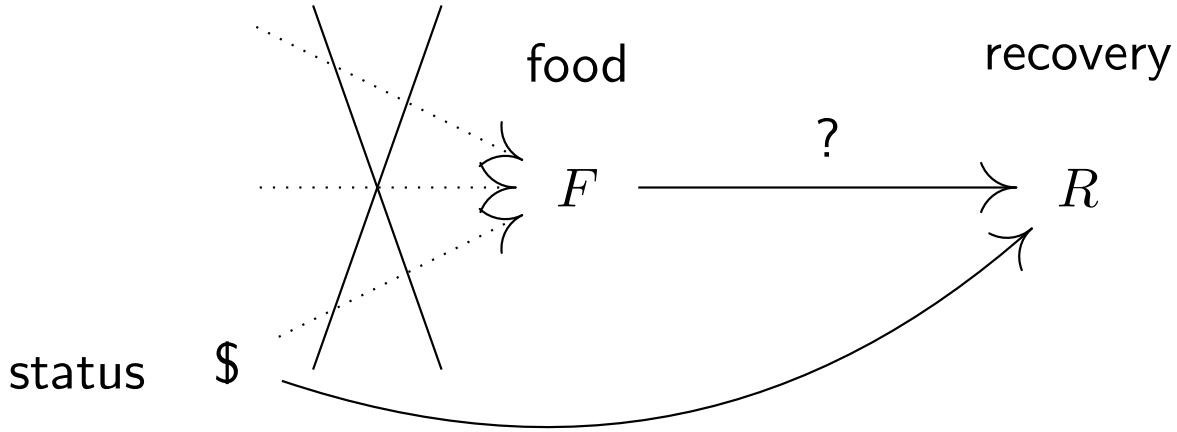
James Lind (1716–94):

Randomize! F and R dependent $\Longrightarrow$ there is a causal link!



James Lind (1716–94):

Randomize! F and R dependent $\Longrightarrow$ there is a causal link!



"On the 20th of May 1747, I selected twelve patients in the scurvy, on board the Salisbury [...] Two were ordered each a quart of cyder a day. Two others took twenty-five drops of elixir vitriol three times a day [...] Two others took two spoonfuls of vinegar three times a day [...] Two of the worst patients were put on a course of sea-water [...] Two others had each two oranges and one lemon given them every day [...] The two remaining patients, took [...] an electary recommended by a [...] surgeon [...] The consequence was, that the most sudden and visible good effects were perceived from the use of oranges and lemons;"

Example: smoking

BRITISH MEDICAL JOURNAL

LONDON SATURDAY SEPTEMBER 30 1950

SMOKING AND CARCINOMA OF THE LUNG

PRELIMINARY REPORT

BY

RICHARD DOLL, M.D., M.R.C.P.

Member of the Statistical Research Unit of the Medical Research Council

AND

A. BRADFORD HILL, Ph.D., D.Sc.

Professor of Medical Statistics, London School of Hygiene and Tropical Medicine; Honorary Director of the Statistical Research Unit of the Medical Research Council

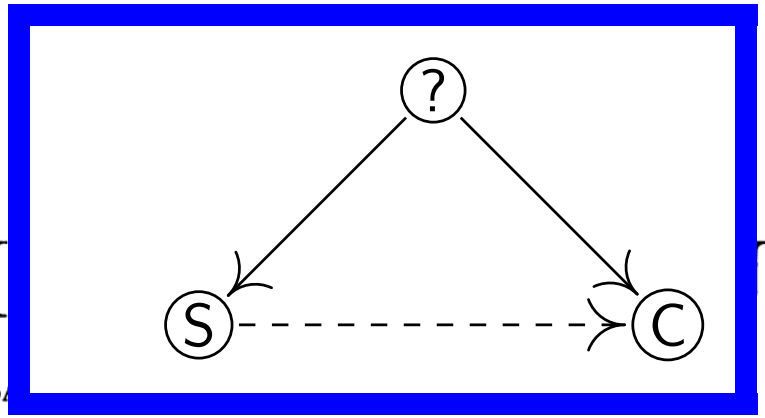
In England and Wales the phenomenal increase in the number of deaths attributed to cancer of the lung provides one of the most striking changes in the pattern of mortality recorded by the Registrar-General. For example, in the quarter of a century between 1922 and 1947 the annual number of deaths recorded increased from 612 to 9,287, an roughly fifteenfold. This remarkable increase is

whole explanation, although no one would deny that it may well have been contributory. As a corollary, it is right and proper to seek for other causes.

Possible Causes of the Increase

Two main causes have from time to time been put for-

Example: smoking



BRITISH M RNAL

LONDON S

SMOKING AND CARCINOMA OF THE LUNG

PRELIMINARY REPORT

BY

RICHARD DOLL, M.D., M.R.C.P.

Member of the Statistical Research Unit of the Medical Research Council

AND

A. BRADFORD HILL, Ph.D., D.Sc.

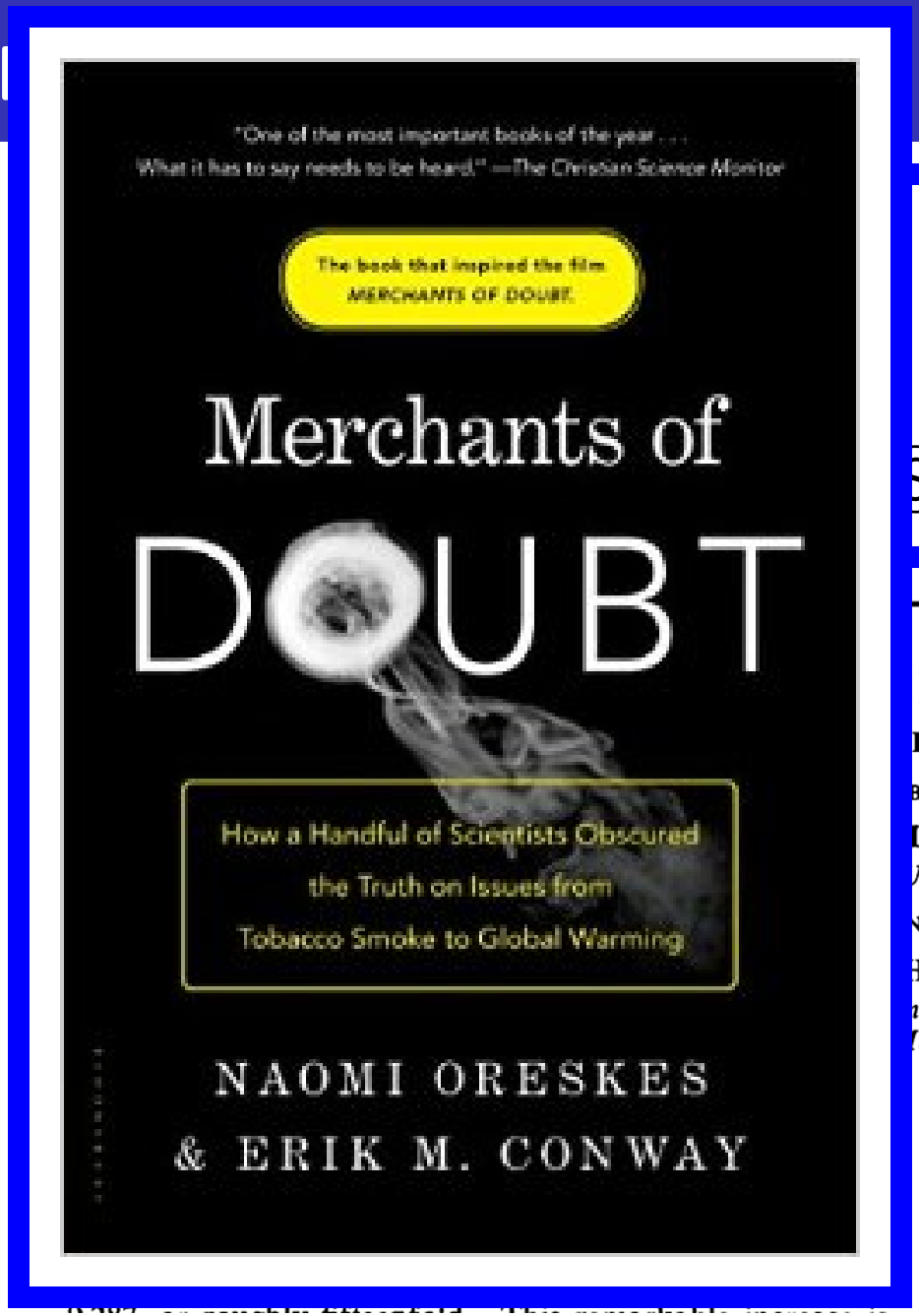
Professor of Medical Statistics, London School of Hygiene and Tropical Medicine; Honorary Director of the Statistical Research Unit of the Medical Research Council

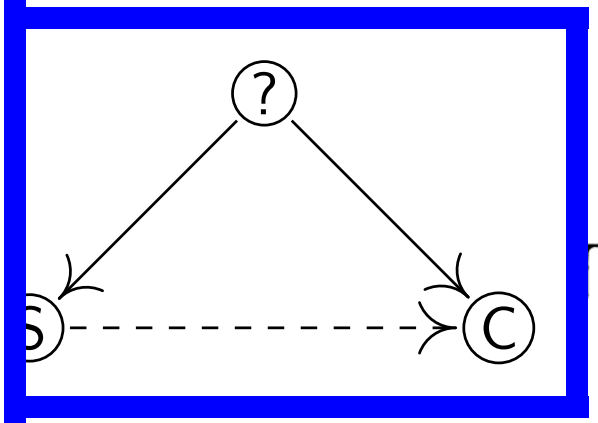
In England and Wales the phenomenal increase in the number of deaths attributed to cancer of the lung provides one of the most striking changes in the pattern of mortality recorded by the Registrar-General. For example, in the quarter of a century between 1922 and 1947 the annual number of deaths recorded increased from 612 to 9,287, an roughly fifteenfold. This remarkable increase is

whole explanation, although no one would deny that it may well have been contributory. As a corollary, it is right and proper to seek for other causes.

Possible Causes of the Increase

Two main causes have from time to time been put for-





RNAL

NOMA OF THE LUNG

RY REPORT

BY

L, M.D., M.R.C.P.

Unit of the Medical Research Council

ND

HILL, Ph.D., D.Sc.

d Tropical Medicine; Honorary Director of the Statistical
edical Research Council

whole explanation, although no one would deny that it may well have been contributory. As a corollary, it is right and proper to seek for other causes.

Possible Causes of the Increase

Two main causes have from time to time been put for-

Definition (Equivalence of causal models)

Two models are called

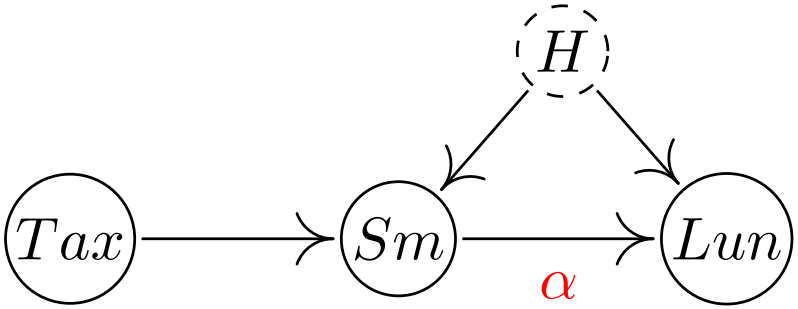
$$\{\textbf{probabilistically} \,/\, \textbf{interventionally}\} \ \textbf{equivalent}$$

if they entail the same

$$\{\text{observational} \,/\, \text{observational \& interventional}\}$$

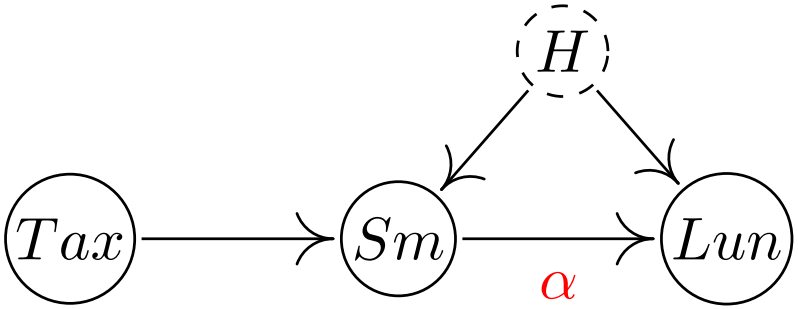
distributions. Here, it suffices to consider interventions that set a variable X_j to a fully supported $\tilde{N}_j$ ("randomized experiments").

Consider this graph



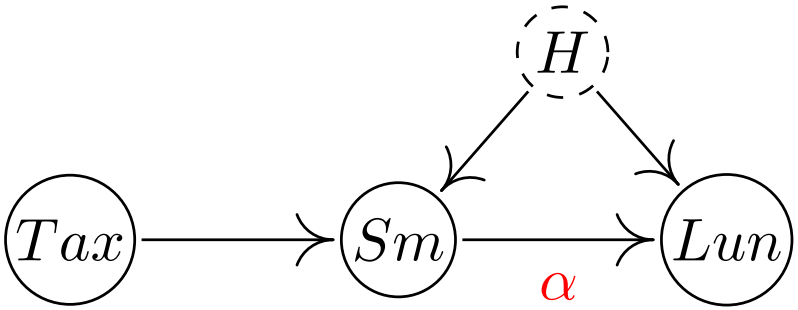
$$Lun = \alpha Sm + \beta H + N$$

Consider this graph



$$Lun = \alpha Sm + \beta H + N$$

An instrumental variable (here: tax) can fix the problem!



$$Lun = \alpha Sm + \beta H + N$$

Summary Part I:

- What if interested in iid prediction, i.e., observational data? Don't worry (too much) about causality!

Summary Part I:

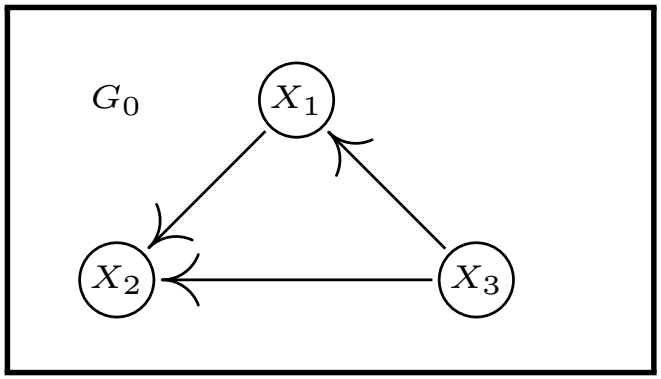
- What if interested in iid prediction, i.e., observational data? Don't worry (too much) about causality!
- But often, we are interested in a system's behaviour under intervention.

Summary Part I:

- What if interested in iid prediction, i.e., observational data? Don't worry (too much) about causality!
- But often, we are interested in a system's behaviour under intervention.
- SCMs entail graphs, obs. distr., interventions and counterfactuals.

$$X_1 := f_1(X_3, N_1)$$
$$X_2 := f_2(X_1, X_3, N_2)$$
$$X_3 := f_3(N_3)$$

- N_i jointly independent
- G_0 has no cycles

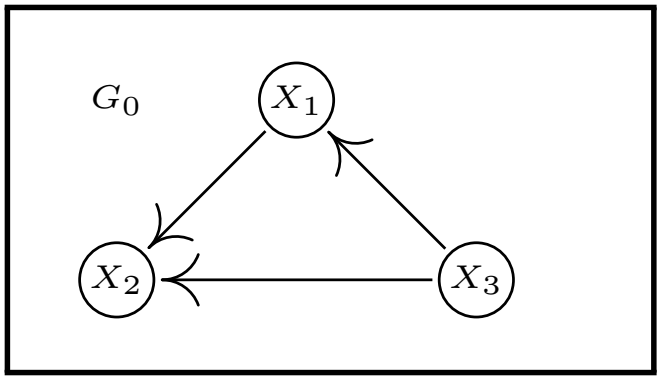


Summary Part I:

- What if interested in iid prediction, i.e., observational data? Don't worry (too much) about causality!
- But often, we are interested in a system's behaviour under intervention.
- SCMs entail graphs, obs. distr., interventions and counterfactuals.

$$X_1 := f_1(X_3, N_1)$$
$$X_2 := f_2(X_1, X_3, N_2)$$
$$X_3 := f_3(N_3)$$

- N_i jointly independent
- G_0 has no cycles



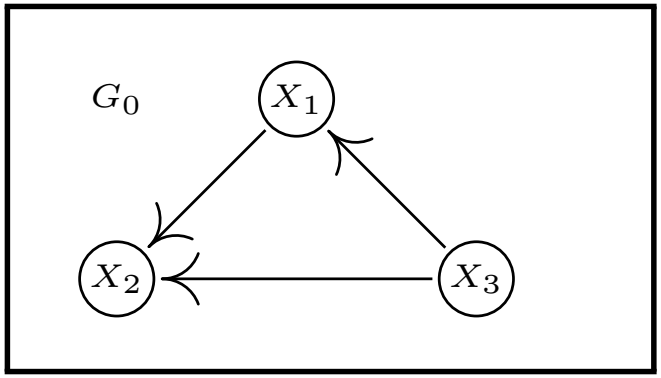
- adjusting: graph + observational distribution $\rightsquigarrow$ interventions

Summary Part I:

- What if interested in iid prediction, i.e., observational data? Don't worry (too much) about causality!
- But often, we are interested in a system's behaviour under intervention.
- SCMs entail graphs, obs. distr., interventions and counterfactuals.

$$X_1 := f_1(X_3, N_1)$$
$$X_2 := f_2(X_1, X_3, N_2)$$
$$X_3 := f_3(N_3)$$

- N_i jointly independent
- G_0 has no cycles



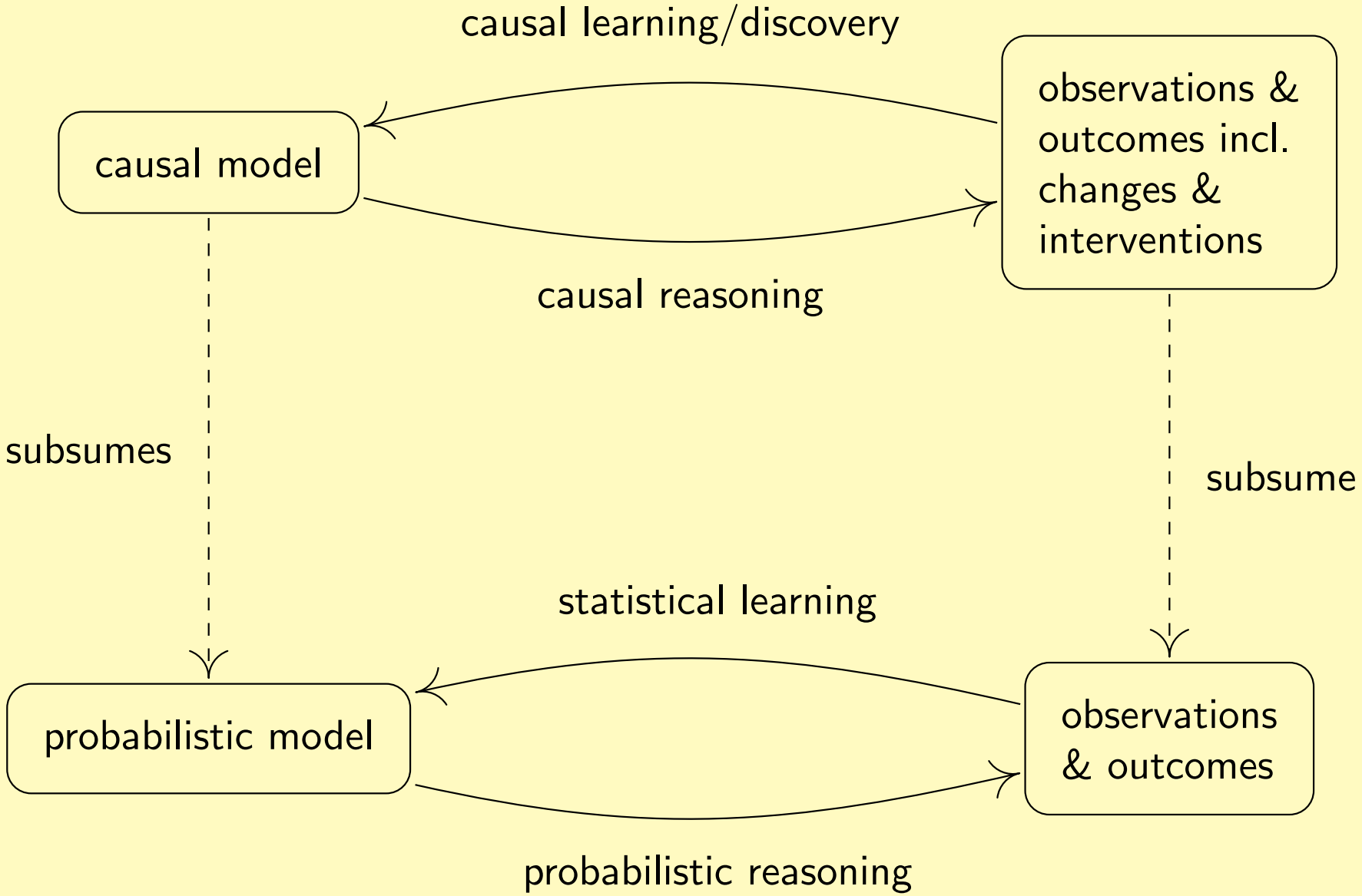
- adjusting: graph + observational distribution $\rightsquigarrow$ interventions
- instrumental variables: may help if there are hidden variables

Part II: Structure Learning or Causal Discovery

- There are different SCMs inducing the same obs. but different int. distributions.

- There are different SCMs inducing the same obs. but different int. distributions.
- Need assumptions! (such as faithfulness, restricted SCMs, ...)

- There are different SCMs inducing the same obs. but different int. distributions.
- Need assumptions! (such as faithfulness, restricted SCMs, ...)
- or heterogeneous data


causal learning/discovery
causal model
observations & outcomes incl. changes & interventions
causal reasoning
subsumes
subsume
statistical learning
probabilistic model
observations & outcomes
probabilistic reasoning

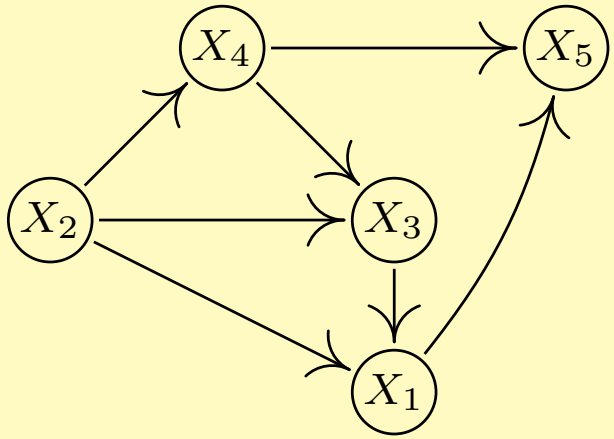
The Problem of Causal Discovery:

observed iid data from $P(X_1, \ldots, X_5)$

X_1	X_2	X_3	X_4	X_5
3.4	−0.3	5.8	−2.1	2.2
1.7	−0.2	7.0	−1.2	0.4
−2.4	−0.1	4.3	−0.7	3.5
2.3	−0.3	5.5	−1.1	−4.4
3.5	−0.2	3.9	−0.9	−3.9
⋮	⋮	⋮	⋮	⋮

$\xrightarrow{?}$

causal model, e.g. DAG $\mathcal{G}$



Correlation (Dependence) does not imply causation

Correlation (Dependence) does not imply causation ... but:

Correlation (Dependence) does not imply causation ... but:

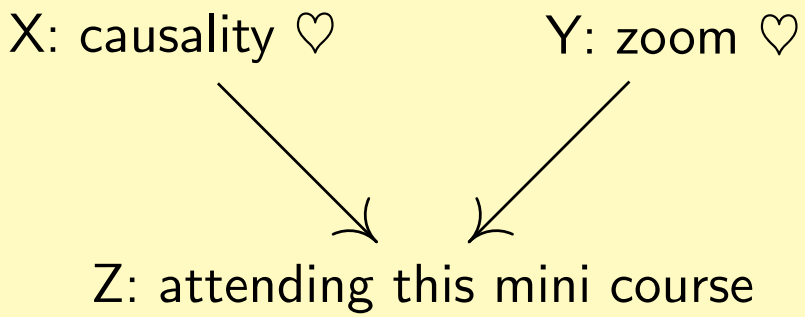
Reichenbach's common cause principle.
Assume that $X \not\perp\!\!\!\perp Y$. Then

- X "causes" Y,
- Y "causes" X,
- there is a hidden common "cause" or
- combination of the above.

Correlation (Dependence) does not imply causation ... but:

Reichenbach's common cause principle.
Assume that $X \not\perp\!\!\!\perp Y$. Then

- X "causes" Y,
- Y "causes" X,
- there is a hidden common "cause" or
- combination of the above.
- (In practice implicit conditioning also happens:

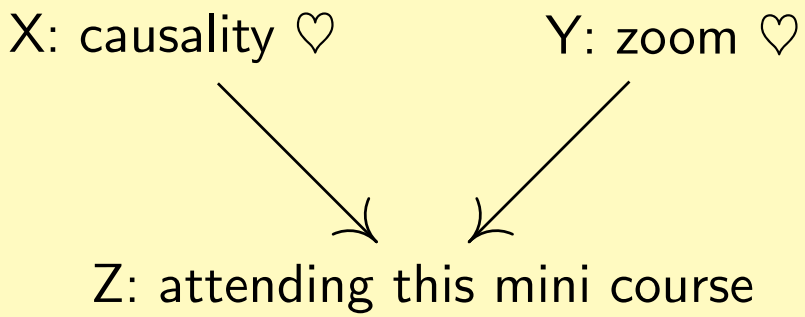


aka "selection bias").

Correlation (Dependence) does not imply causation ... but:

Reichenbach's common cause principle.
Assume that $X \not\perp\!\!\!\perp Y$. Then

- X "causes" Y,
- Y "causes" X,
- there is a hidden common "cause" or
- combination of the above.
- (In practice implicit conditioning also happens:



aka "selection bias"). Formalization of this idea...

Definition

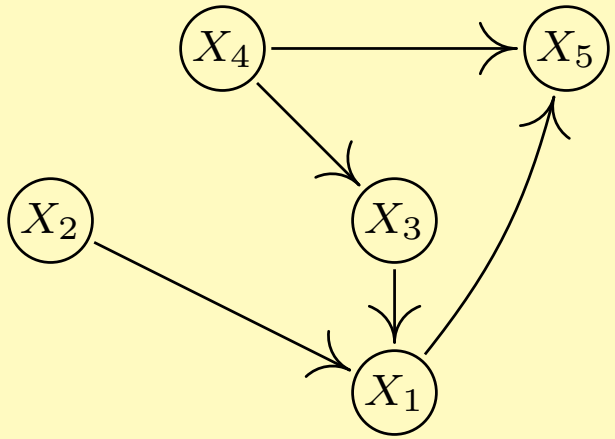
P is Markov w.r.t. G if

$$\underbrace{X \text{ and } Y \text{ are } d\text{-separated by } \mathcal{S} \text{ in } G}_{\text{properties of graph}} \quad \Rightarrow \quad \underbrace{X \perp\!\!\!\perp Y \,|\, \mathcal{S}}_{\text{properties in } P}$$

Definition: graphs

$G = (V, E)$ with $E \subseteq V \times V$. The rest is as in real life!

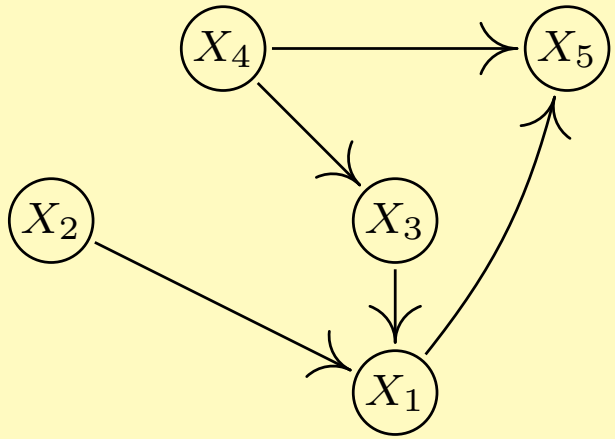
- parents, children, descendants, ancestors, ...
- paths, directed paths
- immoralities (or v-structures)
- d-separation (see next)
- ...



Definition: d-separation

X_i and X_j are d-separated by $\mathcal{S}$ if all paths between X_i and X_j are blocked by $\mathcal{S}$.

Check, whether all paths blocked!!

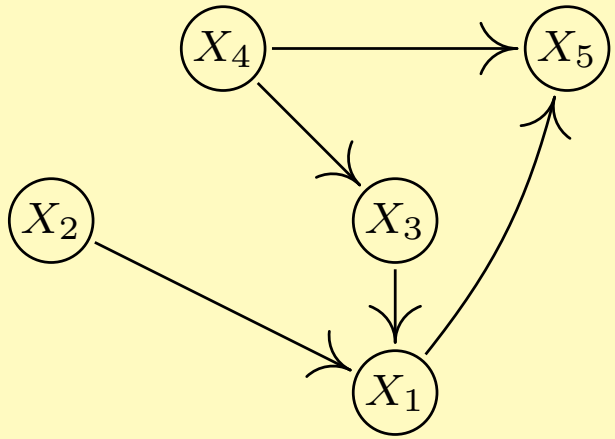


X_2 and X_5 are d-sep. by $\{X_1, X_4\}$
X_4 and X_1 are d-sep. by $\{X_2, X_3\}$
X_2 and X_4 are d-sep. by $\{\}$
X_4 and X_1 are NOT d-sep. by $\{X_3, X_5\}$

Definition: d-separation

X_i and X_j are d-separated by $\mathcal{S}$ if all paths between X_i and X_j are blocked by $\mathcal{S}$.

Check, whether all paths blocked!!



$\circ \cdots \rightarrow \circ \rightarrow \cdots \circ$ blocks a path.

$\circ \cdots \leftarrow \circ \rightarrow \cdots \circ$ blocks a path.

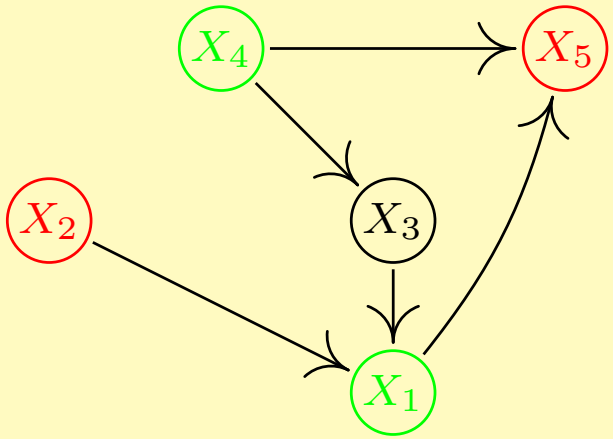
$\circ \cdots \rightarrow \circ \leftarrow \cdots \circ$ blocks a path.

X_2 and X_5 are d-sep. by $\{X_1, X_4\}$
X_4 and X_1 are d-sep. by $\{X_2, X_3\}$
X_2 and X_4 are d-sep. by $\{\}$
X_4 and X_1 are NOT d-sep. by $\{X_3, X_5\}$

Definition: d-separation

X_i and X_j are d-separated by $\mathcal{S}$ if all paths between X_i and X_j are blocked by $\mathcal{S}$.

Check, whether all paths blocked!!



$\circ \cdots \rightarrow \circ \rightarrow \cdots \circ$ blocks a path.

$\circ \cdots \leftarrow \circ \rightarrow \cdots \circ$ blocks a path.

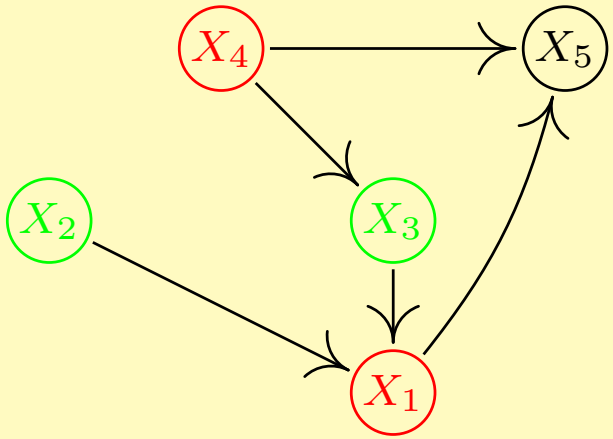
$\circ \cdots \rightarrow \circ \leftarrow \cdots \circ$ blocks a path.

X_2 and X_5 are d-sep. by $\{X_1, X_4\}$
X_4 and X_1 are d-sep. by $\{X_2, X_3\}$
X_2 and X_4 are d-sep. by $\{\}$
X_4 and X_1 are NOT d-sep. by $\{X_3, X_5\}$

Definition: d-separation

X_i and X_j are d-separated by $\mathcal{S}$ if all paths between X_i and X_j are blocked by $\mathcal{S}$.

Check, whether all paths blocked!!



$\circ \cdots \rightarrow \circ \rightarrow \cdots \circ$ blocks a path.

$\circ \cdots \leftarrow \circ \rightarrow \cdots \circ$ blocks a path.

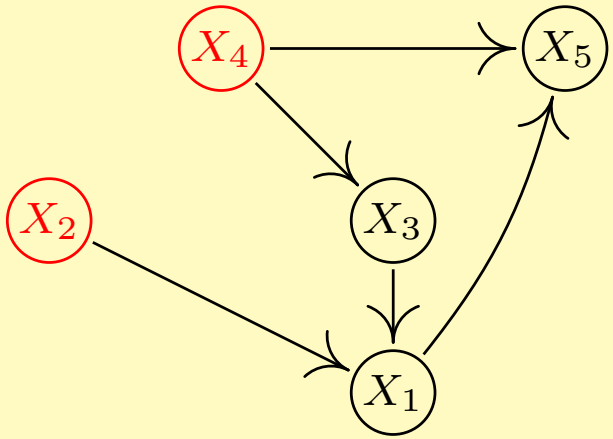
$\circ \cdots \rightarrow \circ \leftarrow \cdots \circ$ blocks a path.

X_2 and X_5 are d-sep. by $\{X_1, X_4\}$
X_4 and X_1 are d-sep. by $\{X_2, X_3\}$
X_2 and X_4 are d-sep. by $\{\}$
X_4 and X_1 are NOT d-sep. by $\{X_3, X_5\}$

Definition: d-separation

X_i and X_j are d-separated by $\mathcal{S}$ if all paths between X_i and X_j are blocked by $\mathcal{S}$.

Check, whether all paths blocked!!



$\circ \cdots \rightarrow \circ \rightarrow \cdots \circ$ blocks a path.

$\circ \cdots \leftarrow \circ \rightarrow \cdots \circ$ blocks a path.

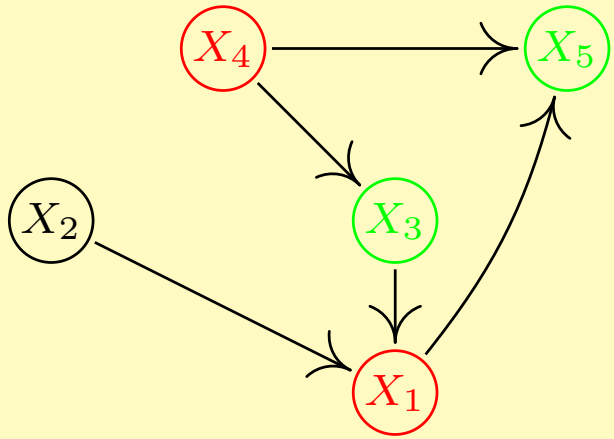
$\circ \cdots \rightarrow \circ \leftarrow \cdots \circ$ blocks a path.

X_2 and X_5 are d-sep. by $\{X_1, X_4\}$
X_4 and X_1 are d-sep. by $\{X_2, X_3\}$
X_2 and X_4 are d-sep. by $\{\}$
X_4 and X_1 are NOT d-sep. by $\{X_3, X_5\}$

Definition: d-separation

X_i and X_j are d-separated by $\mathcal{S}$ if all paths between X_i and X_j are blocked by $\mathcal{S}$.

Check, whether all paths blocked!!



$\circ \cdots \rightarrow \circ \rightarrow \cdots \circ$ blocks a path.

$\circ \cdots \leftarrow \circ \rightarrow \cdots \circ$ blocks a path.

$\circ \cdots \rightarrow \circ \leftarrow \cdots \circ$ blocks a path.

X_2 and X_5 are d-sep. by $\{X_1, X_4\}$
X_4 and X_1 are d-sep. by $\{X_2, X_3\}$
X_2 and X_4 are d-sep. by $\{\}$
X_4 and X_1 are NOT d-sep. by $\{X_3, X_5\}$

Definition

P satisfies the (global) Markov condition w.r.t. G if

$$\underbrace{X \text{ and } Y \text{ are } d\text{-separated by } \mathcal{S} \text{ in } G}_{\text{properties of graph}} \quad \Rightarrow \quad \underbrace{X \perp\!\!\!\perp Y \,|\, \mathcal{S}}_{\text{properties in } P}$$

Definition

P satisfies the (global) Markov condition w.r.t. G if

$$\underbrace{X \text{ and } Y \text{ are } d\text{-separated by } \mathcal{S} \text{ in } G}_{\text{properties of graph}} \quad \Rightarrow \quad \underbrace{X \perp\!\!\!\perp Y \,|\, \mathcal{S}}_{\text{properties in } P}$$

Proposition

Let the distribution P be Markov wrt a causal graph G. Then, Reichenbach's common cause principle is satisfied.

Proof: dependent variables must be d-connected.

Definition

P satisfies the (global) Markov condition w.r.t. G if

$$\underbrace{X \text{ and } Y \text{ are } d\text{-separated by } \mathcal{S} \text{ in } G}_{\text{properties of graph}} \quad \Rightarrow \quad \underbrace{X \perp\!\!\!\perp Y \,|\, \mathcal{S}}_{\text{properties in } P}$$

Definition

P satisfies the (global) Markov condition w.r.t. G if

$$\underbrace{X \text{ and } Y \text{ are } d\text{-separated by } \mathcal{S} \text{ in } G}_{\text{properties of graph}} \quad \Rightarrow \quad \underbrace{X \perp\!\!\!\perp Y \,|\, \mathcal{S}}_{\text{properties in } P}$$

Definition

P satisfies faithfulness w.r.t. G if

$$\underbrace{X \text{ and } Y \text{ are } d\text{-separated by } \mathcal{S} \text{ in } G}_{\text{properties of graph}} \quad \Leftarrow \quad \underbrace{X \perp\!\!\!\perp Y \,|\, \mathcal{S}}_{\text{properties in } P}$$

Idea 1: independence-based methods

Exercise-IndBased:

a) Assume $P^{(X,Y,Z)}$ is Markov and faithful wrt. G. Assume all(!) conditional independences are

$$X \perp\!\!\!\perp Z \,|\, \emptyset$$

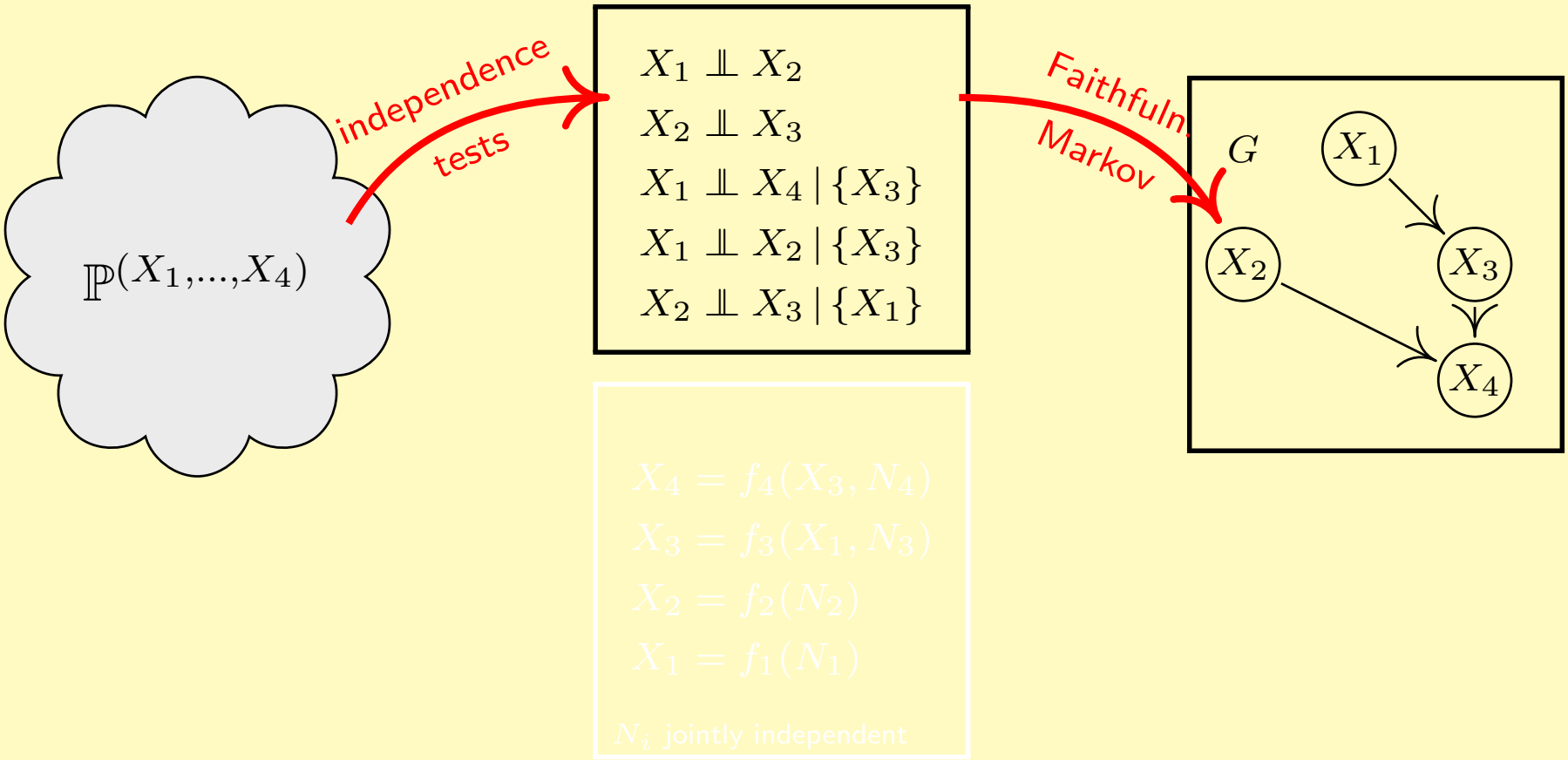
(plus symmetric statements). What is G?

b) Assume $P^{(W,X,Y,Z)}$ is Markov and faithful wrt. G. Assume all(!) conditional independences are

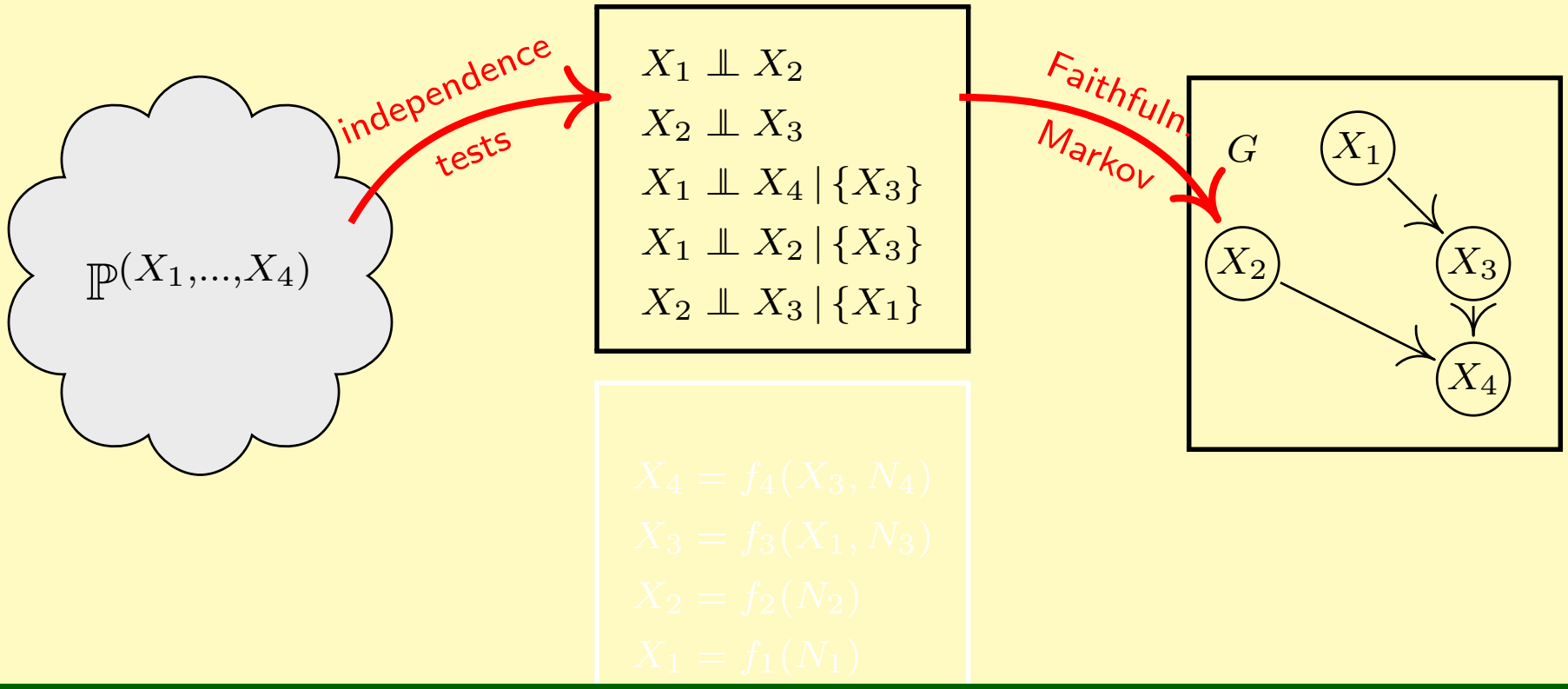
$$\begin{aligned} (Y,Z) &\perp\!\!\!\perp W \,|\, \emptyset \\ W &\perp\!\!\!\perp Y \,|\, (X,Z) \\ (W,X) &\perp\!\!\!\perp Y \,|\, Z \end{aligned}$$

(plus symmetric and trivially implied statements). What is G?

Idea 1: independence-based methods



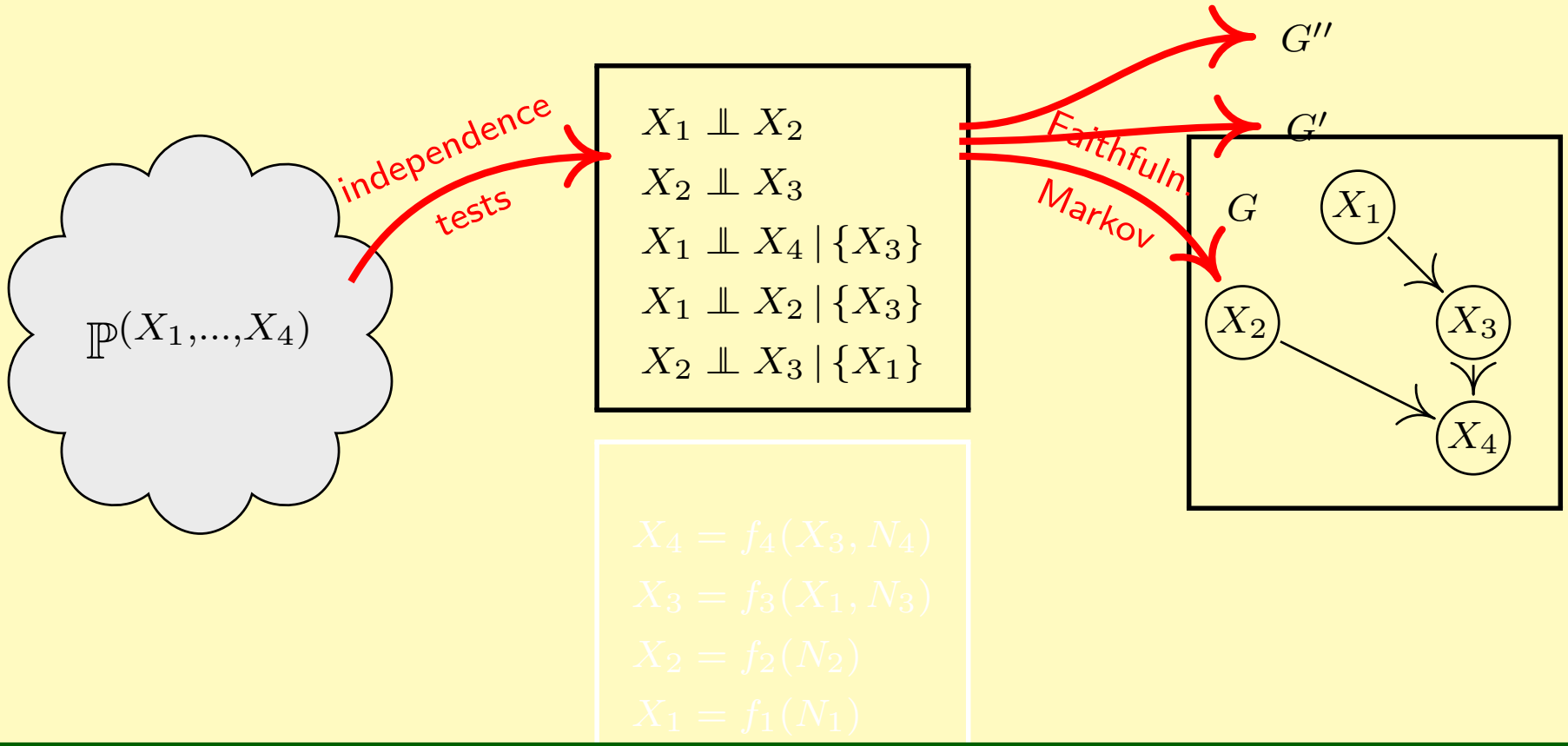
Idea 1: independence-based methods



Method: IC (Pearl 2009); PC, FCI (Spirtes et al., 2000)

1. Find all (cond.) independences from the data.
2. Select the DAG(s) that corresponds to these independences.

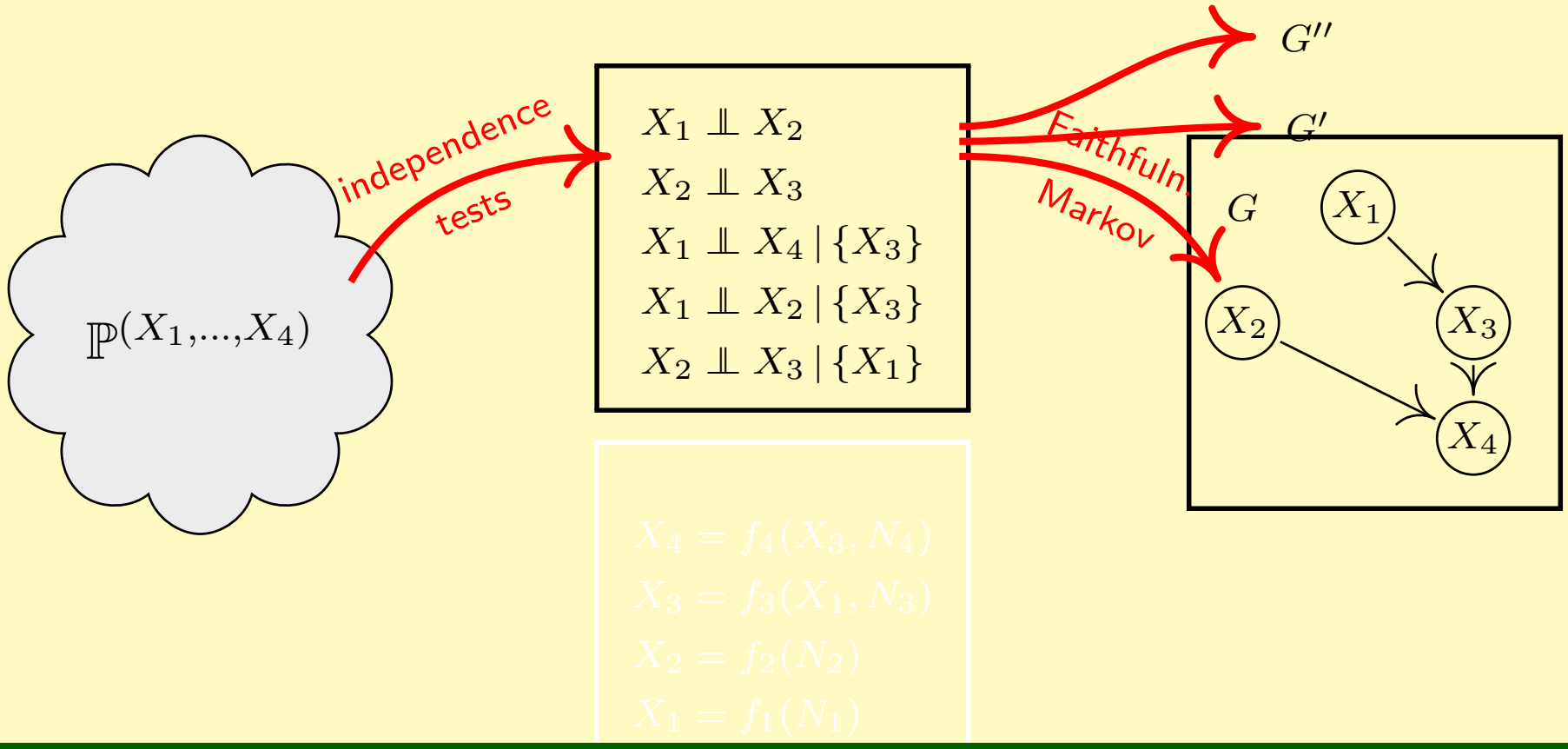
Idea 1: independence-based methods



Method: IC (Pearl 2009); PC, FCI (Spirtes et al., 2000)

1. Find all (cond.) independences from the data.
2. Select the DAG(s) that corresponds to these independences.

Idea 1: independence-based methods



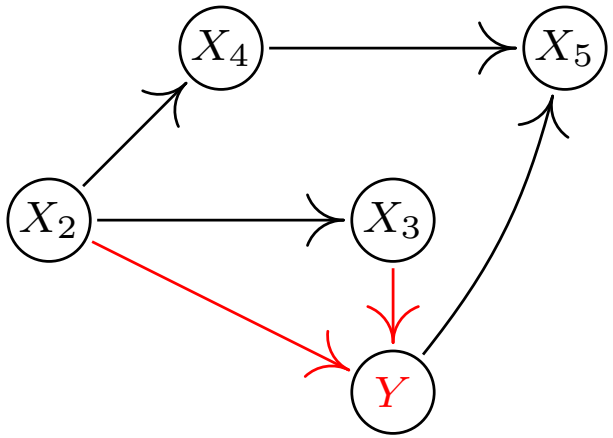
Method: ~~IC (Pearl 2009);~~ PC, FCI (Spirtes et al., 2000)

1. ~~Find all (cond.) independences from the data.~~ Be smart.
2. Select the DAG(s) that corresponds to these independences.

The Problem of Causal Discovery:

observed data $\xrightarrow{?}$ causal model

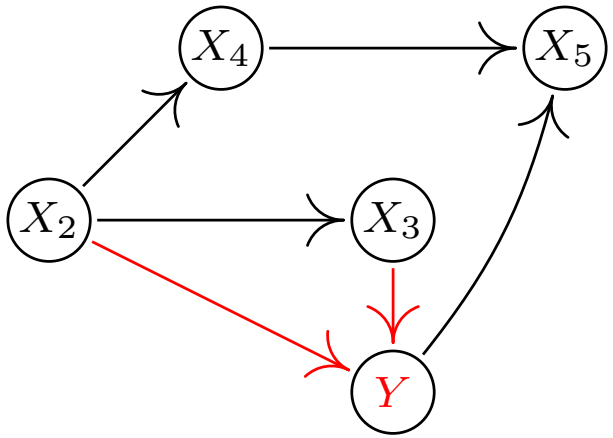
Y	X_2	X_3	X_4	X_5
3.4	−0.3	5.8	−2.1	2.2
1.7	−0.2	7.0	−1.2	0.4
−2.4	−0.1	4.3	−0.7	3.5
2.3	−0.3	5.5	−1.1	−4.4
3.5	−0.2	3.9	−0.9	−3.9
⋮	⋮	⋮	⋮	⋮



The Problem of Causal Discovery:

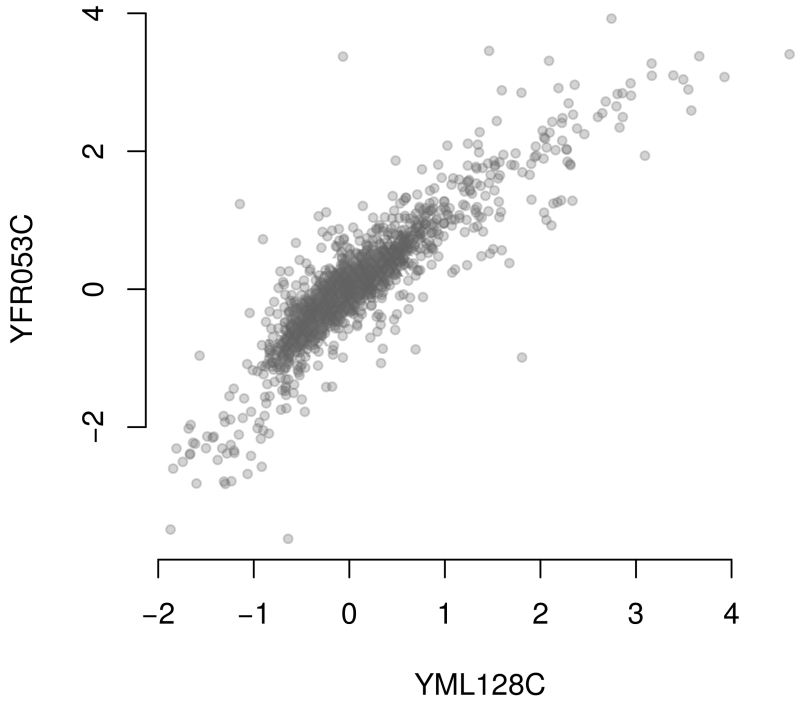
observed data $\xrightarrow{?}$ causal model

Y	X_2	X_3	X_4	X_5
3.4	−0.3	5.8	−2.1	2.2
1.7	−0.2	7.0	−1.2	0.4
−2.4	−0.1	4.3	−0.7	3.5
2.3	−0.3	5.5	−1.1	−4.4
3.5	−0.2	3.9	−0.9	−3.9
⋮	⋮	⋮	⋮	⋮



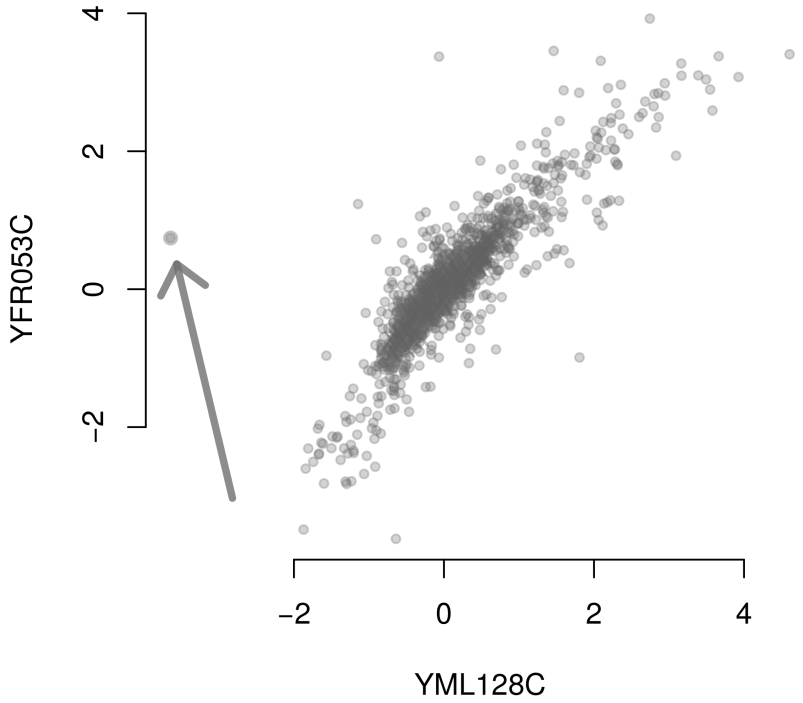
Here: Find the direct causes of Y!

Choose the predictor with the strongest correlation...



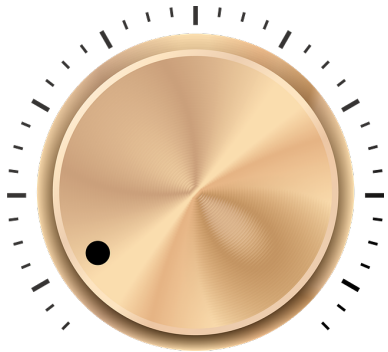
data from: Kemmeren et al. 2014

...and check the corresponding intervention on that predictor:



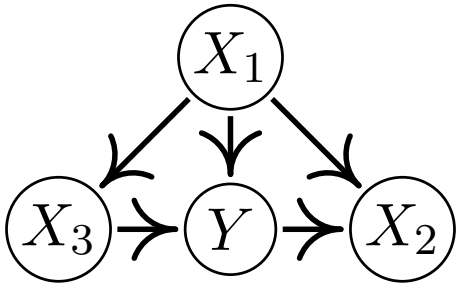
data from: Kemmeren et al. 2014

invariance  prediction

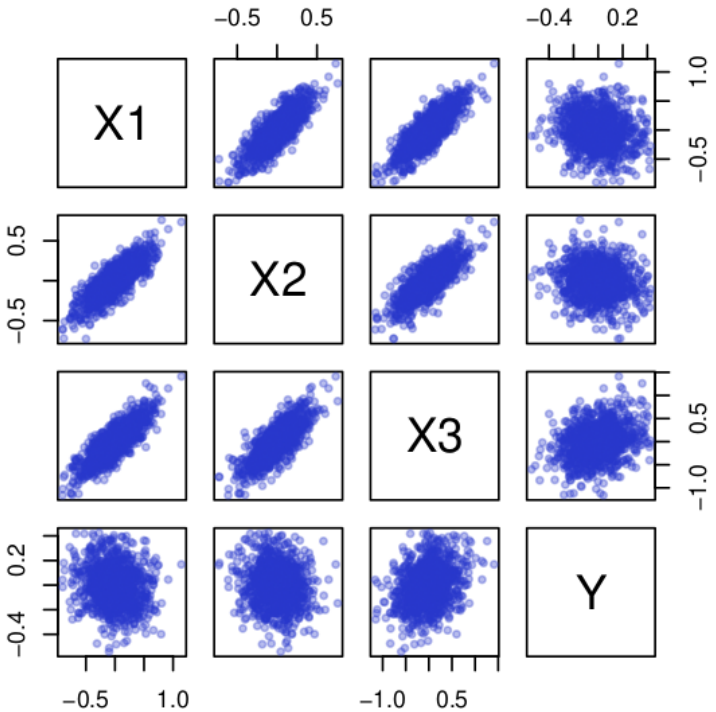
invariance  prediction

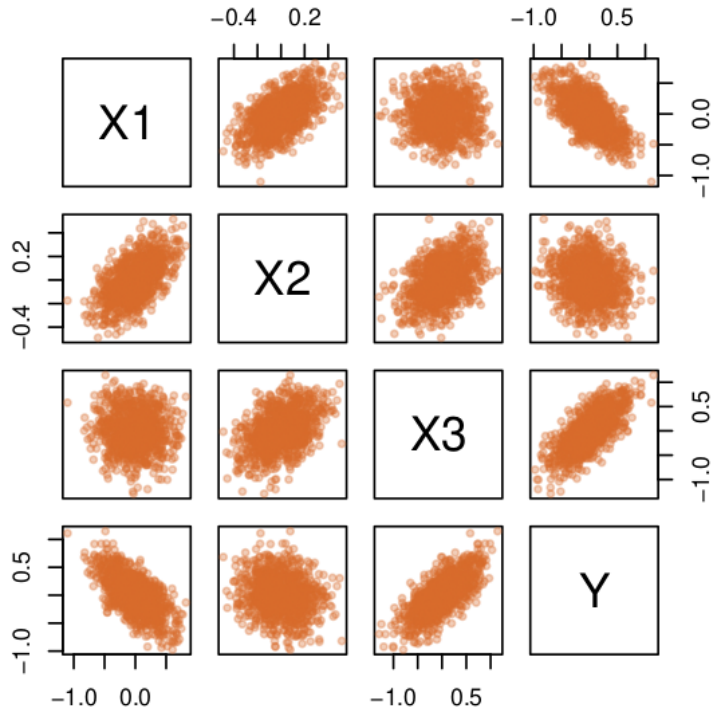
Invariant Causal Prediction

unknown:



known:





linear model

```
> linmod <- lm( Y ~ X)
> summary(linmod)


Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.305e-05  2.067e-03    0.04    0.968
X1          -5.490e-01  9.725e-03  -56.46   <2e-16 ***
X2          -4.078e-01  1.810e-02  -22.52   <2e-16 ***
X3           6.821e-01  6.896e-03   98.91   <2e-16 ***
```

ICP (R-package InvariantCausalPrediction)

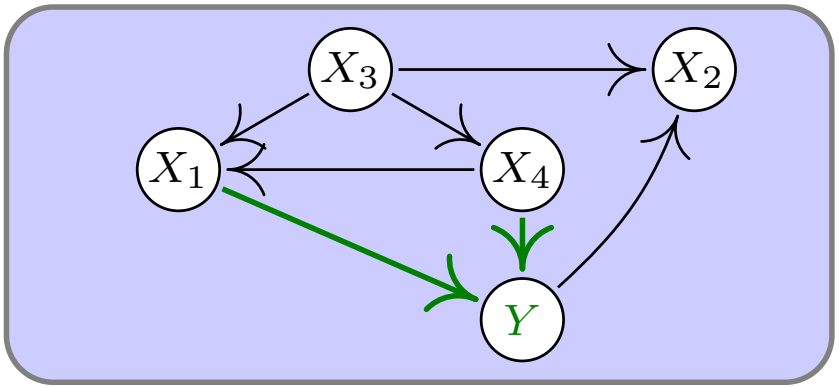
```
> ExpInd
```

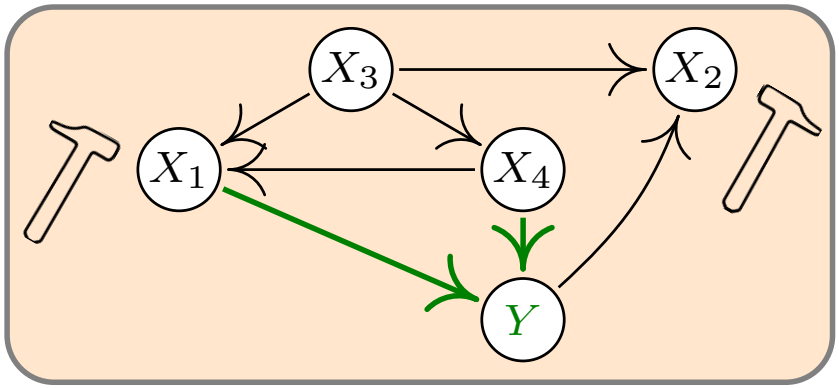
[1]1111111111111111111111111111111...2222222222222...

```
> icp <- ICP(X,Y,ExpInd)
```

```
            LOWER BOUND  UPPER BOUND  MAXIMIN EFFECT  P-VALUE
X1                -0.71        -0.52           -0.52   <1e-09 ***
X2                -0.46         0.00            0.00     0.55
X3                 0.58         0.70            0.58   <1e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

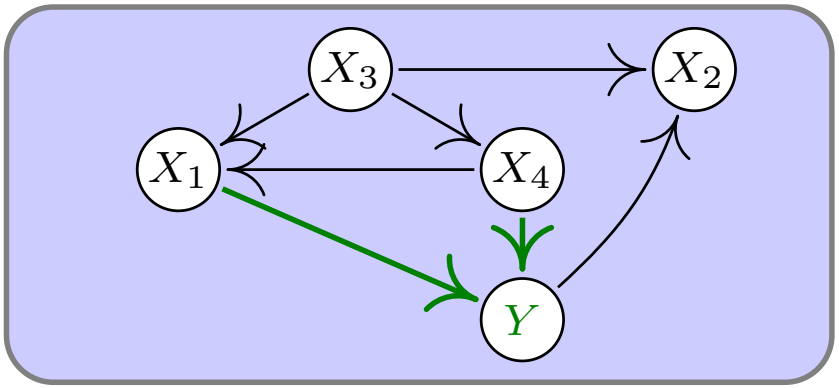
Fundamental assumption: $X_1, X_4 \to Y$ is invariant under interventions.

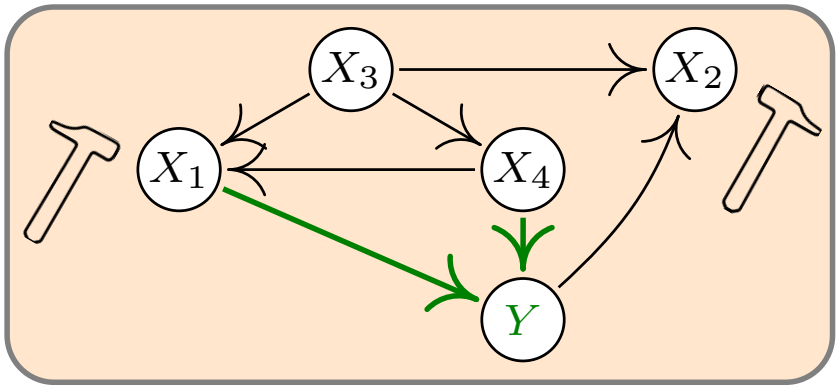




cf. modularity, autonomy, Haavelmo 1944, Aldrich 1989, Pearl 2009, ...

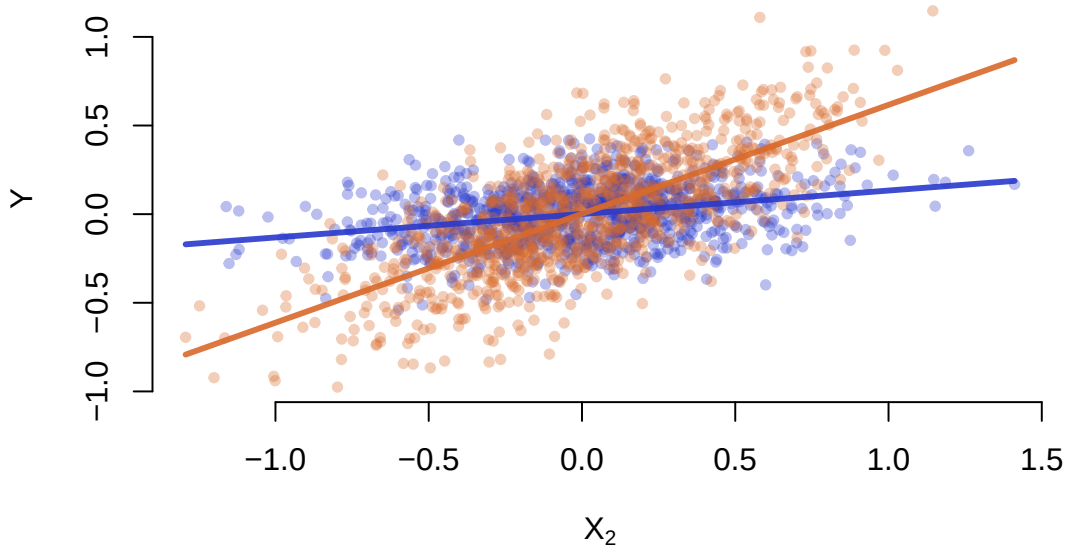
Fundamental assumption: $X_1, X_4 \to Y$ is invariant under interventions.



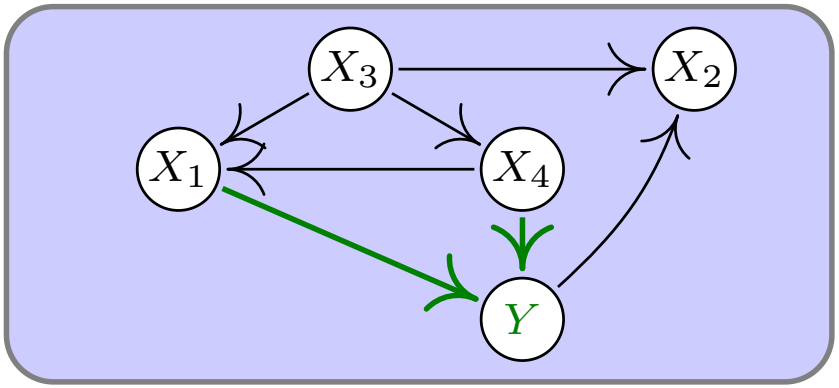


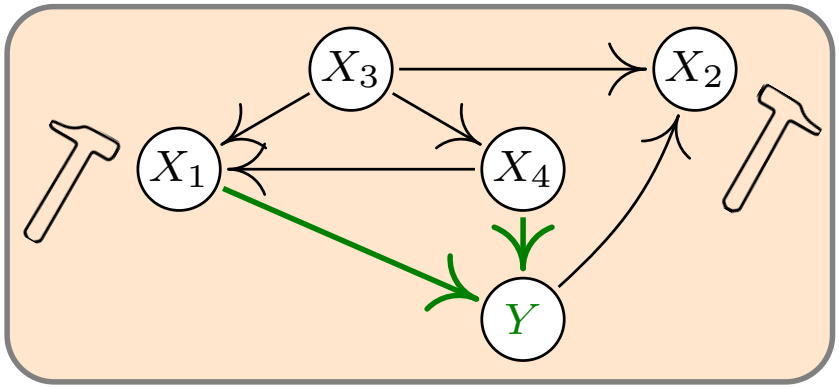
cf. modularity, autonomy, Haavelmo 1944, Aldrich 1989, Pearl 2009, ...

Not all sets of predictors yield an invariant model. Here: $\{2\}$.



Fundamental assumption: $X_1, X_4 \to Y$ is invariant under interventions.





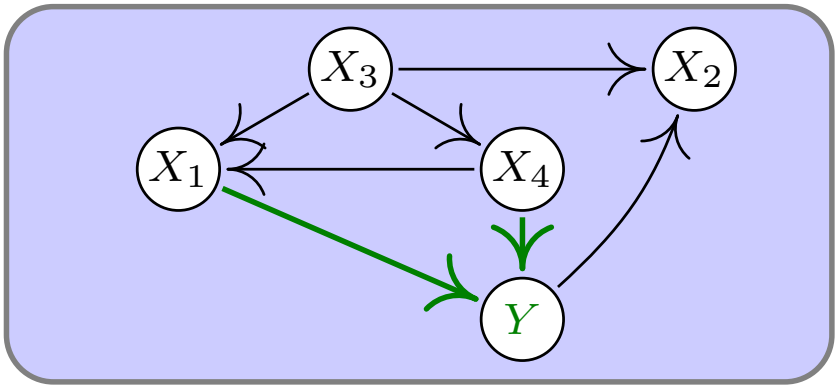
cf. modularity, autonomy, Haavelmo 1944, Aldrich 1989, Pearl 2009, ...

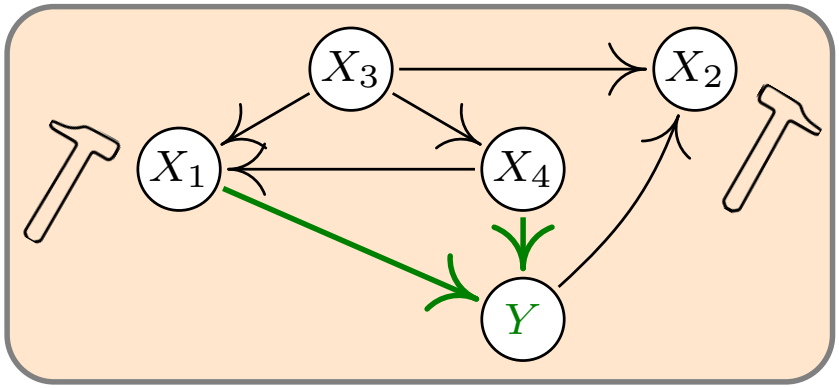
Key idea: Use and data and search for invariant models.

set	$\emptyset$	$\{1\}$	$\{2\}$	$\{3\}$	$\cdots$	$\{1,4\}$	$\{2,4\}$	$\cdots$	$\{1,3,4\}$
invariance	✗	✗	✗	✗	$\cdots$	✓	✗	$\cdots$	✓

$$\hat{S} := \bigcap_{S \text{ invariant}} S = \{1,4\}$$

Fundamental assumption: $X_1, X_4 \to Y$ is invariant under interventions.





cf. modularity, autonomy, Haavelmo 1944, Aldrich 1989, Pearl 2009, ...

Key idea: Use and data and search for invariant models.

set	$\emptyset$	$\{1\}$	$\{2\}$	$\{3\}$	$\cdots$	$\{1,4\}$	$\{2,4\}$	$\cdots$	$\{1,3,4\}$
invariance	✗	✗	✗	✗	$\cdots$	✓	✗	$\cdots$	✓

$$\hat{S} := \bigcap_{S \text{ invariant}} S = \{1,4\}$$

JP, Bühlmann, Meinshausen, JRSS-B 2016 (with discussion): $P(\hat{S} \subseteq S^*) \geq 1 - \alpha$. (ICP.ipynb)

Exercise-ICP.R

Given $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathbb{R}^d \times \mathbb{R}$ and environments $\mathcal{E}$.

Invariance $H_{0,S}$:

- for all $i = 1, \ldots, n$: $\quad Y_i = X_{S,i} \cdot \gamma + \varepsilon_i$.
- $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d.
- X_i can have an arbitrary distribution

Given $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathbb{R}^d \times \mathbb{R}$ and environments $\mathcal{E}$.

Invariance $H_{0,S}$:

- for all $i = 1, \ldots, n$: $\quad Y_i = X_{S,i} \cdot \gamma + \varepsilon_i$.
- $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d.
- X_i can have an arbitrary distribution

Environments $\mathcal{E}$ have elements
$e_1 = \{1, 2, 3, \ldots, 40\}$, $e_2 = \{41, \ldots, 100\}$, $e_3 = \{101, \ldots, n\}$, for example.

Given $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathbb{R}^d \times \mathbb{R}$ and environments $\mathcal{E}$.

Invariance $H_{0,S}$:

- for all $i = 1, \ldots, n$: $\quad Y_i = X_{S,i} \cdot \gamma + \varepsilon_i$.
- $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d.
- X_i can have an arbitrary distribution

Environments $\mathcal{E}$ have elements
$e_1 = \{1, 2, 3, \ldots, 40\}$, $e_2 = \{41, \ldots, 100\}$, $e_3 = \{101, \ldots, n\}$, for example.

Relation to causality:
Environments: different interventions (not on Y). Then, $H_{0,PA(Y)}$ holds.

cf. modularity, autonomy, Haavelmo 1944, Aldrich 1989, Pearl 2009, Schölkopf et al. 2012, ...

Given $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathbb{R}^d \times \mathbb{R}$ and environments $\mathcal{E}$.

Invariance $H_{0,S}$:

- for all $i = 1, \ldots, n$: $\quad Y_i = X_{S,i} \cdot \gamma + \varepsilon_i$.
- $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d.
- X_i can have an arbitrary distribution

Environments $\mathcal{E}$ have elements
$e_1 = \{1, 2, 3, \ldots, 40\}$, $e_2 = \{41, \ldots, 100\}$, $e_3 = \{101, \ldots, n\}$, for example.

Relation to causality:
Environments: different interventions (not on Y). Then, $H_{0,PA(Y)}$ holds.

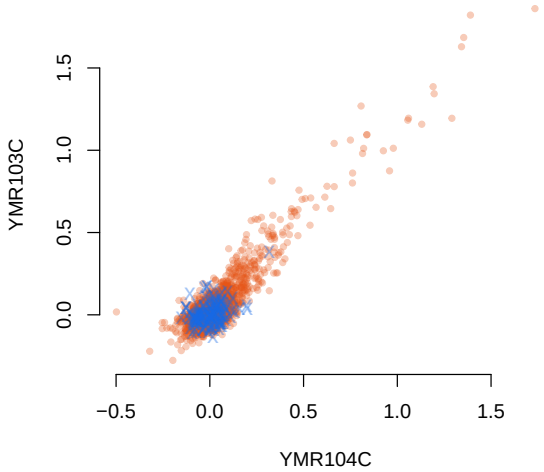
cf. modularity, autonomy, Haavelmo 1944, Aldrich 1989, Pearl 2009, Schölkopf et al. 2012, ...

Theorem (PBM 2016)

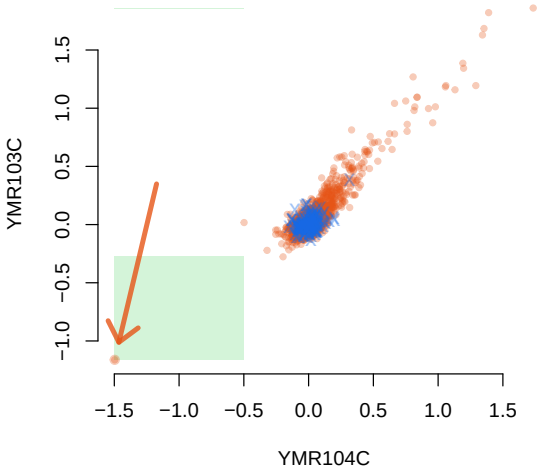
Assume H_{0,S^*} satisfied for some S^*. For any test level α we obtain

$$P(\hat{S} \subseteq S^*) \geq 1 - \alpha\,.$$

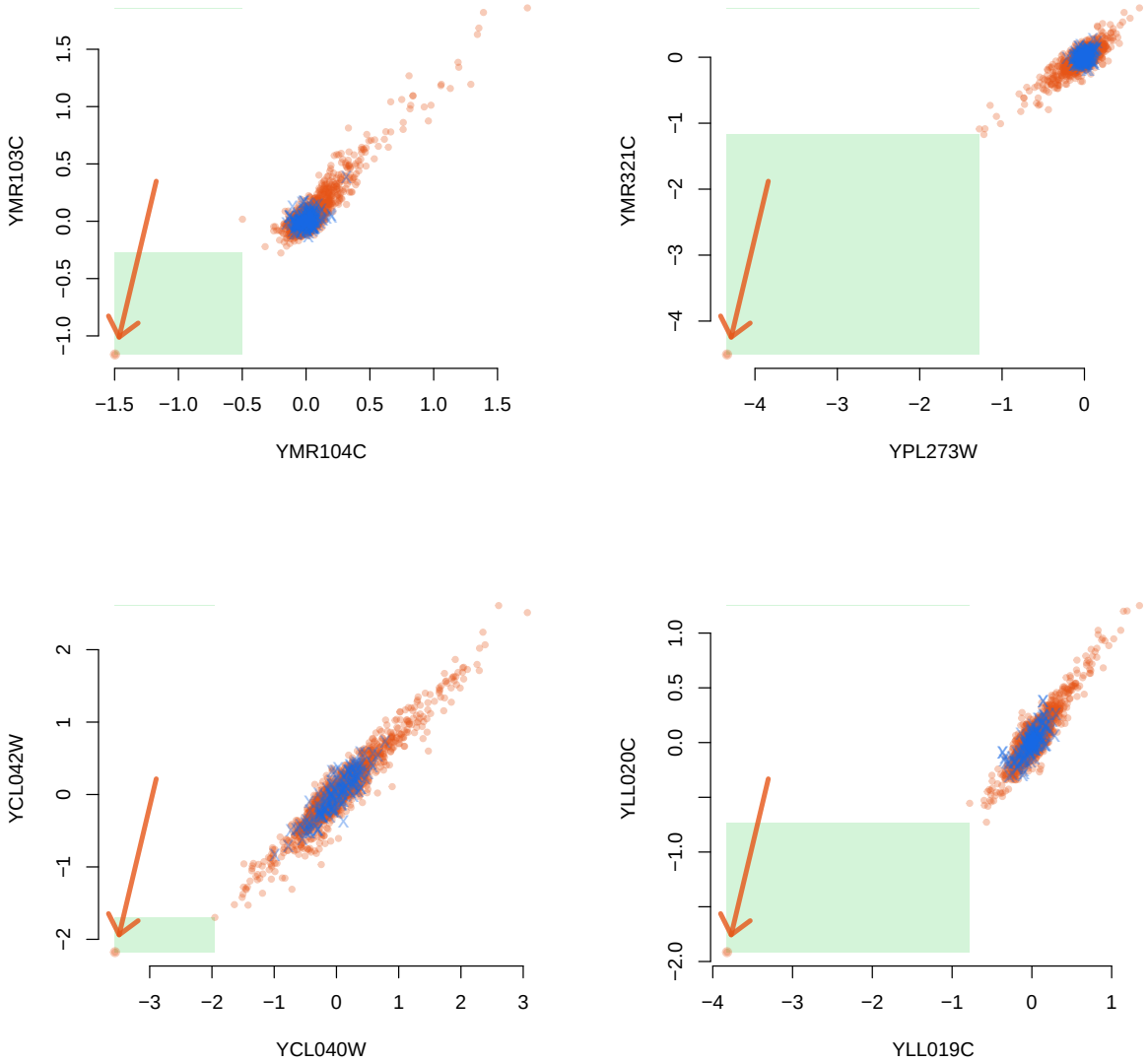
Predictors that are inferred to be causal by ICP...



...and the corresponding interventions on the predictor

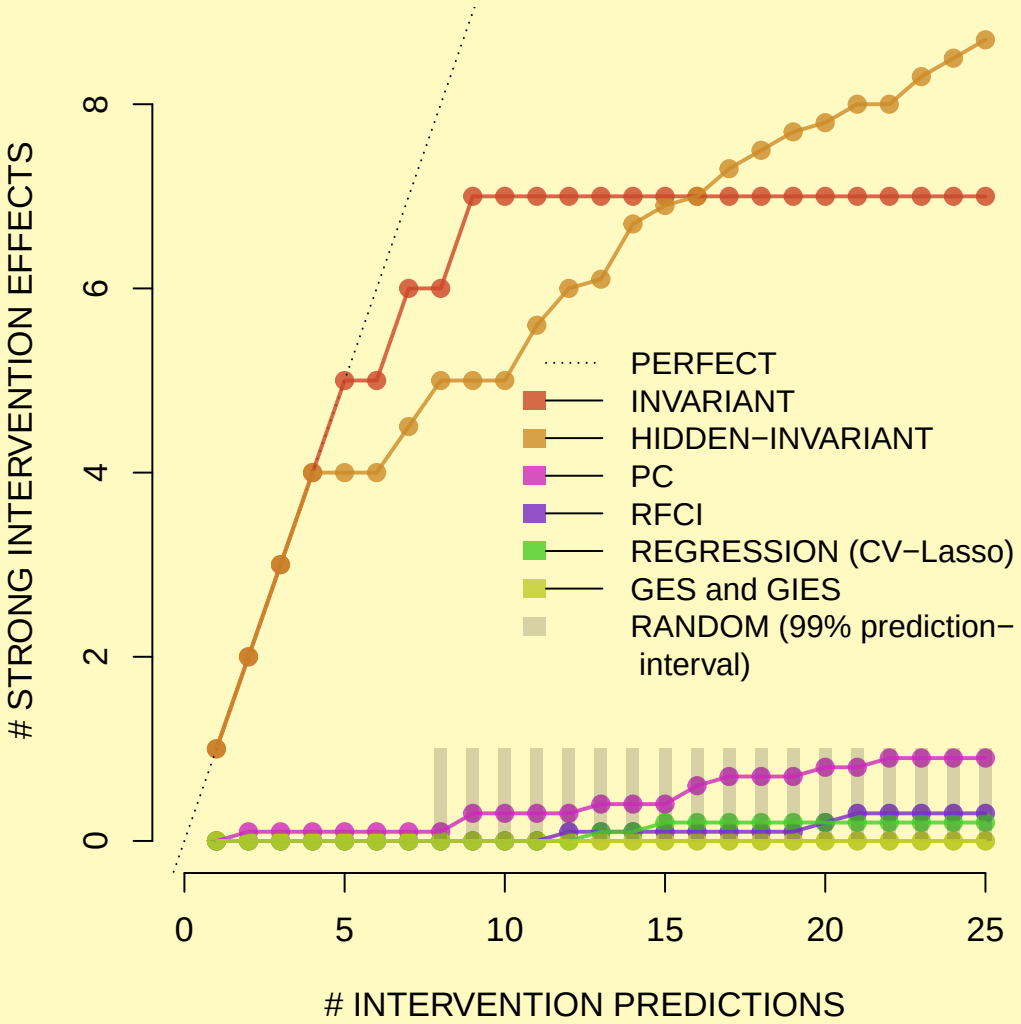


...and the corresponding interventions on the predictor



Kemmeren et al. 2014

Yeast data (Kemmeren et al., 2014)



Meinshausen et al., PNAS 2016

So far: invariance with respect to

anchor = environments

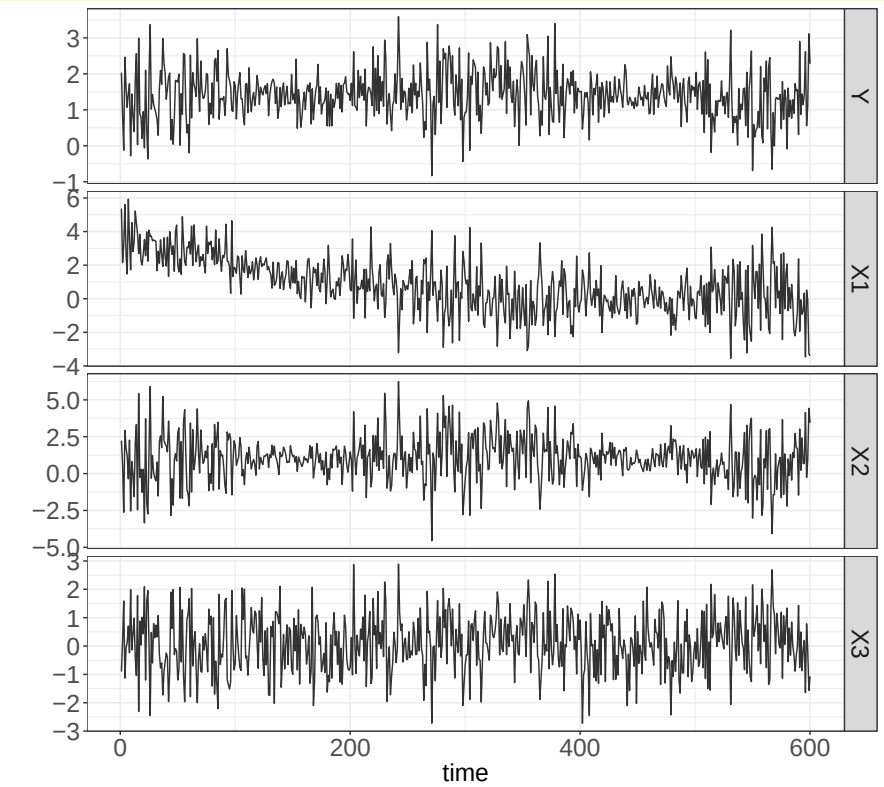
Also possible:

anchor = time

Suppose there is time.

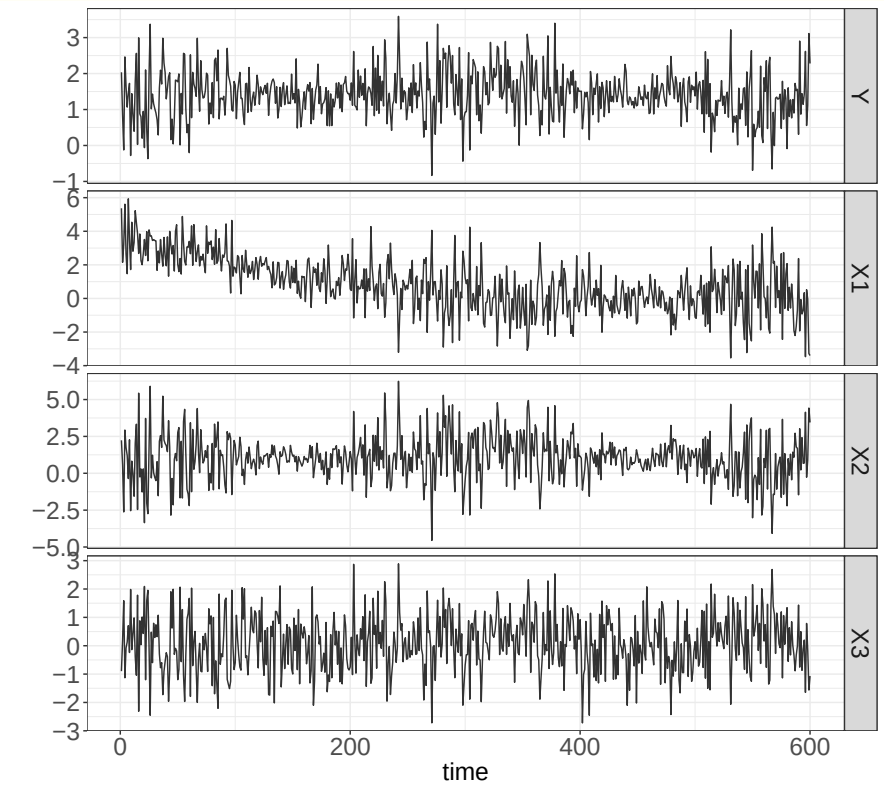
t	X_1	X_2	X_3	X_4	Y
1	3.4	−0.3	5.8	−2.1	2.2
2	1.7	−0.2	7.0	−1.2	0.4
3	−2.4	−0.1	4.3	−0.7	3.5
4	2.3	−0.3	5.5	−1.1	−4.4
5	3.5	−0.2	3.9	−0.9	−3.9
⋮	⋮	⋮	⋮	⋮	⋮

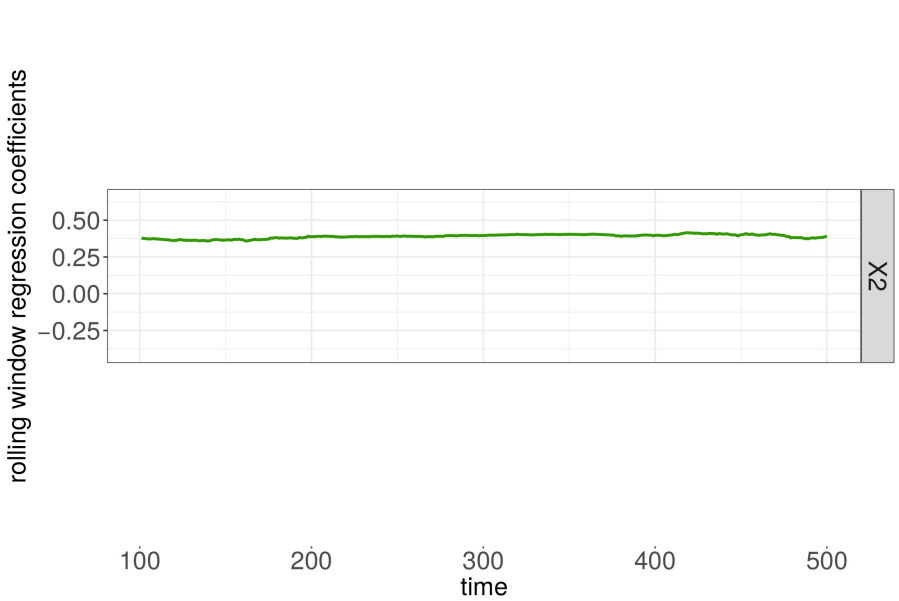
Regressing on (X_1, X_2, X_3):

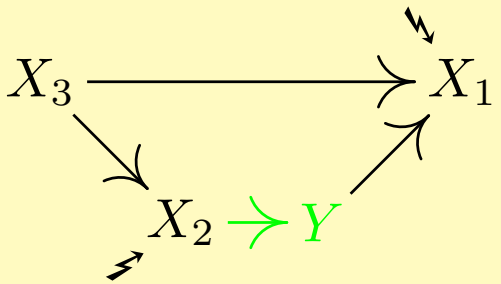


The coefficients change.

Regressing on X_1, X_2, and X_3:

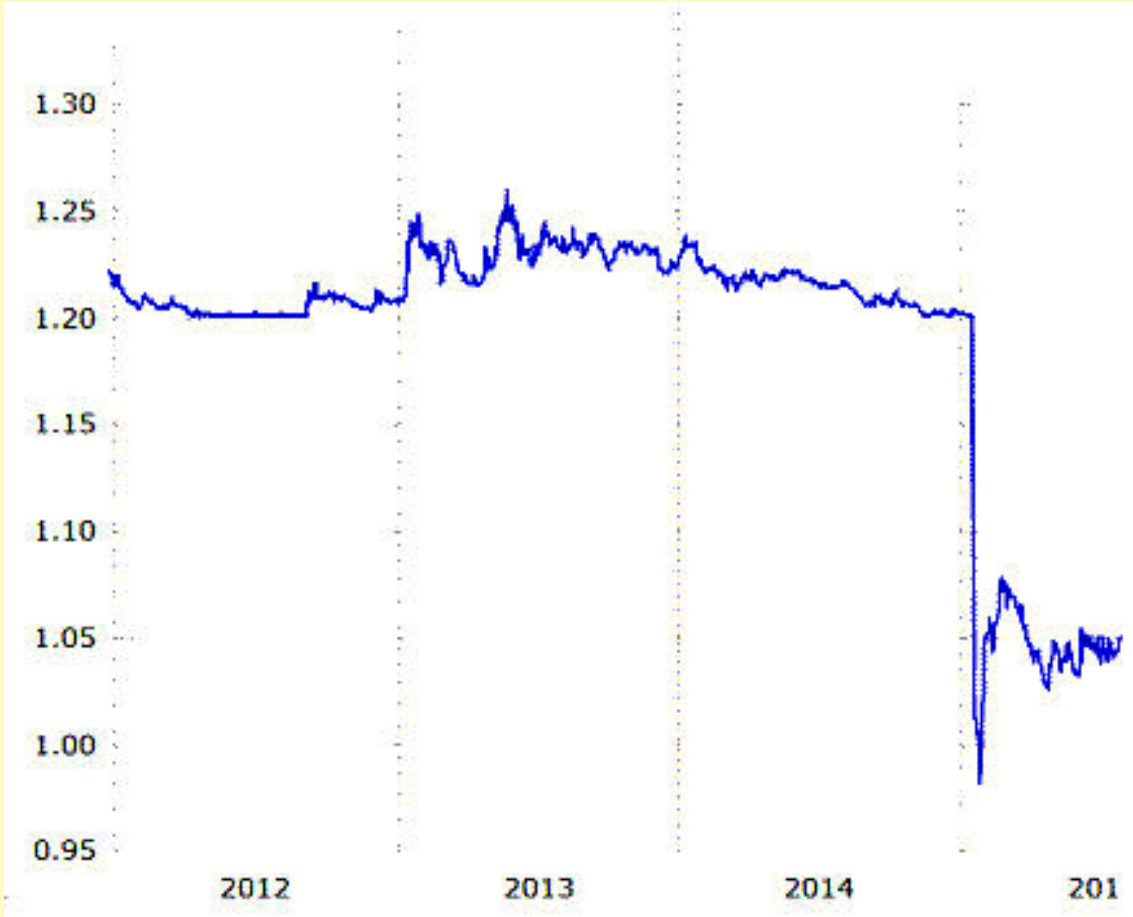






X_2 yields an invariant model. Ground truth:

How much CHF do I need to pay for buying 1 EUR?



http://www.fremdwaehrungskonto.info/wp-content/uploads/2015/07/CHF-EUR-Kursentwicklung-2011-2015.gif

monthly data Swiss National Bank Jan 1999 - Jan 2017

	description
Y	exchange rate Euro to Swiss Franks
X^1	change in average call money rate
X^2	log returns of foreign currency investments of the SNB
X^3	log returns of reserve positions at Intern. Monetary Fund of the SNB
X^4	log returns of monetary assistance loans of the SNB
X^5	log returns of Swiss Frank securities of the SNB
X^6	log returns of remaining assets of the SNB
X^7	log returns of Swiss GDP
X^8	log returns of Euro zone GDP
X^9	inflation rate for Switzerland

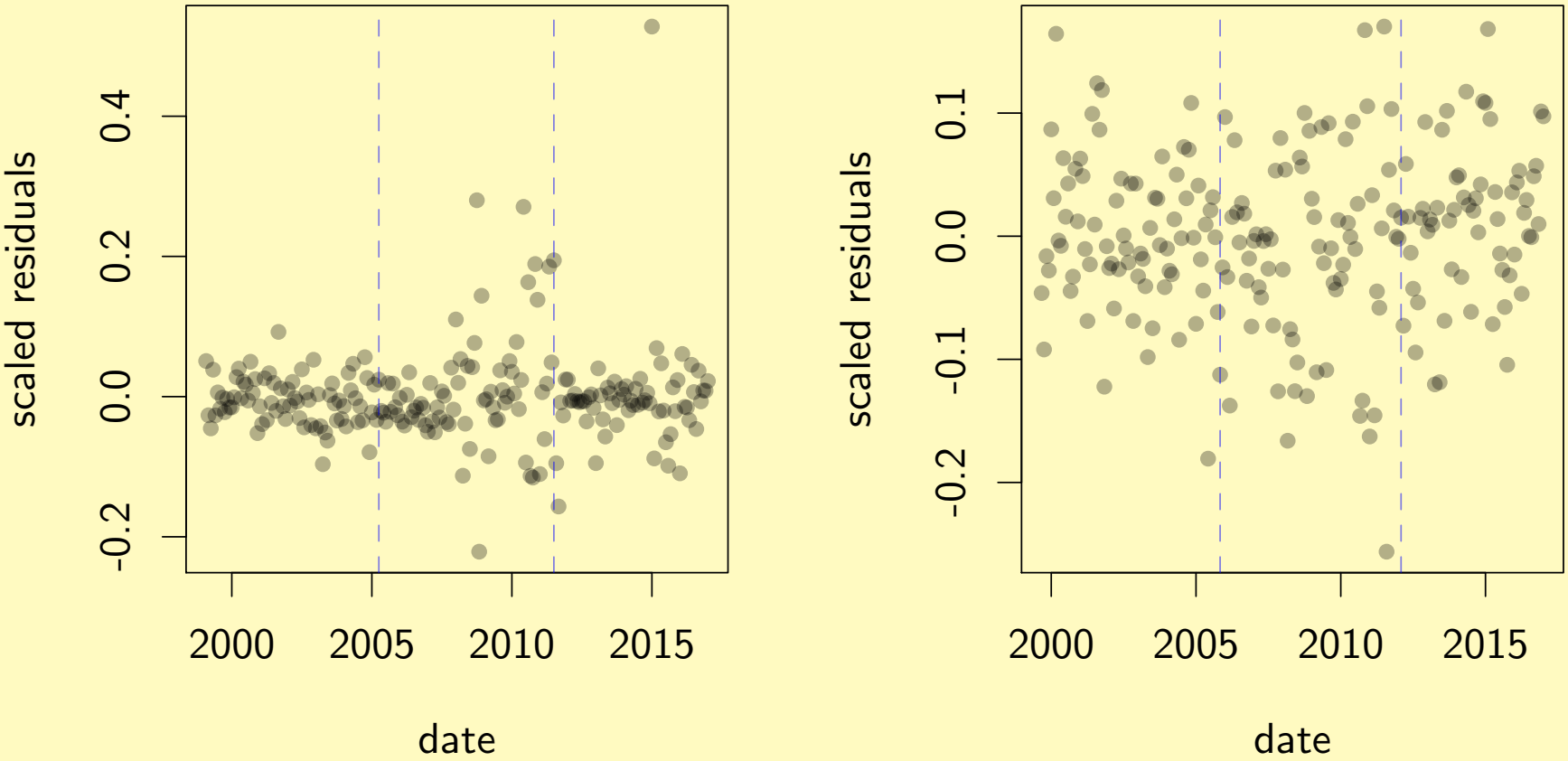


Figure: left plot (not invariant) and right plot (invariant)

monthly data Swiss National Bank Jan 1999 - Jan 2017

	description
Y	exchange rate Euro to Swiss Franks
X^1	change in average call money rate
X^2	log returns of foreign currency investments of the SNB
X^3	log returns of reserve positions at Intern. Monetary Fund of the SNB
X^4	log returns of monetary assistance loans of the SNB
X^5	log returns of Swiss Frank securities of the SNB
X^6	log returns of remaining assets of the SNB
X^7	log returns of Swiss GDP
X^8	log returns of Euro zone GDP
X^9	inflation rate for Switzerland

monthly data Swiss National Bank Jan 1999 - Jan 2017

	description
Y	exchange rate Euro to Swiss Franks
X^1	change in average call money rate
X^2	log returns of foreign currency investments of the SNB
X^3	log returns of reserve positions at Intern. Monetary Fund of the SNB
X^4	log returns of monetary assistance loans of the SNB
X^5	log returns of Swiss Frank securities of the SNB
X^6	log returns of remaining assets of the SNB
X^7	log returns of Swiss GDP
X^8	log returns of Euro zone GDP
X^9	inflation rate for Switzerland

Pfister, Bühlmann, JP, JASA 2018:
Non-inv. models rejected if $\sqrt{\log n/n} = o(a_n)$, where a_n is largest difference in noise variances.

Discrete environments (gene data):

JP, Meinshausen, Bühlmann: *Causal inference using invariant prediction: identification and confidence intervals*, JRSSB 2016

Discrete environments (gene data):

JP, Meinshausen, Bühlmann: *Causal inference using invariant prediction: identification and confidence intervals*, JRSSB 2016

No environments (finance data):  = time

Pfister, Bühlmann, JP: *Invariant causal pred. for seq. data*, JASA 2018

Discrete environments (gene data):

JP, Meinshausen, Bühlmann: *Causal inference using invariant prediction: identification and confidence intervals*, JRSSB 2016

No environments (finance data):  = time

Pfister, Bühlmann, JP: *Invariant causal pred. for seq. data*, JASA 2018

Nonlinear relations (fertility data):

Heinze-Deml, JP, Meinshausen: *Invariant Causal Prediction for Nonlinear Models*, Journal of Causal Inference 2018

Discrete environments (gene data):

JP, Meinshausen, Bühlmann: *Causal inference using invariant prediction: identification and confidence intervals*, JRSSB 2016

No environments (finance data):  = time

Pfister, Bühlmann, JP: *Invariant causal pred. for seq. data*, JASA 2018

Nonlinear relations (fertility data):

Heinze-Deml, JP, Meinshausen: *Invariant Causal Prediction for Nonlinear Models*, Journal of Causal Inference 2018

Discrete hidden variables (Earth system data):

Christiansen, JP: *Invariance-based Causal Discovery in the Presence of Discrete Hidden Variables*, JMLR 2020

Di

JP,

No

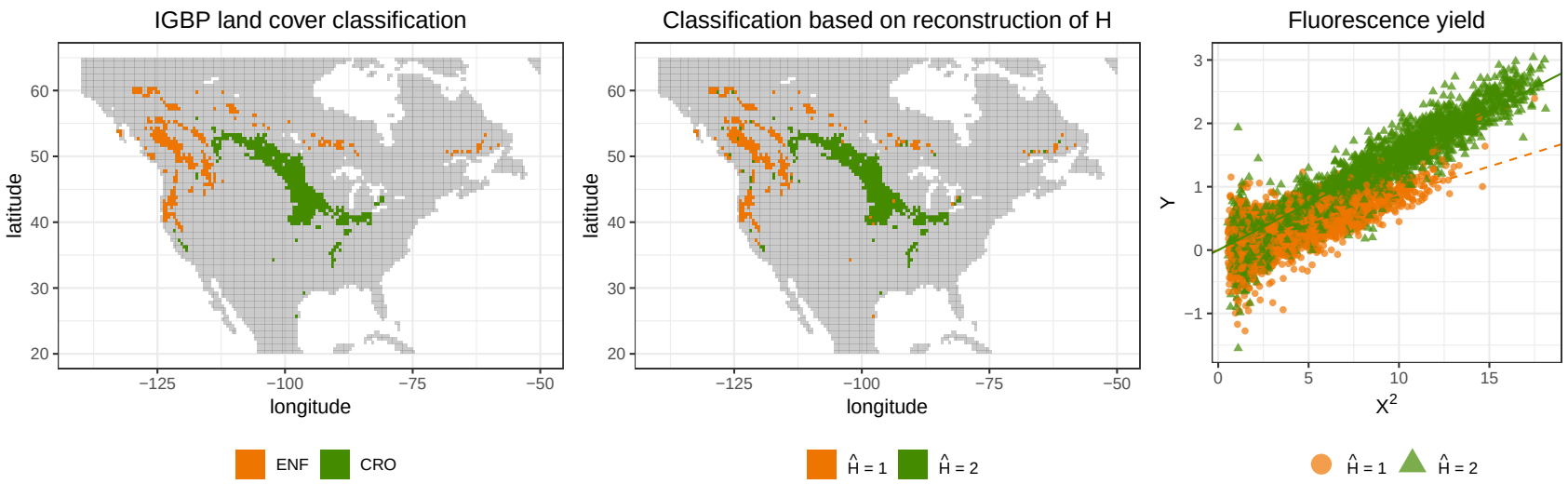
Pfis

No

Hei

Di

Chr



Discrete environments (gene data):

JP, Meinshausen, Bühlmann: *Causal inference using invariant prediction: identification and confidence intervals*, JRSSB 2016

No environments (finance data):  = time

Pfister, Bühlmann, JP: *Invariant causal pred. for seq. data*, JASA 2018

Nonlinear relations (fertility data):

Heinze-Deml, JP, Meinshausen: *Invariant Causal Prediction for Nonlinear Models*, Journal of Causal Inference 2018

Discrete hidden variables (Earth system data):

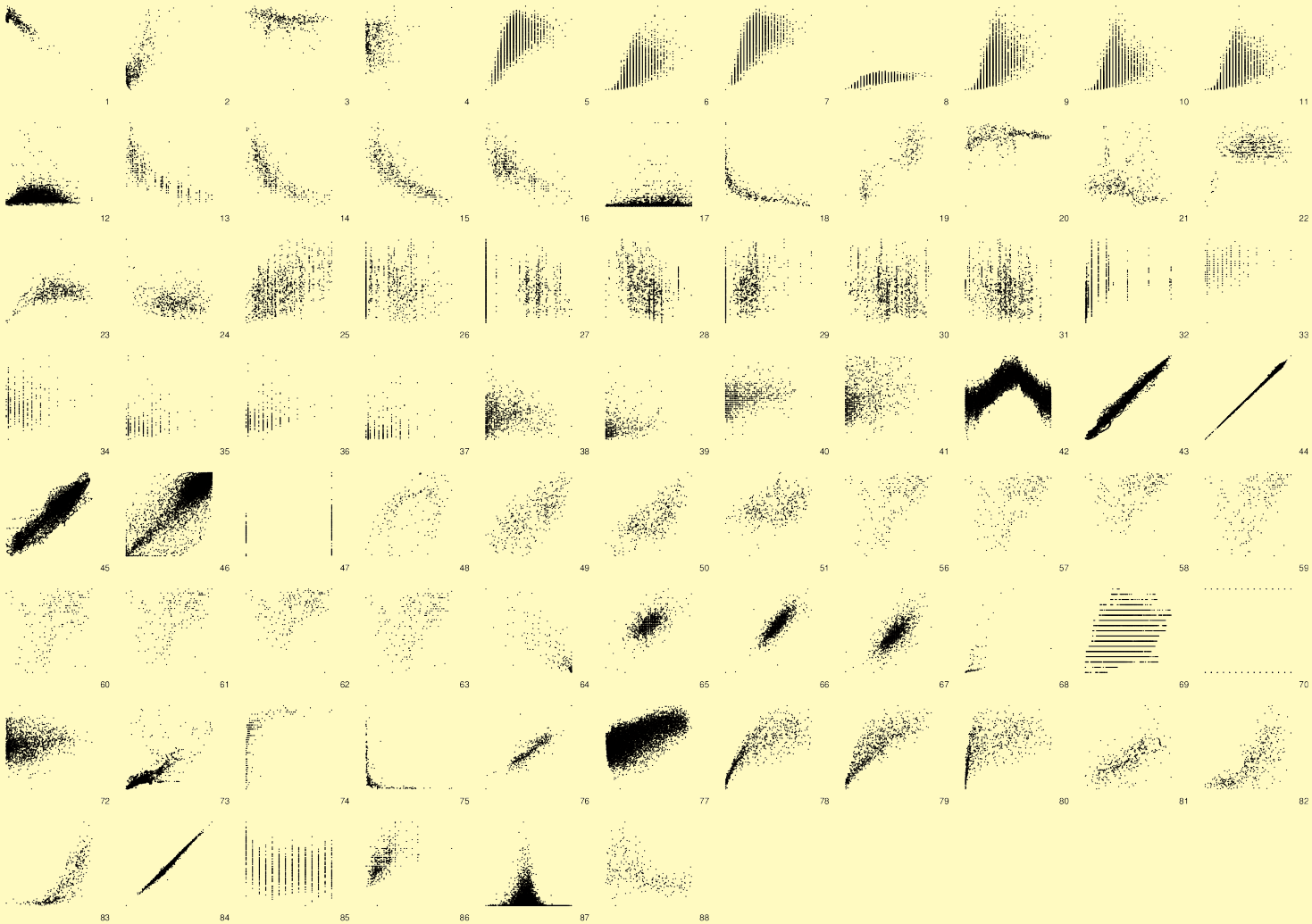
Christiansen, JP: *Invariance-based Causal Discovery in the Presence of Discrete Hidden Variables*, JMLR 2020

Survival data (registry data):

Laksafoss, JP: *Causal Methods for Survival Analysis*, work in progress

What can we do with two variables and no environments?
(In general, nothing is possible.)

Idea 3: restricted structural causal models



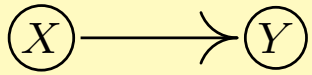
Mooij, JP, Janzing, Zscheischler, Schölkopf: *Disting. cause from effect using obs. data: methods and benchm.*, JMLR 2016

Idea 3: restricted structural causal models

Consider a distribution entailed by

$$Y = f(X) + N_Y$$

with $N_Y, X \overset{ind}{\sim} \mathcal{N}$

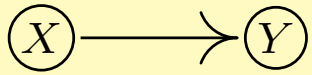


Idea 3: restricted structural causal models

Consider a distribution entailed by

$$Y = f(X) + N_Y$$

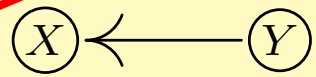
with $N_Y, X \overset{ind}{\sim} \mathcal{N}$



Then, if f is nonlinear, there is no

$$X = g(Y) + M_X$$

with $M_X, Y \overset{ind}{\sim} \mathcal{N}$



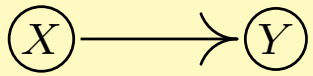
JP, J. Mooij, D. Janzing and B. Schölkopf: *Causal Discovery with Continuous Additive Noise Models, JMLR 2014*

Idea 3: restricted structural causal models

Consider a distribution entailed by

$$Y = X^3 + N_Y$$

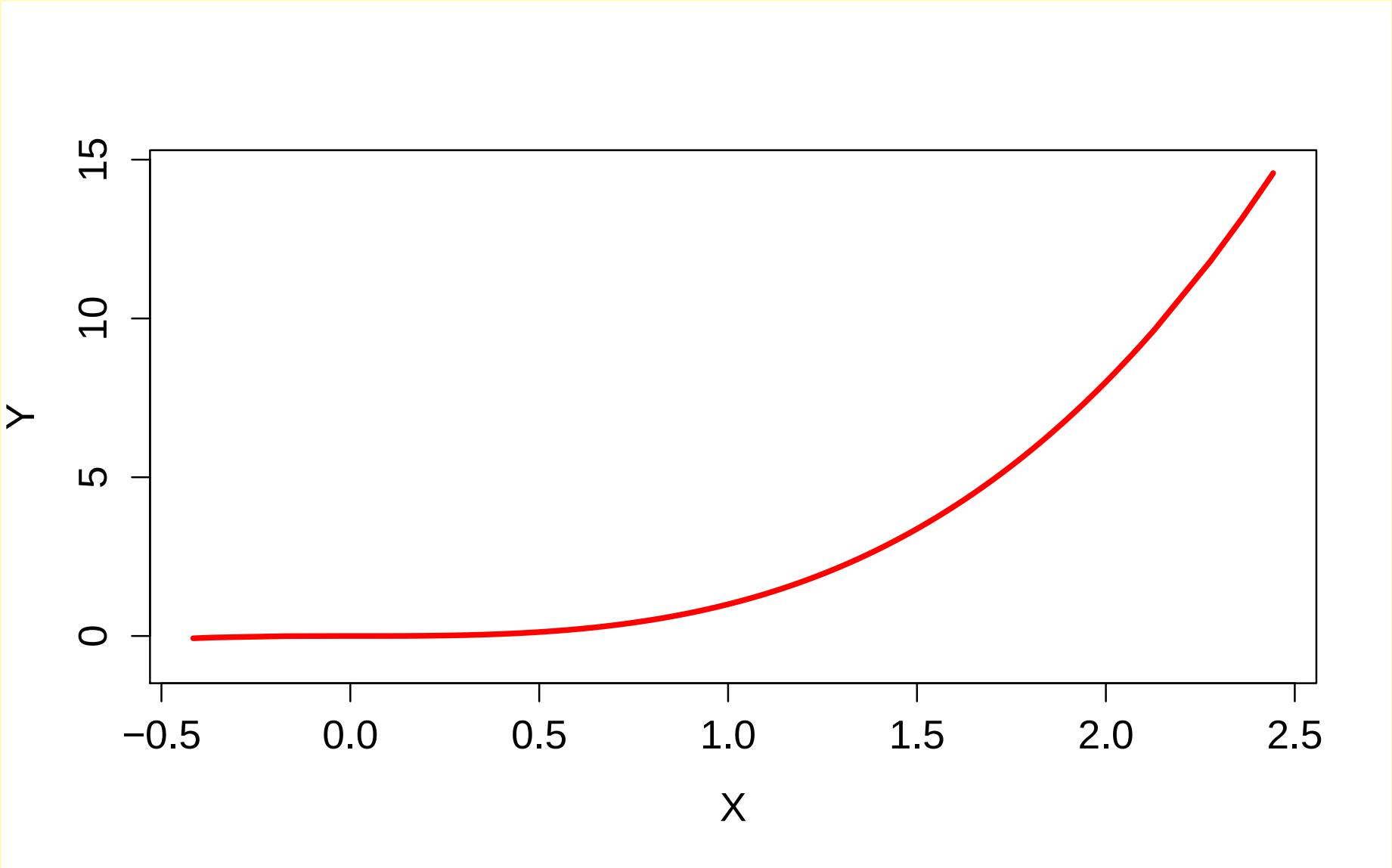
with $N_Y, X \overset{ind}{\sim} \mathcal{N}$



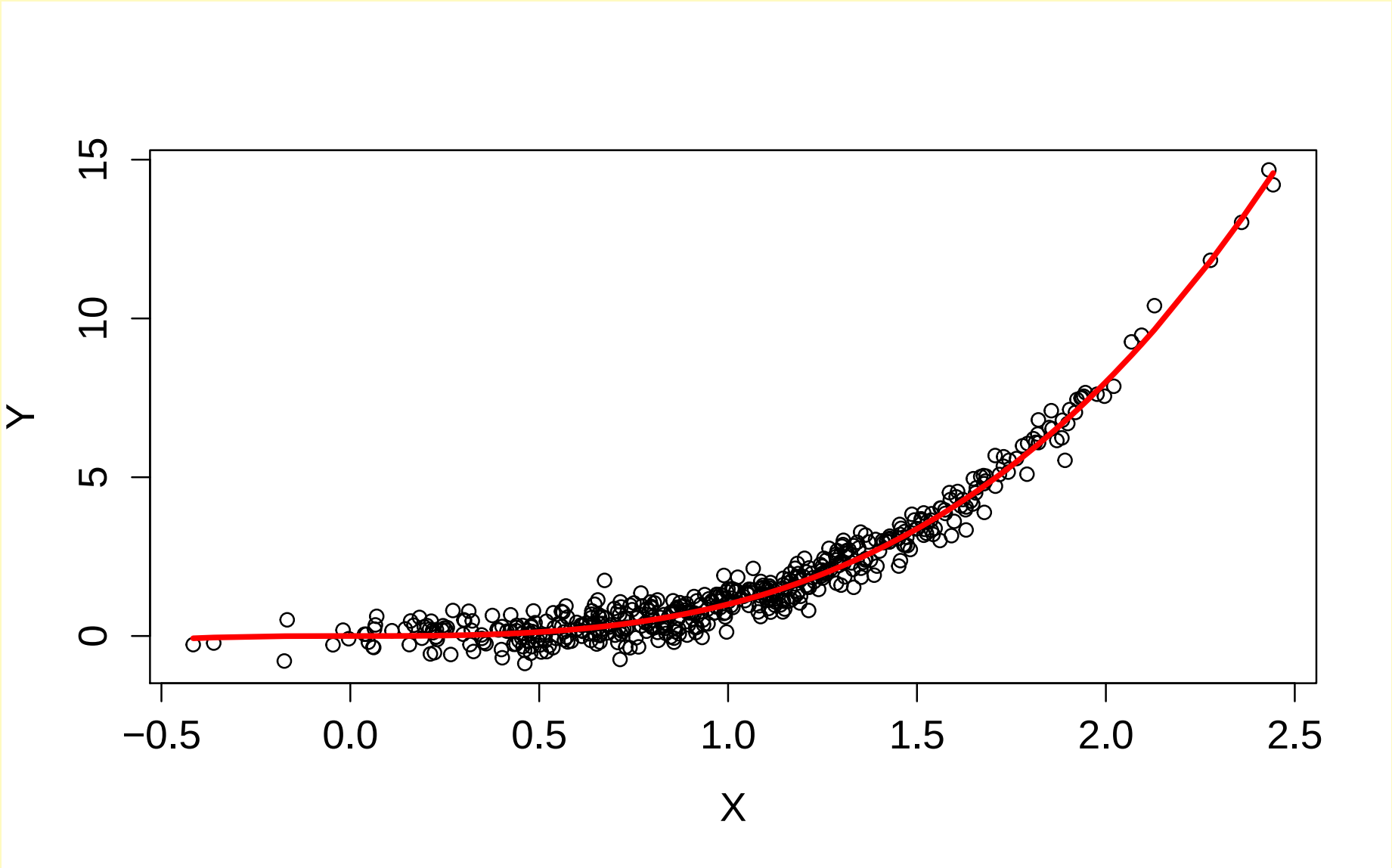
with

$$X \sim \mathcal{N}(1, 0.5^2)$$
$$N_Y \sim \mathcal{N}(0, 0.4^2)$$

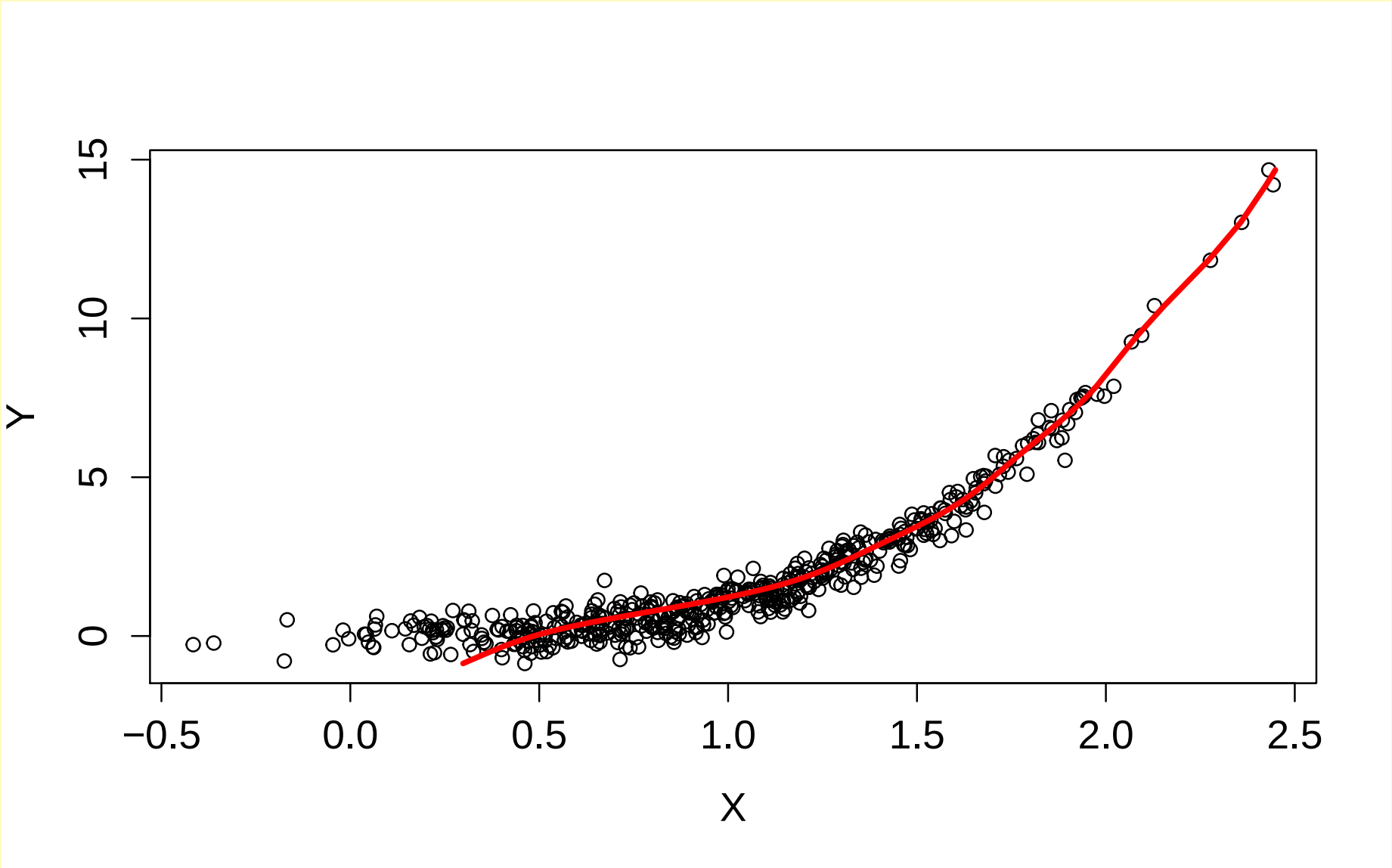
Idea 3: restricted structural causal models



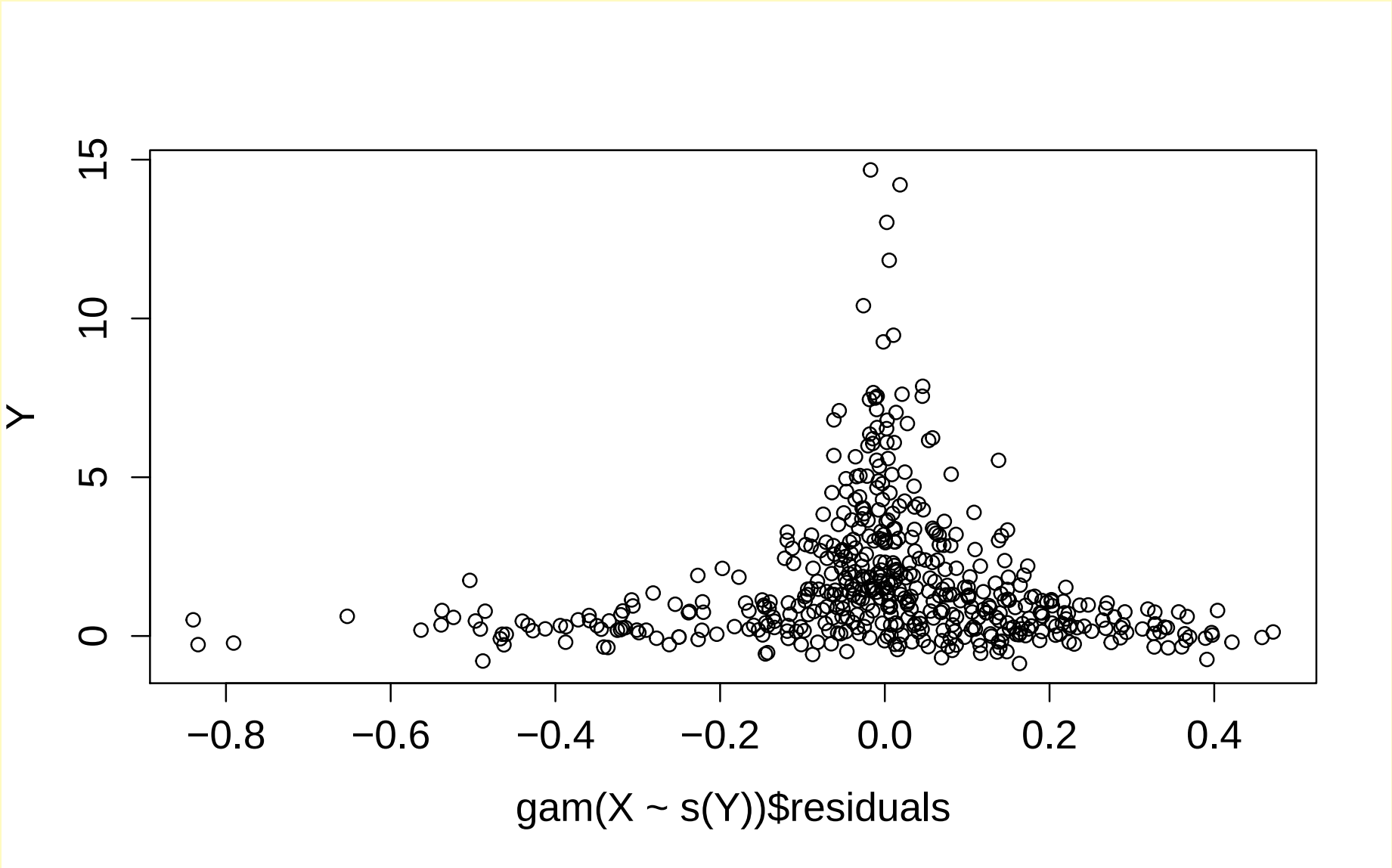
Idea 3: restricted structural causal models



Idea 3: restricted structural causal models



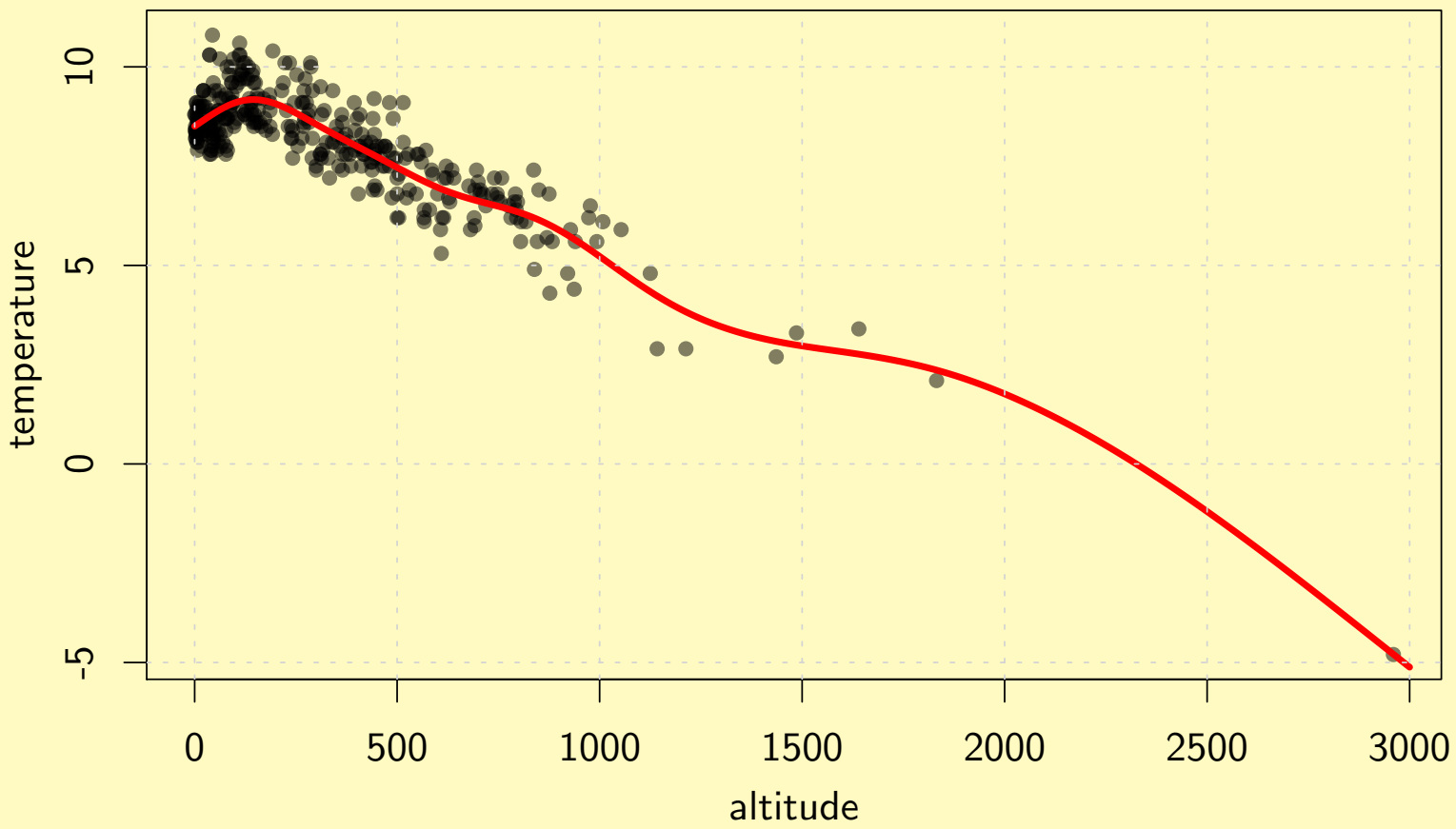
Idea 3: restricted structural causal models



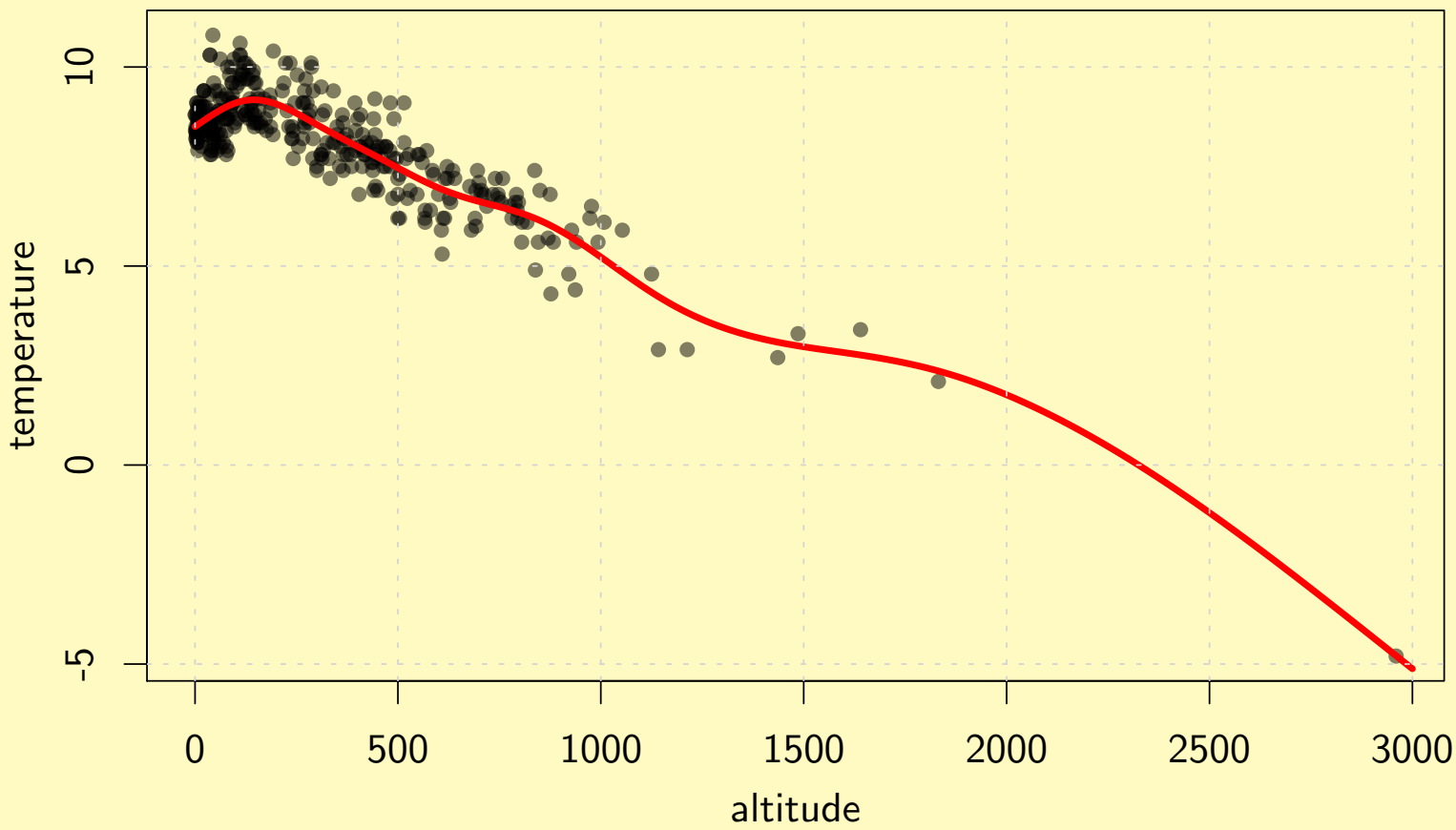
Idea 3: restricted structural causal models

Method... Exercise-ANM-biv.R

Example: altitude and temperature



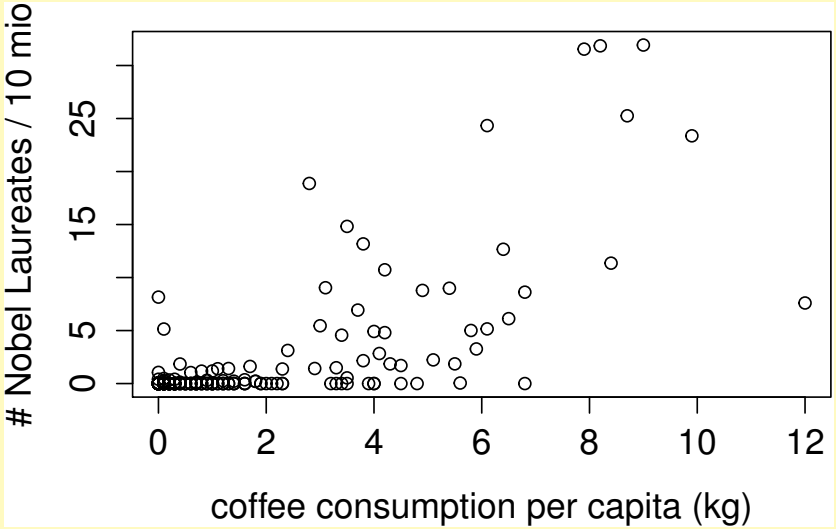
Example: altitude and temperature



p-value forward: 0.024

p-value backward: 0.0000000000019

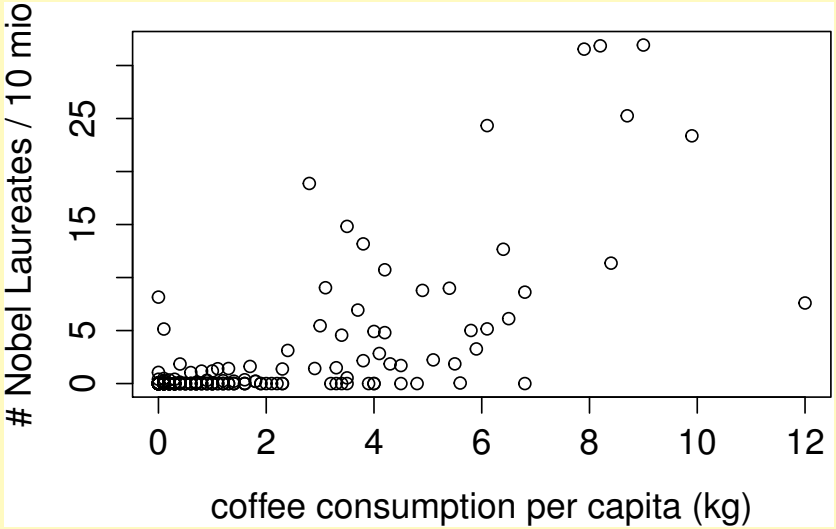
Example: coffee



Correlation: 0.698

p-value: $< 2.2 \cdot 10^{-16}$

Example: coffee

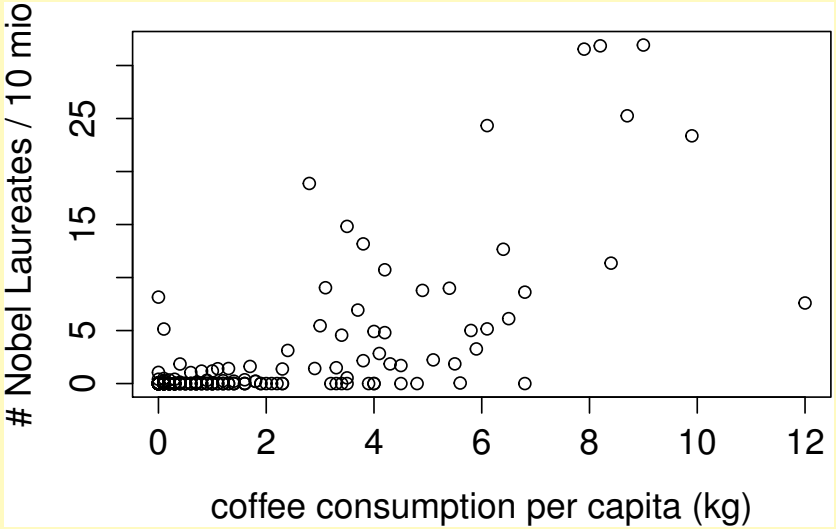


Correlation: 0.698
p-value: $< 2.2 \cdot 10^{-16}$

Coffee $\rightarrow$ Nobel Prize: Dependent residuals (p-value of $5.1 \cdot 10^{-78}$).
Nobel Prize $\rightarrow$ Coffee: Dependent residuals (p-value of $3.1 \cdot 10^{-12}$).

$\Rightarrow$ Model class too small? Causally insufficient?

Example: coffee

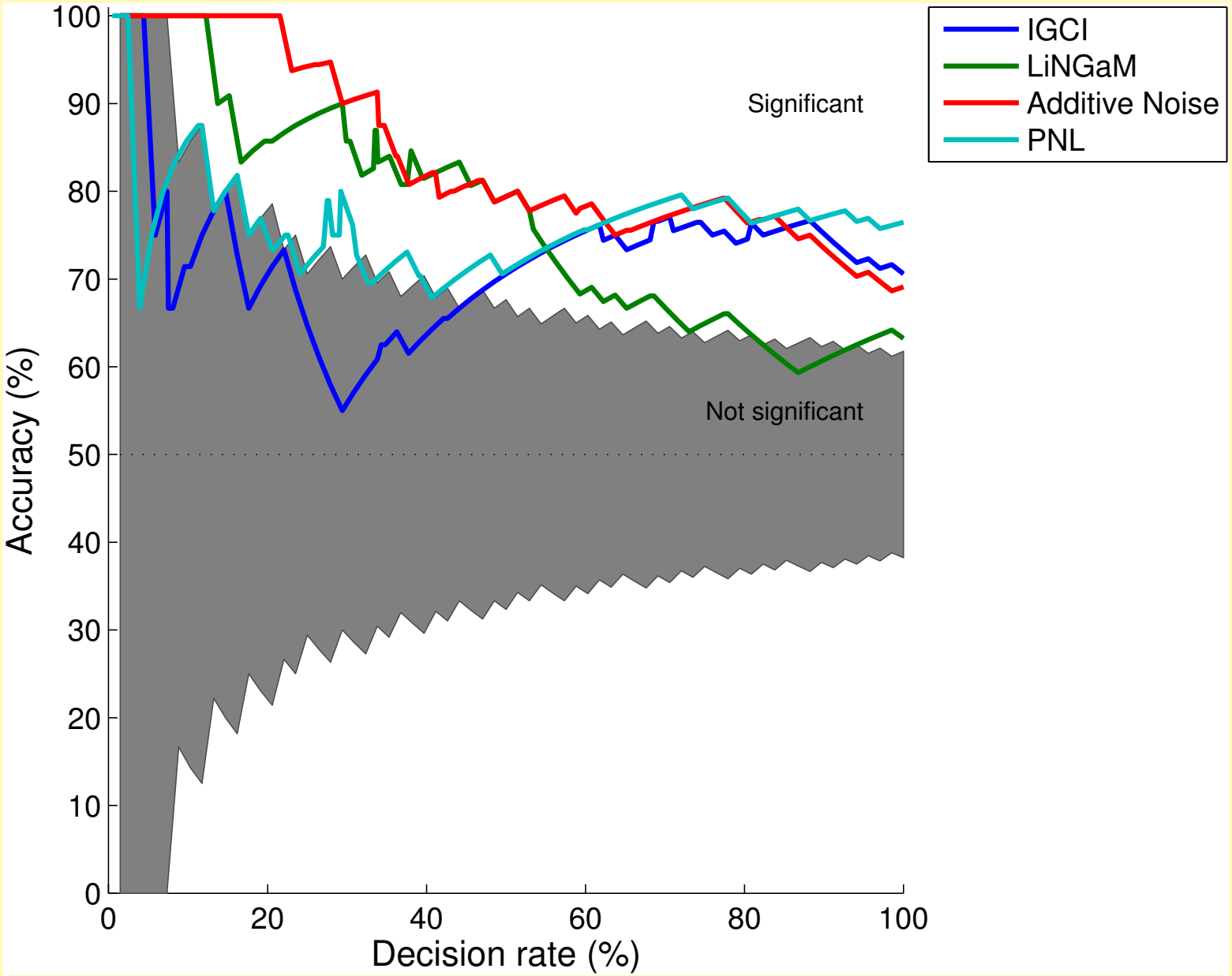


Correlation: 0.698
p-value: $< 2.2 \cdot 10^{-16}$

Coffee $\rightarrow$ Nobel Prize: Dependent residuals (p-value of $5.1 \cdot 10^{-78}$).
Nobel Prize $\rightarrow$ Coffee: Dependent residuals (p-value of $3.1 \cdot 10^{-12}$).

$\Rightarrow$ Model class too small? Causally insufficient?
Question: When is a p-value too small?

Real Data: cause-effect pairs



Idea 3: restricted structural causal models

Slightly surprising:

$$\text{identifiability for two variables} \rightsquigarrow \text{identifiability for } d \text{ variables}$$

Peters et al.: *Identifiability of Causal Graphs using Functional Models*, UAI 2011

Idea 3: restricted structural causal models

Slightly surprising:

$$\text{identifiability for two variables} \rightsquigarrow \text{identifiability for } d \text{ variables}$$

Peters et al.: *Identifiability of Causal Graphs using Functional Models*, UAI 2011

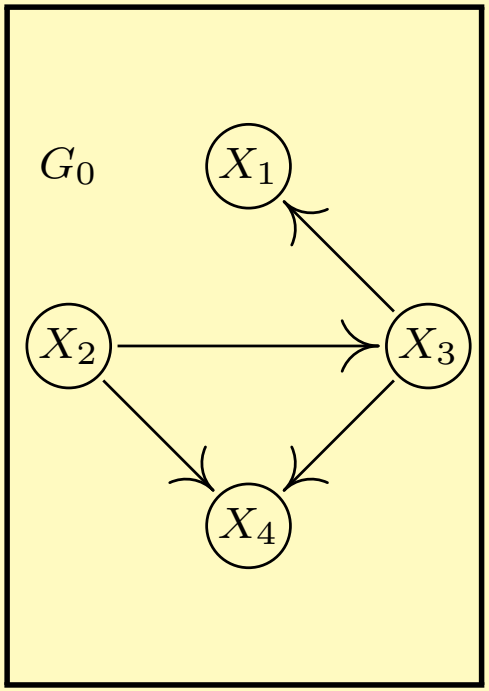
Most important counterexample: linear Gaussian.

Idea 3: restricted structural causal models

Assume $P(X_1, \dots, X_4)$ has been entailed by

$$X_1 = f_1(X_3, N_1)$$
$$X_2 = N_2$$
$$X_3 = f_3(X_2, N_3)$$
$$X_4 = f_4(X_2, X_3, N_4)$$

- N_i jointly independent
- G_0 has no cycles



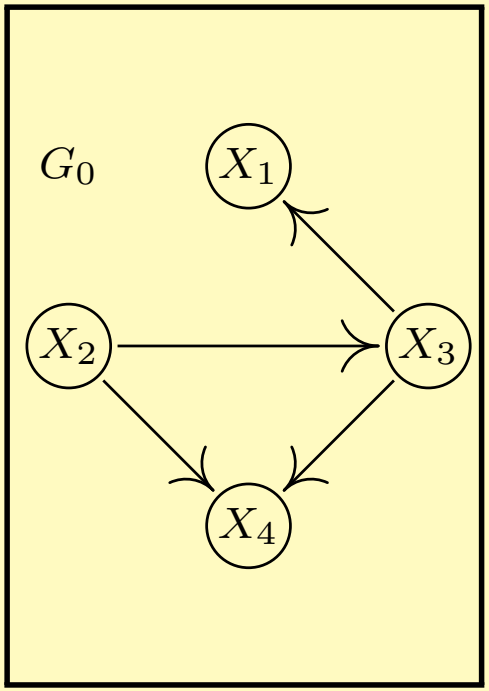
Structural equation model.
Can the DAG be recovered from $P(X_1, \dots, X_4)$?

Idea 3: restricted structural causal models

Assume $P(X_1, \dots, X_4)$ has been entailed by

$$X_1 = f_1(X_3, N_1)$$
$$X_2 = N_2$$
$$X_3 = f_3(X_2, N_3)$$
$$X_4 = f_4(X_2, X_3, N_4)$$

- N_i jointly independent
- G_0 has no cycles



Structural equation model.
Can the DAG be recovered from $P(X_1, \dots, X_4)$? **No.**

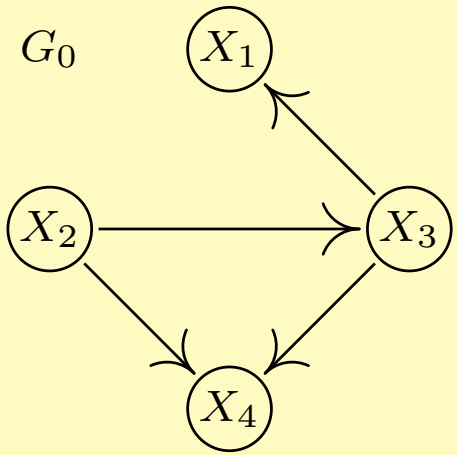
JP, J. Mooij, D. Janzing and B. Schölkopf: *Causal Discovery with Continuous Additive Noise Models, JMLR 2014*
P. Bühlmann, JP, J. Ernest: *CAM: Causal add. models, high-dim. order search and penalized regr., Annals of Statistics 2014*

Idea 3: restricted structural causal models

Assume $P(X_1, \ldots, X_4)$ has been entailed by

$$X_1 = f_1(X_3) + N_1$$
$$X_2 = N_2$$
$$X_3 = f_3(X_2) + N_3$$
$$X_4 = f_4(X_2, X_3) + N_4$$

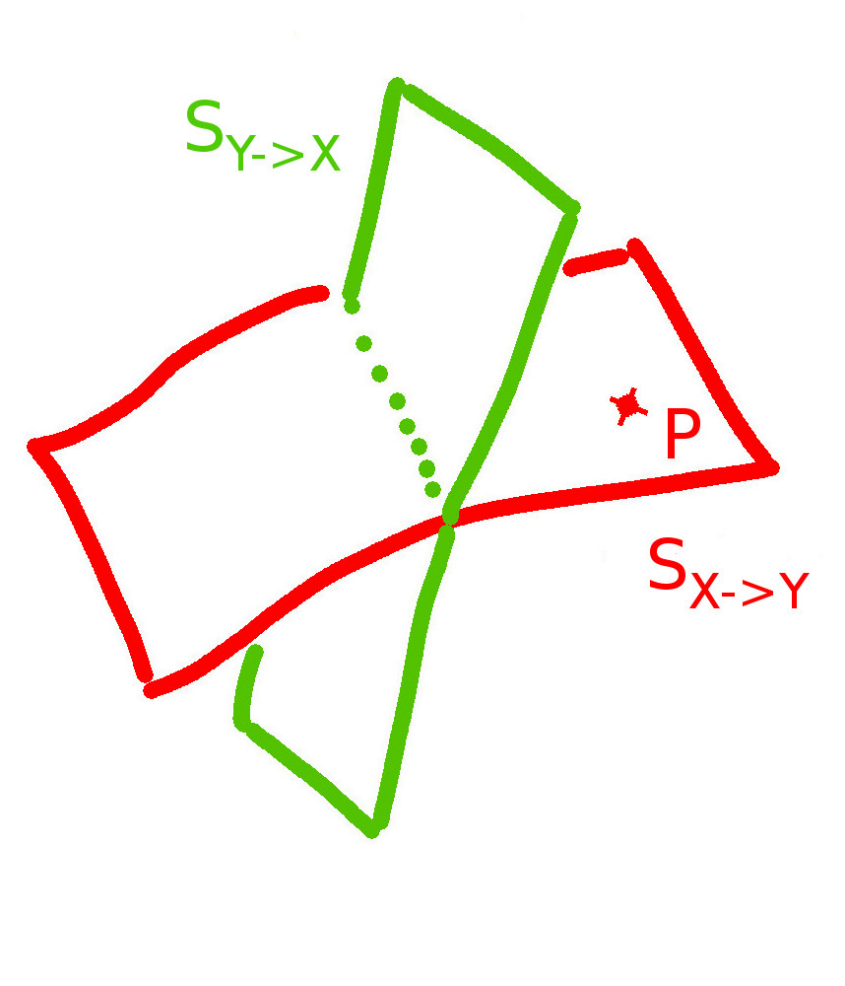
- $N_i \sim \mathcal{N}(0, \sigma_i^2)$ jointly independent
- G_0 has no cycles



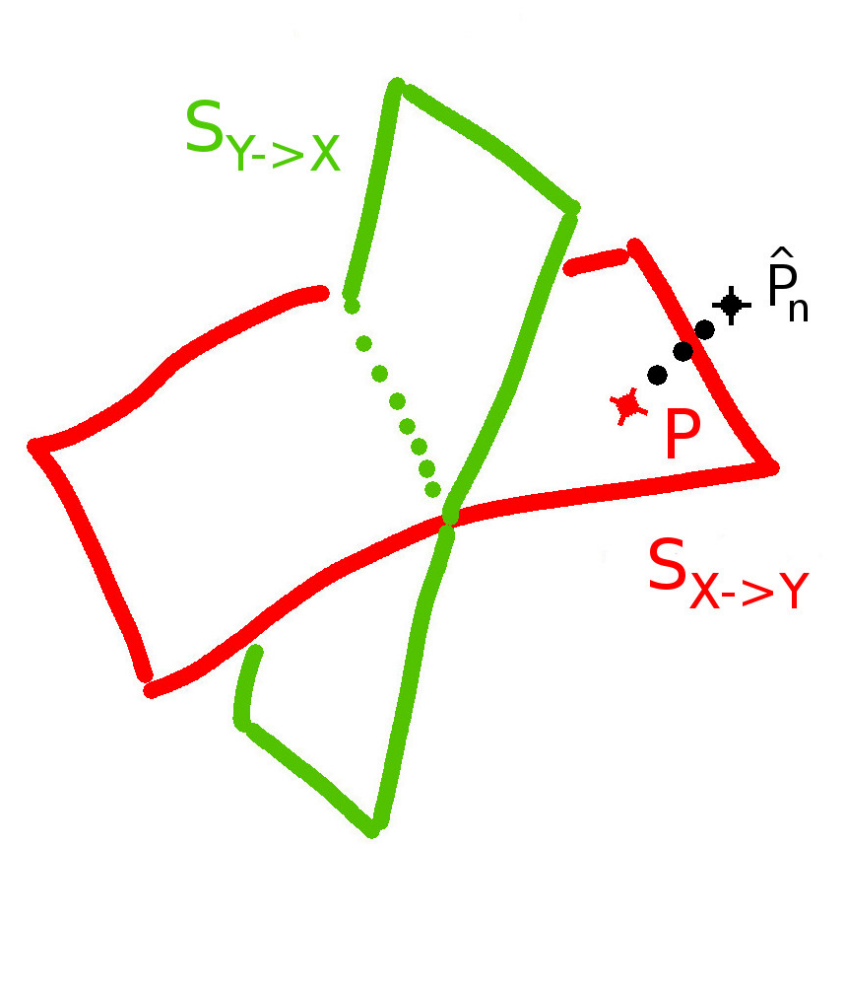
Additive noise model with Gaussian noise.
Can the DAG be recovered from $P(X_1, \ldots, X_4)$? **Yes iff f_i nonlinear.**

JP, J. Mooij, D. Janzing and B. Schölkopf: *Causal Discovery with Continuous Additive Noise Models, JMLR 2014*
P. Bühlmann, JP, J. Ernest: *CAM: Causal add. models, high-dim. order search and penalized regr.*, Annals of Statistics 2014

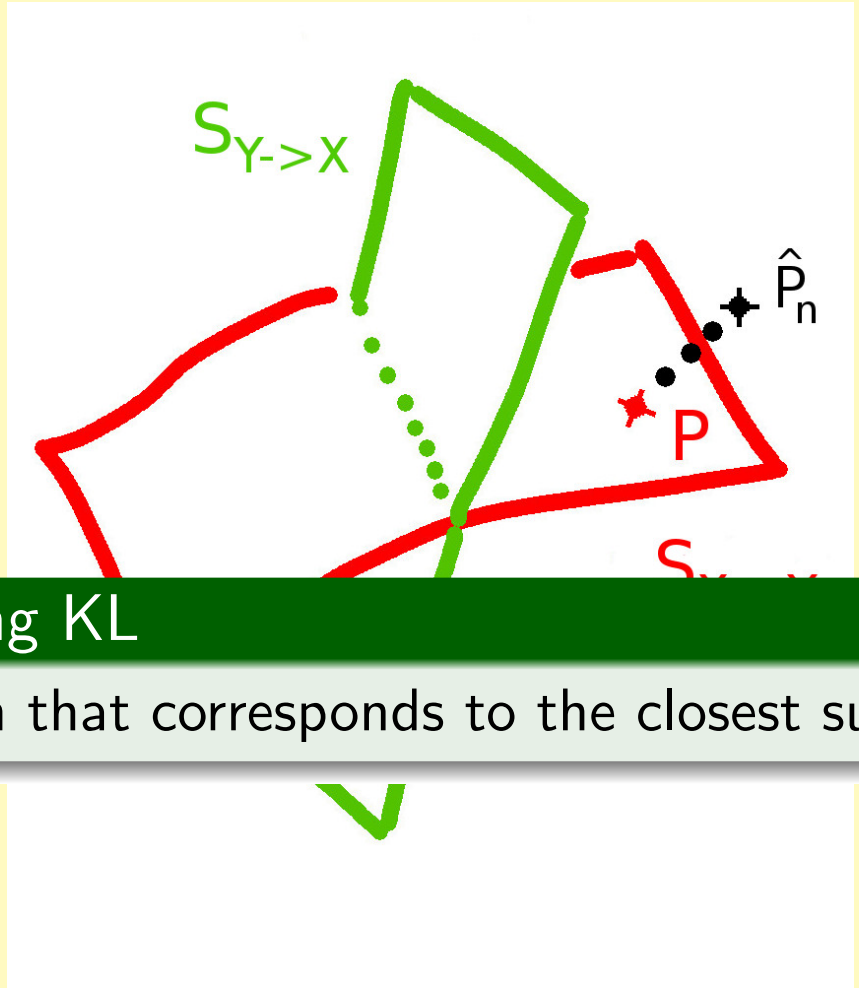
Idea 3: restricted structural causal models



Idea 3: restricted structural causal models



Idea 3: restricted structural causal models



Method: Minimizing KL

Choose the direction that corresponds to the closest subspace...

Idea 3: restricted structural causal models

Consider model classes

$$\mathcal{S}_G := \{Q : Q \text{ entailed by a causal additive model (CAM) with DAG } G\}$$

Define

$$\hat{G}_n := \underset{\text{DAG } G}{\operatorname{argmin}} \inf_{Q \in \mathcal{S}_G} \mathrm{KL}(\hat{P}_n \,||\, Q)$$

Idea 3: restricted structural causal models

Consider model classes

$$\mathcal{S}_G := \{Q : Q \text{ entailed by a causal additive model (CAM) with DAG } G\}$$

Define

$$\hat{G}_n := \underset{\text{DAG } G}{\operatorname{argmin}} \inf_{Q \in \mathcal{S}_G} \mathrm{KL}(\hat{P}_n \,||\, Q)$$

$$\overset{\text{max. likelihood}}{=} \underset{\text{DAG } G}{\operatorname{argmin}} \sum_{i=1}^{p} \log \hat{\mathrm{var}}(\mathrm{residuals}_{\mathbf{PA}_i^G \to X_i})$$

Idea 3: restricted structural causal models

Consider model classes

$$\mathcal{S}_G := \{Q : Q \text{ entailed by a causal additive model (CAM) with DAG } G\}$$

Define

$$\hat{G}_n := \underset{\text{DAG } G}{\operatorname{argmin}} \inf_{Q \in \mathcal{S}_G} \mathrm{KL}(\hat{P}_n \,||\, Q)$$

$$\overset{\text{max. likelihood}}{=} \underset{\text{DAG } G}{\operatorname{argmin}} \sum_{i=1}^{p} \log \hat{\mathrm{var}}(\mathrm{residuals}_{\mathbf{PA}_i^G \to X_i})$$

Wait, there is no penalization on the number of edges!

Idea 3: restricted structural causal models

Consider model classes

$$\mathcal{S}_G := \{Q : Q \text{ entailed by a causal additive model (CAM) with DAG } G\}$$

Define

$$\hat{G}_n := \underset{\text{DAG } G}{\operatorname{argmin}} \inf_{Q \in \mathcal{S}_G} \mathrm{KL}(\hat{P}_n \,||\, Q)$$

$$\overset{\text{max. likelihood}}{=} \underset{\text{DAG } G}{\operatorname{argmin}} \sum_{i=1}^{p} \log \hat{\mathrm{var}}(\mathrm{residuals}_{\mathbf{PA}_i^G \to X_i})$$

Wait, there is no penalization on the number of edges!
Wait again, there are too many DAGs!

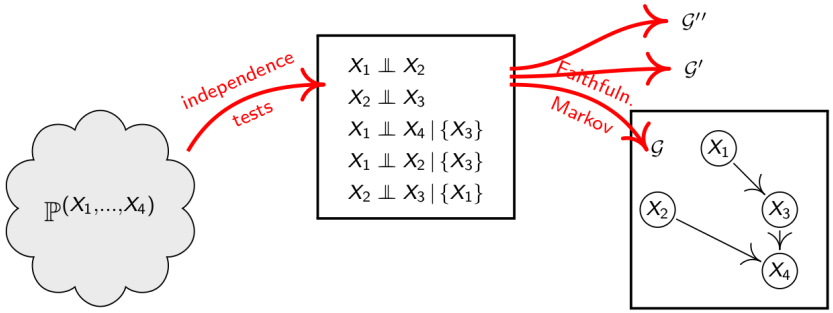
Idea 3: restricted structural causal models

d	number of DAGs with d nodes
1	1
2	3
3	25
4	543
5	29281
6	3781503
7	1138779265
8	783702329343
9	1213442454842881
10	4175098976430598143
11	31603459396418917607425
12	521939651343829405020504063
13	18676600744432035186664816926721
14	1439428141044398334941790719839535103
15	237725265553410354992180218286376719253505
16	83756670773733320287699303047996412235223138303
17	62707921196923889899446452602494921906963551482675201
18	9942119532215951589522891459235452451655502687858830501478 3
19	33277190122710759173617757331126112588358307625842190258354677 3505
20	2344880451051088988152559855229099188899081192234291298795803236068491263
21	3469876828358875002875932843018108822231394454043860171902755911344658607767 5521
22	1075822921725761493652956179327624326573727662809185218104090000500559527511693495107583

https://oeis.org/A003024/b003024.txt

Summary Part II:

- Idea 1: independence-based methods (single environment)



- Idea 2: invariant prediction (the more heterogeneity the better!)

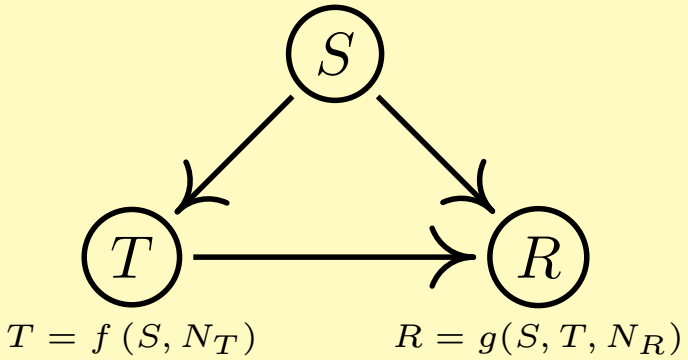
- Idea 3: additive noise (single environment)

$$X_1 = f_1(X_3) + N_1$$
$$X_2 = N_2$$
$$X_3 = f_3(X_2) + N_3$$
$$X_4 = f_4(X_2, X_3) + N_4$$

Part III: Applications to Machine Learning

Idea 1: Reinforcement Learning

Recall the kidney stones:



$$p(r,t,s) = p(r \mid t,s) \cdot \quad p(t \mid s) \quad \cdot p(s)$$

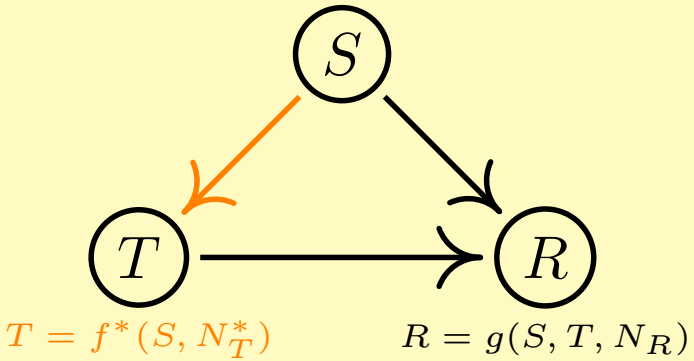
$$p_3^*(r,t,s) = p(r \mid t,s) \cdot \quad \underbrace{p^*(t \mid s)}_{p^*(t \mid s) = ?} \quad \cdot p(s)$$

Question: What would happen if...?

Question: What is $\sup_{p^*} \mathbf{E}_{p_3^*} R$?

Idea 1: Reinforcement Learning

Recall the kidney stones:



$$p(r,t,s) = p(r \mid t,s) \cdot \quad p(t \mid s) \quad \cdot p(s)$$

$$p_3^*(r,t,s) = p(r \mid t,s) \cdot \quad \underbrace{p^*(t \mid s)}_{p^*(t \mid s)=?} \quad \cdot p(s)$$

Question: What would happen if...?

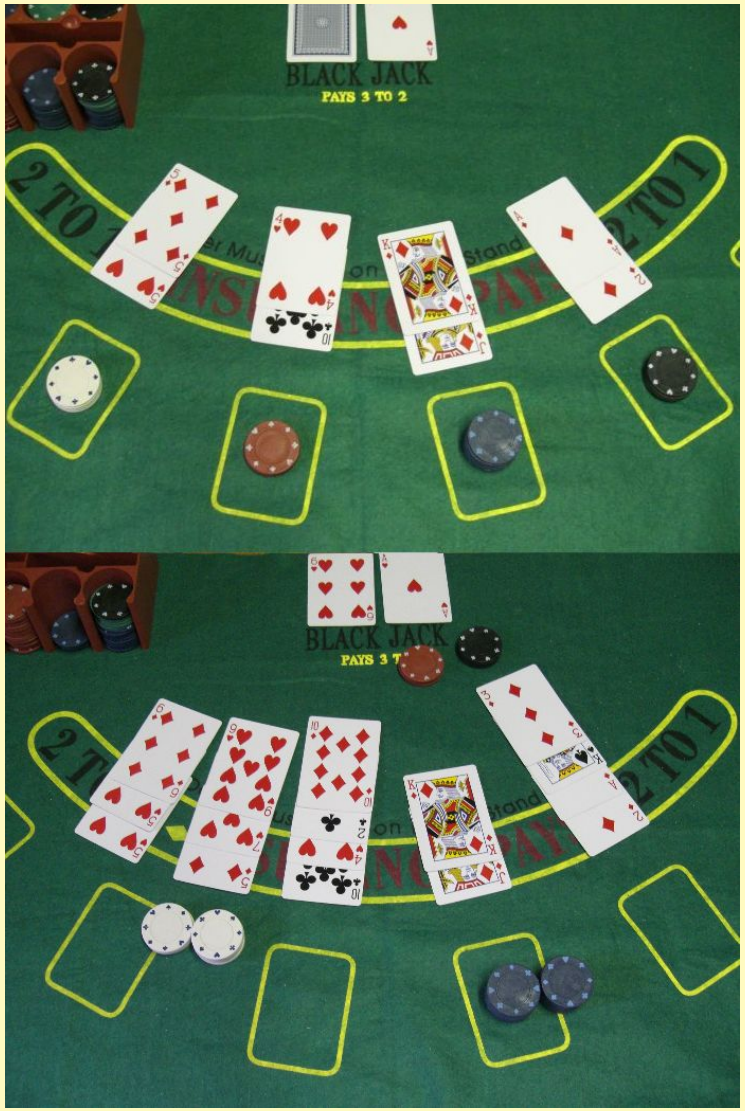
Question: What is $\sup_{p^*} \mathbf{E}_{p^*} R$?

Idea 1: Reinforcement Learning

(some) Rules:

- **Dealing**: player two cards, dealer one card (all face up).
- **Goal**: more points in hand. Face cards: 10, ace either 1 or 11 points.
- **Player's moves**: *hit* (take card, but try ≤ 21), *stand*, *double down*, *split* (in case of pair).
- **Dealer's moves**: deterministic, does not stand before ≥ 17 points.
- **Blackjack**: ace and face card $\rightarrow 1.5 \cdot$bet.

Idea 1: Reinforcement Learning



https://de.wikipedia.org/wiki/Black_Jack.JPG

Idea 1: Reinforcement Learning

Objects of Interest:

- sample from $p = p(X, Y, Z)$ (games),
- function of interest $\ell = \ell(X, Y, Z)$ (money) and
- p^* replacing $p(y \,|\, x) \to p^*(y \,|\, x)$ (strategy $=$ decisions $\mid$ game state).

Idea 1: Reinforcement Learning

Objects of Interest:

- sample from $p = p(X, Y, Z)$ (games),
- function of interest $\ell = \ell(X, Y, Z)$ (money) and
- p^* replacing $p(y \,|\, x) \to p^*(y \,|\, x)$ (strategy $=$ decisions $\mid$ game state).

Questions:

- What is $\mathbf{E}_{p^*} \ell$?

Idea 1: Reinforcement Learning

Objects of Interest:

- sample from $p = p(X, Y, Z)$ (games),
- function of interest $\ell = \ell(X, Y, Z)$ (money) and
- p^* replacing $p(y \,|\, x) \to p^*(y \,|\, x)$ (strategy = decisions | game state).

Questions:

- What is $\mathbf{E}_{p^*} \ell$?

Needed:

- Values of X_i, Y_i and $\ell(X_i, Y_i, Z_i)$ (under p)

X_i	Y_i	Z_i	$\ell(X_i, Y_i, Z_i)$
-1.4	2.0	?	2.1
-0.5	0.7	?	2.5
-0.8	1.5	?	2.6
$\vdots$	$\vdots$	$\vdots$	$\vdots$

X_i	Y_i	Z_i	$\ell(X_i, Y_i, Z_i)$
$\heartsuit K, \heartsuit 9$	hit	?	-1
$\clubsuit A, \spadesuit J$	stand	?	1.5
$\spadesuit 10, \heartsuit 8$	stand	?	-1
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Idea 1: Reinforcement Learning

Assume $p(y \,|\, x) \to p^*(y \,|\, x)$.

$$\begin{aligned}
\mathbf{E}_{p^*} \ell &= \int \ell(x,y,z)\; p^*(x,y,z)\; dx\, dy\, dz \\
&= \int \ell(x,y,z)\; \frac{p^*(x,y,z)}{p(x,y,z)}\; p(x,y,z)\; dx\, dy\, dz \\
&= \int \ell(x,y,z)\; \frac{p^*(y \,|\, x)}{p(y \,|\, x)}\; p(x,y,z)\; dx\, dy\, dz
\end{aligned}$$

Idea 1: Reinforcement Learning

Assume $p(y \mid x) \to p^*(y \mid x)$.

$$\begin{aligned}
\mathbf{E}_{p^*} \ell &= \int \ell(x,y,z)\; p^*(x,y,z)\; dx\, dy\, dz \\
&= \int \ell(x,y,z)\; \frac{p^*(x,y,z)}{p(x,y,z)}\; p(x,y,z)\; dx\, dy\, dz \\
&= \int \ell(x,y,z)\; \frac{p^*(y \mid x)}{p(y \mid x)}\; p(x,y,z)\; dx\, dy\, dz
\end{aligned}$$

Use estimator

$$\mathbf{E}_{p^*} \ell \approx \frac{1}{N} \sum_{i=1}^{N} \ell(X_i, Y_i, Z_i)\; \underbrace{\frac{p^*(Y_i \mid X_i)}{p(Y_i \mid X_i)}}_{w_i} = \frac{1}{N} \sum_{i=1}^{N} M_i$$

Idea 1: Reinforcement Learning

Assume $p(y \mid x) \to p^*(y \mid x)$.

$$\begin{aligned}
\mathbf{E}_{p^*} \ell &= \int \ell(x,y,z)\; p^*(x,y,z)\; dx\, dy\, dz \\
&= \int \ell(x,y,z)\; \frac{p^*(x,y,z)}{p(x,y,z)}\; p(x,y,z)\; dx\, dy\, dz \\
&= \int \ell(x,y,z)\; \frac{p^*(y \mid x)}{p(y \mid x)}\; p(x,y,z)\; dx\, dy\, dz
\end{aligned}$$

Use estimator

$$\mathbf{E}_{p^*} \ell \approx \frac{1}{N} \sum_{i=1}^{N} \ell(X_i, Y_i, Z_i)\; \underbrace{\frac{p^*(Y_i \mid X_i)}{p(Y_i \mid X_i)}}_{w_i} = \frac{1}{N} \sum_{i=1}^{N} M_i$$

Confidence intervals available!

Exercise-Blackjack:

a) Given strategies p and p^* in Blackjack. How do we compute the middle term of

$$\mathbf{E}_{p^*}\ell \approx \frac{1}{N}\sum_{i=1}^{N} \ell(X_i, Y_i, Z_i) \underbrace{\frac{p^*(Y_i \mid X_i)}{p(Y_i \mid X_i)}}_{w_i} = \frac{1}{N}\sum_{i=1}^{N} M_i$$

? Think a bit about implementation, too.

X_i	Y_i	Z_i	$\ell(X_i, Y_i, Z_i)$
$\heartsuit K, \heartsuit 9$	hit	?	-1
$\clubsuit A, \spadesuit J$	stand	?	1.5
$\spadesuit 10, \heartsuit 8$	stand	?	-1
$\vdots$	$\vdots$	$\vdots$	$\vdots$

b) What are potential problems?

Idea 1: Reinforcement Learning

$$p(y \mid x) \to p^*(y \mid x)$$

Which p^* is best?

Idea 1: Reinforcement Learning

$$p(y \mid x) \to p^*(y \mid x)$$

Which p^* is best? Parameterize and estimate

$$\nabla_\theta \mathbf{E}_{p_\theta}|_{\theta=\tilde{\theta}}$$

Idea 1: Reinforcement Learning

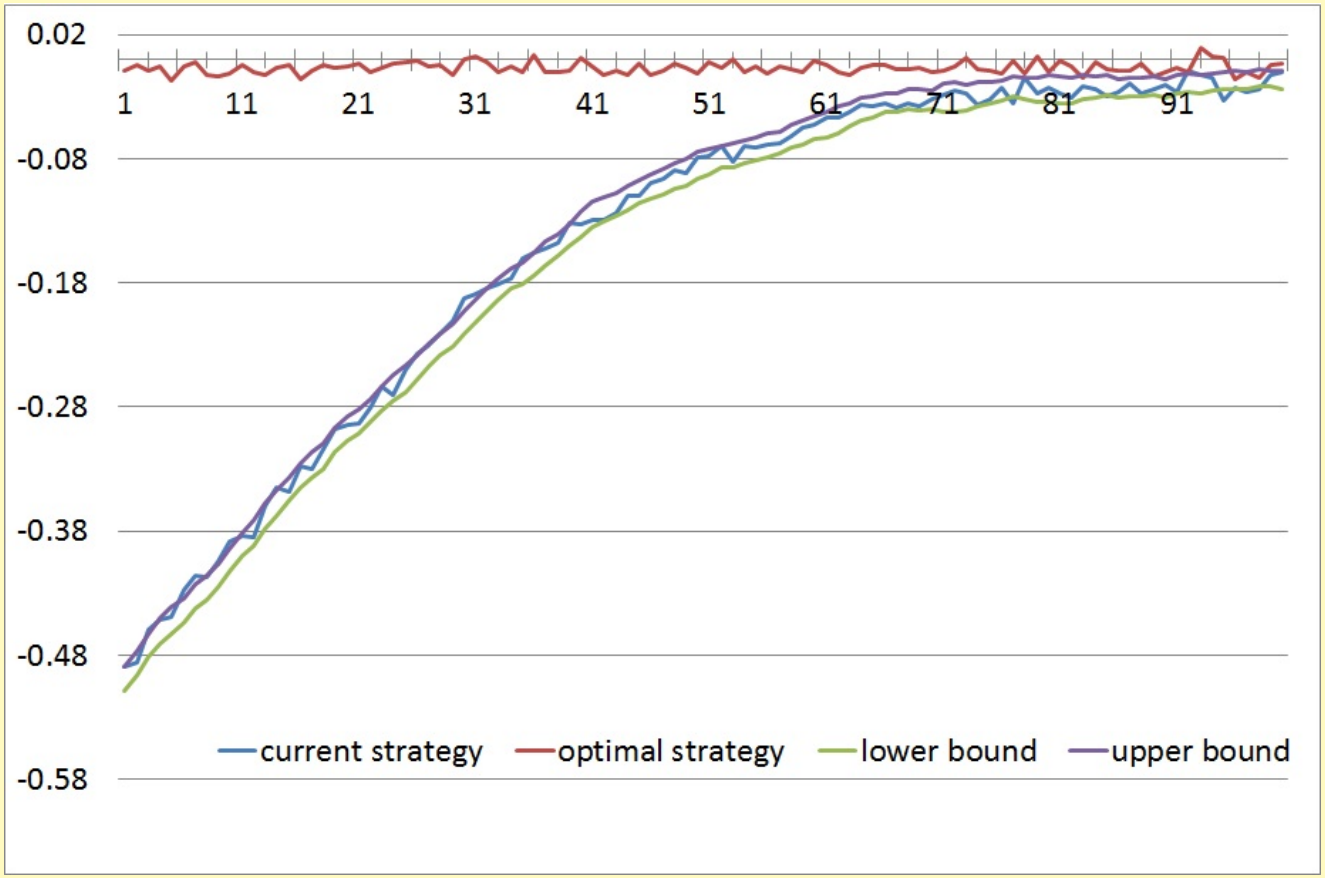
$$p(y \mid x) \to p^*(y \mid x)$$

Which p^* is best? Parameterize and estimate

$$\nabla_\theta \mathbf{E}_{p_\theta}|_{\theta=\tilde{\theta}}$$

Goal:	Optimize $\mathbf{E}_{p_\theta} \ell$
Idea:	Use gradient $\nabla_\theta \mathbf{E}_{p_\theta} \ell$ and optimize step-by-step.
Issues:	confidence intervals, step size, ...

Idea 1: Reinforcement Learning

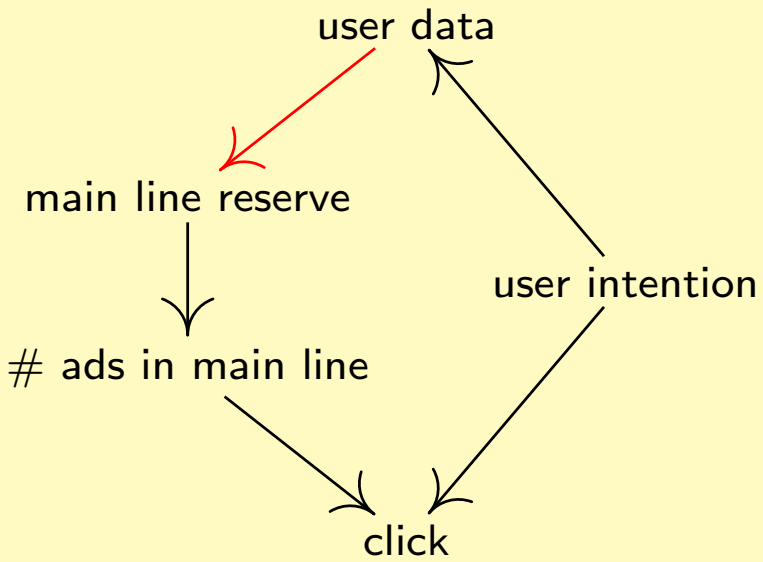


Idea 1: Reinforcement Learning

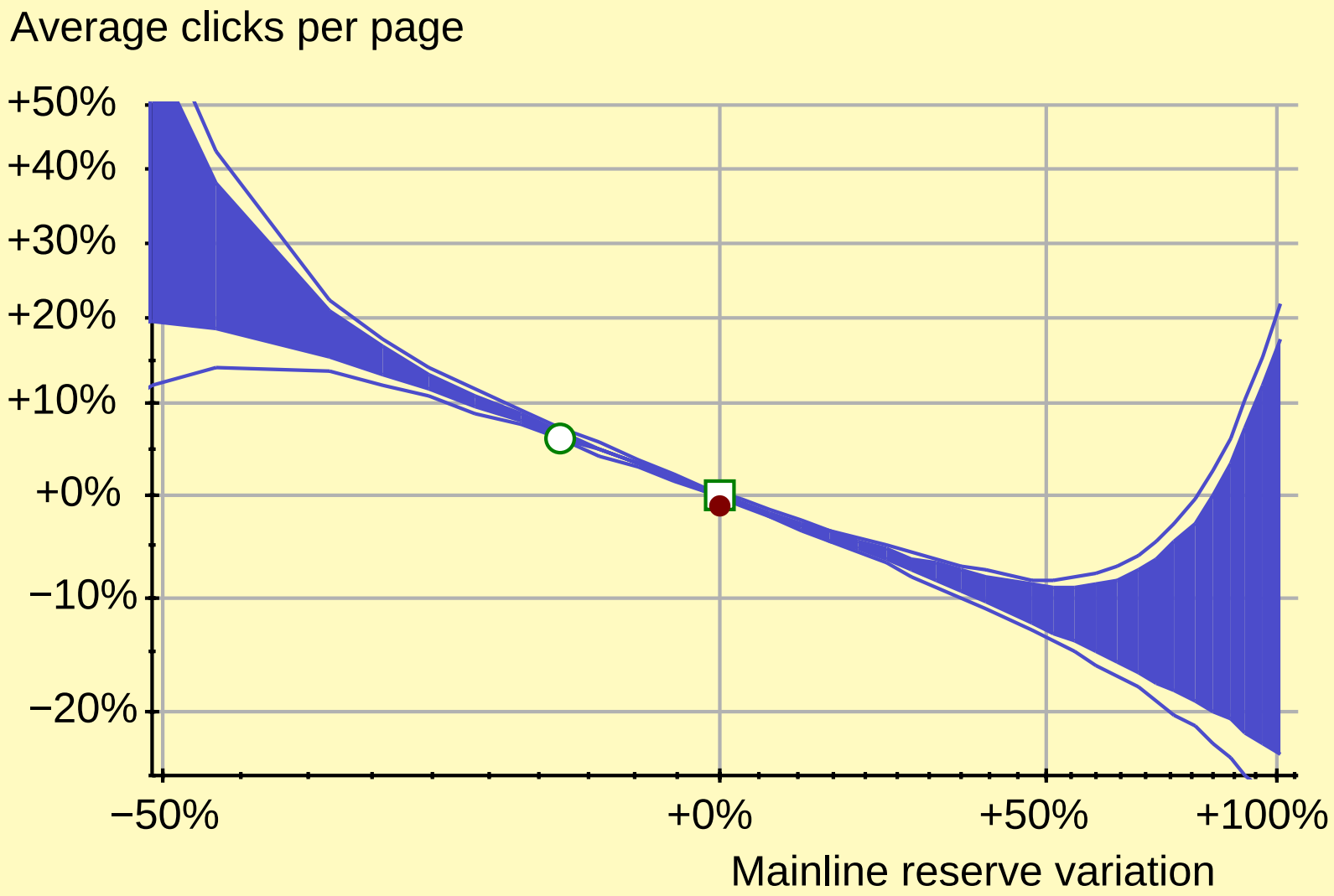
What can we do with $100,000$ samples?

	Online	Offline
reached strategy	$\mathbf{E}_{p^*}\ell \approx -5.1Ct$	$\mathbf{E}_{p^*}\ell \approx -5.8Ct$
irrelevant games	$33,653$	$61,048$
costs	$\$29,300$	$\$51,500$
speed	slow: probabilities	even slower: gradients

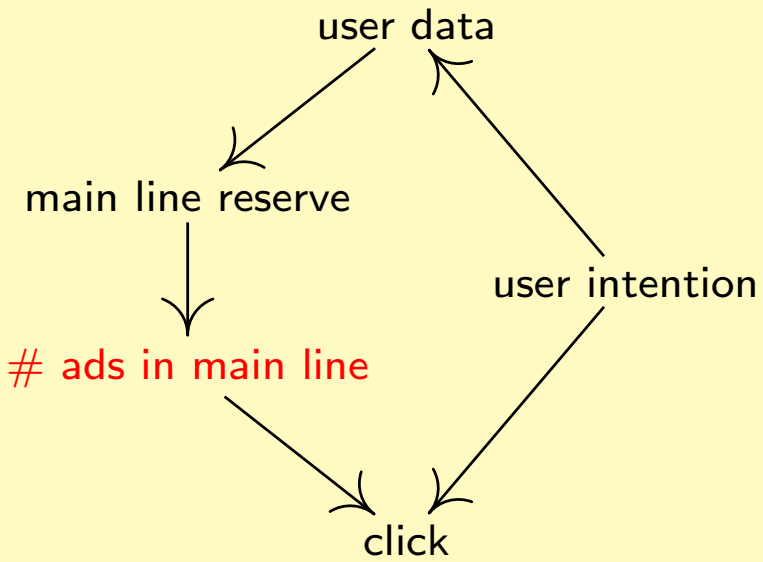
Idea 1: Reinforcement Learning



Idea 1: Reinforcement Learning

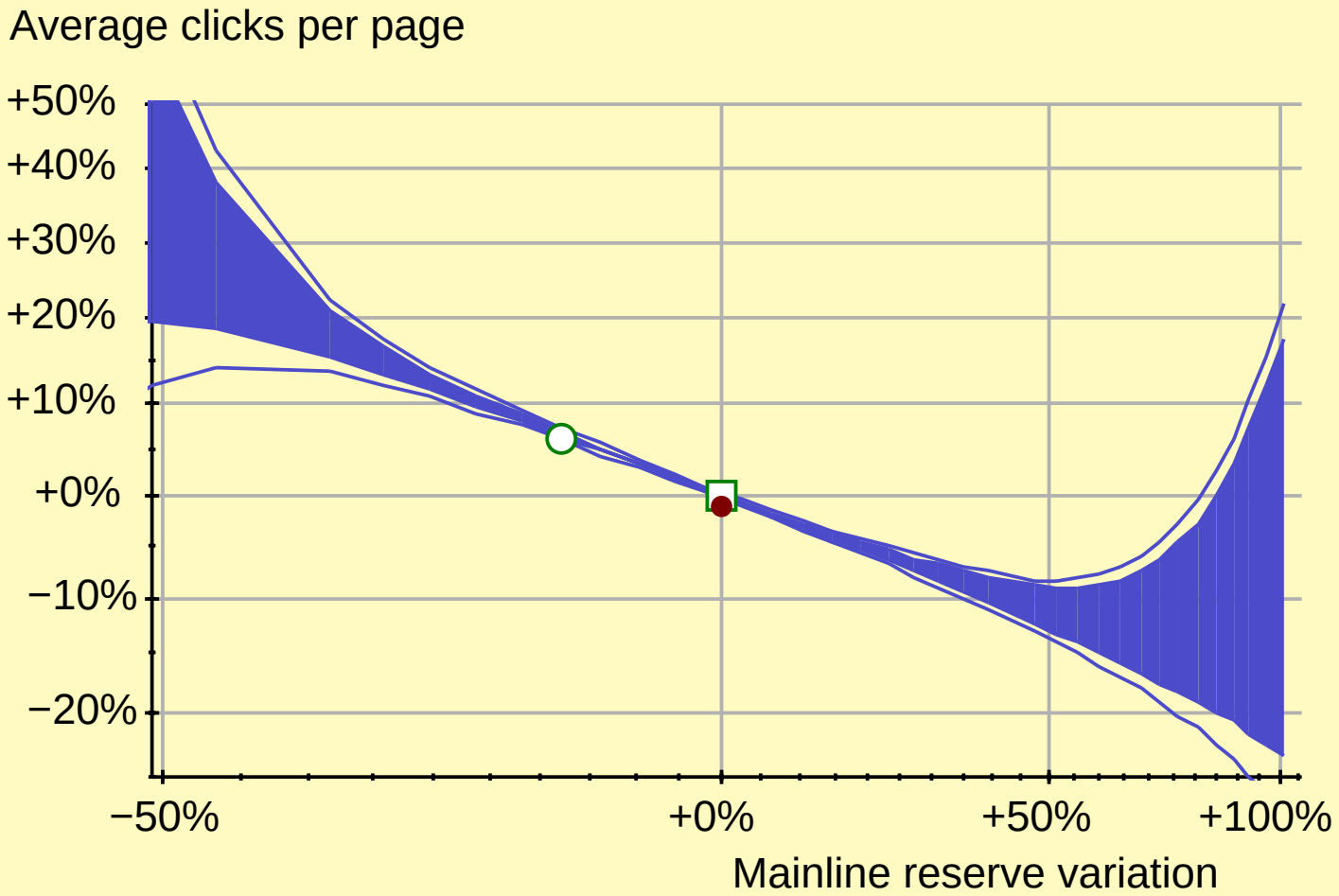


Idea 1: Reinforcement Learning



Idea 1: Reinforcement Learning

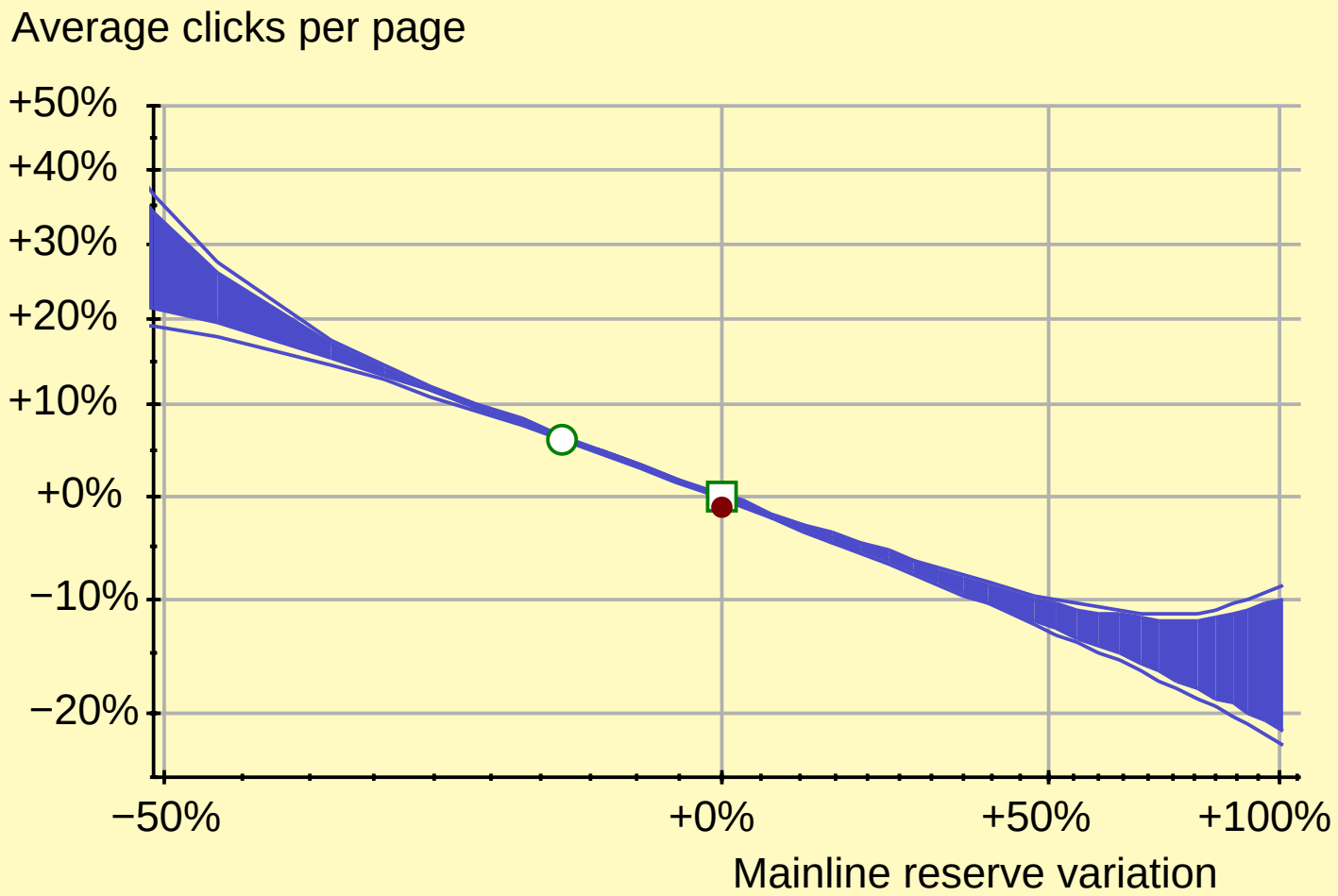
Old: (ads shown in mainline)



Bottou et al.: *Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising*, JMLR 2013

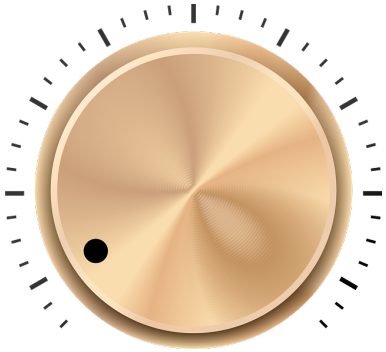
Idea 1: Reinforcement Learning

Using discrete variable (ads shown in mainline):



Bottou et al.: *Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising*, JMLR 2013

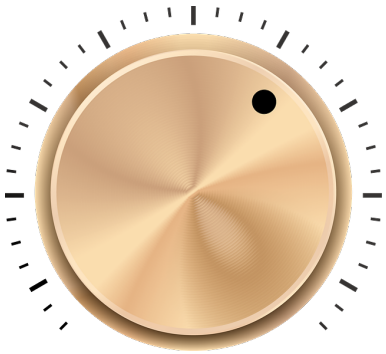
Idea 2: Anchor Regression

prediction  invariance

Idea 2: Anchor Regression

prediction  invariance

Idea 2: Anchor Regression

prediction  invariance

Idea 2: Anchor Regression

Anchor regression
$Y \in \mathbf{R}^1$: target
$X \in \mathbf{R}^{1\times d}$: predictors
$A \in \mathbf{R}^{1\times q}$: anchors, $\mathbf{E}_M A = 0$, $\mathbf{E}_M A^t A = Id$

Idea 2: Anchor Regression

Anchor regression
$Y \in \mathbf{R}^1$: target
$X \in \mathbf{R}^{1\times d}$: predictors
$A \in \mathbf{R}^{1\times q}$: anchors, $\mathbf{E}_M A = 0$, $\mathbf{E}_M A^t A = Id$

$$b^{\gamma}_{AR} := \operatorname*{argmin}_b \underbrace{\mathbf{E}_M (Y - Xb)^2}_{\text{prediction}} + \gamma \underbrace{\|\mathbf{E}_M A^t (Y - Xb)\|_2^2}_{\text{invariance}}$$

Idea 2: Anchor Regression

Anchor regression
$Y \in \mathbf{R}^1$: target
$X \in \mathbf{R}^{1\times d}$: predictors
$A \in \mathbf{R}^{1\times q}$: anchors, $\mathbf{E}_M A = 0$, $\mathbf{E}_M A^t A = Id$

$$b^{\gamma}_{AR} := \underset{b}{\operatorname{argmin}} \underbrace{\mathbf{E}_M (Y - Xb)^2}_{\text{prediction}} + \gamma \underbrace{\|\mathbf{E}_M A^t (Y - Xb)\|_2^2}_{\text{invariance}}$$

$\gamma \to 0$:	predictive model
$\gamma \to \infty$:	TSLS (minim. OLS solution if non-identif.)
$\gamma \in (0, \infty)$:	?

Idea 2: Anchor Regression

M, no intervention: $\begin{pmatrix} X \\ Y \\ H \end{pmatrix} := \mathbf{B} \cdot \begin{pmatrix} X \\ Y \\ H \end{pmatrix} + \varepsilon + \mathbf{K}A; \quad A := \epsilon_A$

$M(\delta)$, intervention on A: $\begin{pmatrix} X \\ Y \\ H \end{pmatrix} := \mathbf{B} \cdot \begin{pmatrix} X \\ Y \\ H \end{pmatrix} + \varepsilon + \mathbf{K}\delta; \quad A := \delta.$

$Id - \mathbf{B}$ invertible

Idea 2: Anchor Regression

$$M\text{, no intervention:} \quad \begin{pmatrix} X \\ Y \\ H \end{pmatrix} := \mathbf{B} \cdot \begin{pmatrix} X \\ Y \\ H \end{pmatrix} + \varepsilon + \mathbf{K}A; \quad A := \epsilon_A$$

$$M(\delta)\text{, intervention on } A: \quad \begin{pmatrix} X \\ Y \\ H \end{pmatrix} := \mathbf{B} \cdot \begin{pmatrix} X \\ Y \\ H \end{pmatrix} + \varepsilon + \mathbf{K}\delta; \quad A := \delta.$$

$Id - \mathbf{B}$ invertible

Proposition

We have

$$b^{\gamma}_{AR} = \operatorname*{argmin}_{b} \max_{\delta \in C^{\gamma}} E_{M(\delta)}[(Y - Xb)^2],$$

Idea 2: Anchor Regression

$$M\text{, no intervention:} \quad \begin{pmatrix} X \\ Y \\ H \end{pmatrix} := \mathbf{B} \cdot \begin{pmatrix} X \\ Y \\ H \end{pmatrix} + \varepsilon + \mathbf{K}A; \quad A := \epsilon_A$$

$$M(\delta)\text{, intervention on } A: \quad \begin{pmatrix} X \\ Y \\ H \end{pmatrix} := \mathbf{B} \cdot \begin{pmatrix} X \\ Y \\ H \end{pmatrix} + \varepsilon + \mathbf{K}\delta; \quad A := \delta.$$

$Id - \mathbf{B}$ invertible

Proposition

We have

$$b_{AR}^{\gamma} = \operatorname*{argmin}_b \max_{\delta \in C^{\gamma}} E_{M(\delta)}[(Y - Xb)^2],$$

where

$$C^{\gamma} := \{\delta \text{ such that } \|\delta\|_2 \leq \sqrt{\gamma}\}.$$

D. Rothenhäusler, P. Bühlmann, N. Meinshausen, JP: *Anchor regression: heterogeneous data meets causality*, JRSS-B 2020+,

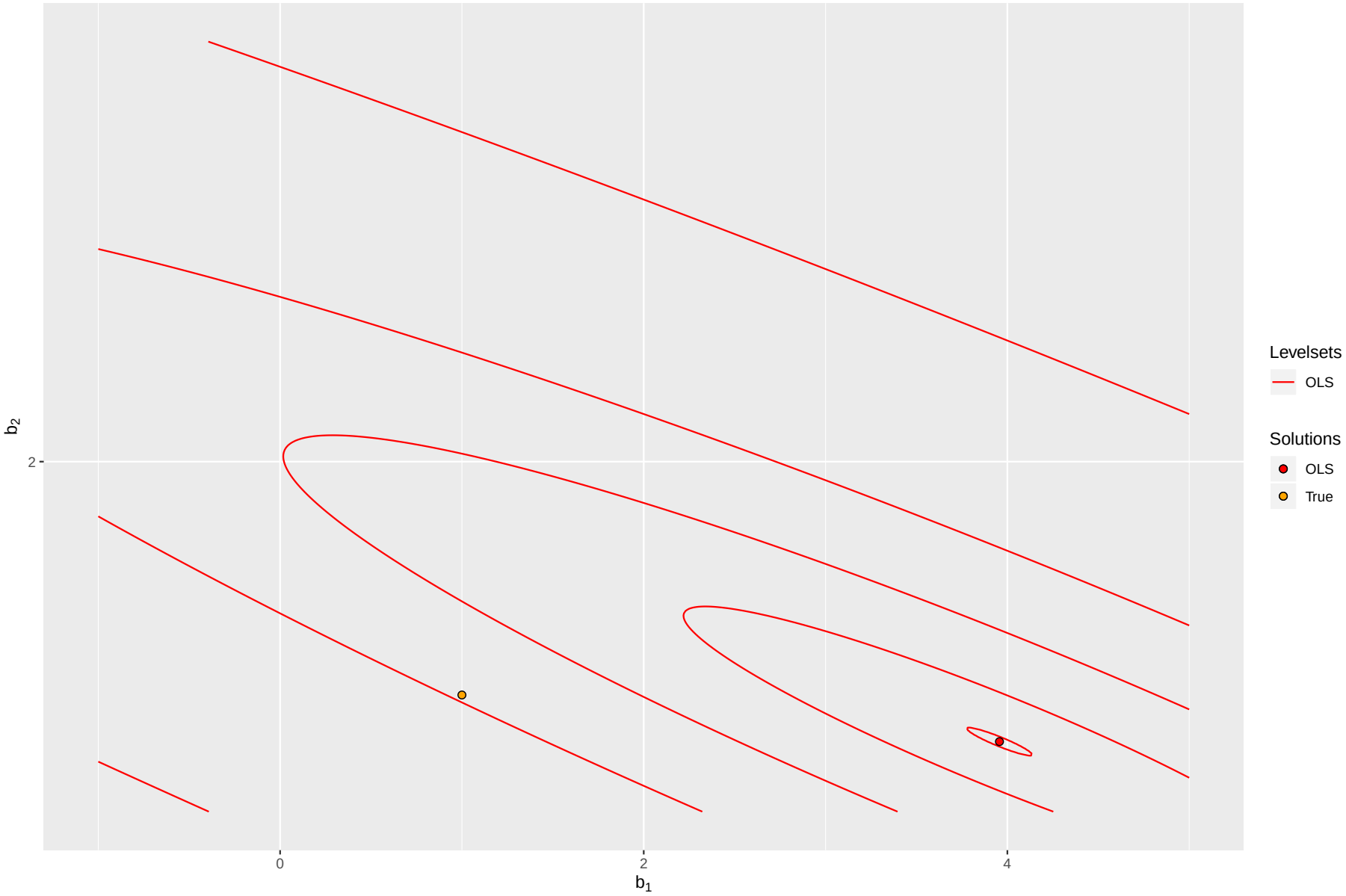
Idea 2: Anchor Regression

What do we do for a finite sample?

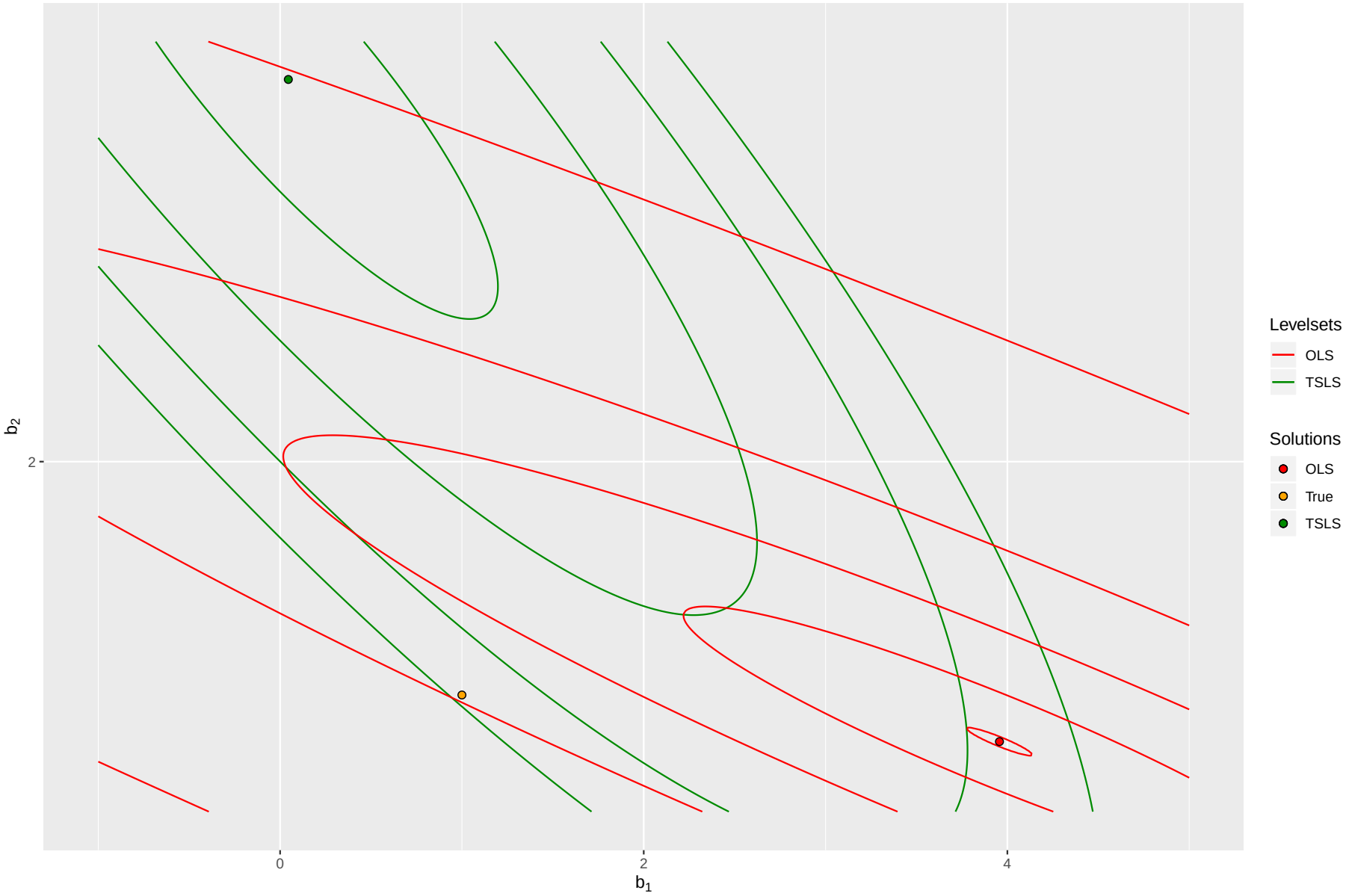
$$b^\gamma := \operatorname*{argmin}_b \underbrace{\mathbf{E}_M(Y - Xb)^2}_{\text{prediction}} + \gamma \underbrace{\|\mathbf{E}_M A^t(Y - Xb)\|_2^2}_{\text{invariance}}$$

$$\hat{b}_n^\gamma := \operatorname*{argmin}_b (\mathbf{Y} - \mathbf{X}b)^\top(\mathbf{Y} - \mathbf{X}b) + \gamma(\mathbf{Y} - \mathbf{X}b)^\top \mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top(\mathbf{Y} - \mathbf{X}b)$$

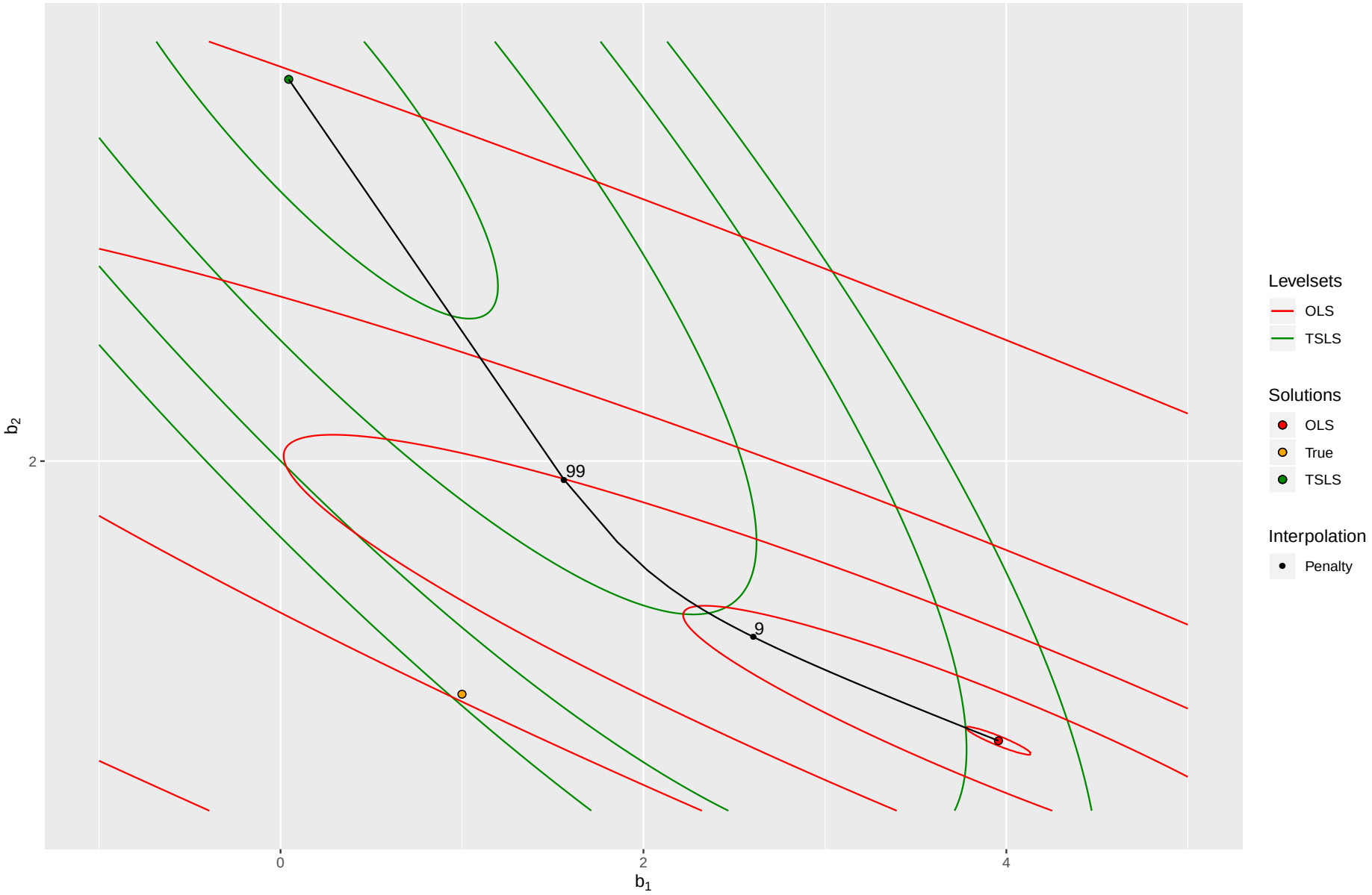
Idea 2: Anchor Regression



Idea 2: Anchor Regression



Idea 2: Anchor Regression



Anchor regr. can be motivated by improved finite sample properties, too.

It is a k-class estimator (cf. also LIML and Fuller).

e.g., Anderson and Rubin 1949 and Theil 1958 and Fuller 1977

Jakobsen and JP (arXiv:2005.03353)

Anchor regr. can be motivated by improved finite sample properties, too.

It is a k-class estimator (cf. also LIML and Fuller).

e.g., Anderson and Rubin 1949 and Theil 1958 and Fuller 1977

Jakobsen and JP (arXiv:2005.03353)

How can we choose the gamma?

Anchor regr. can be motivated by improved finite sample properties, too.

It is a k-class estimator (cf. also LIML and Fuller).

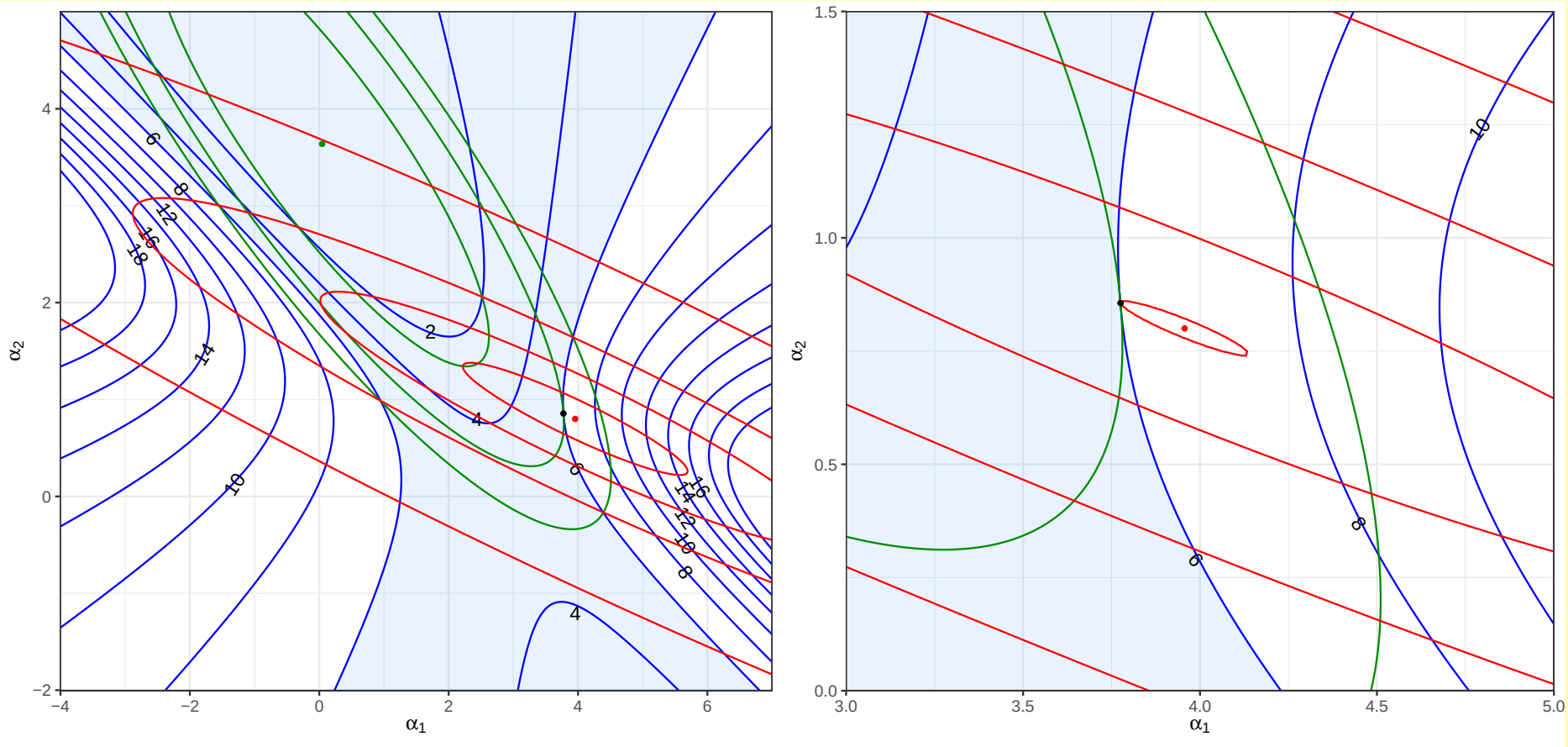
e.g., Anderson and Rubin 1949 and Theil 1958 and Fuller 1977

Jakobsen and JP (arXiv:2005.03353)

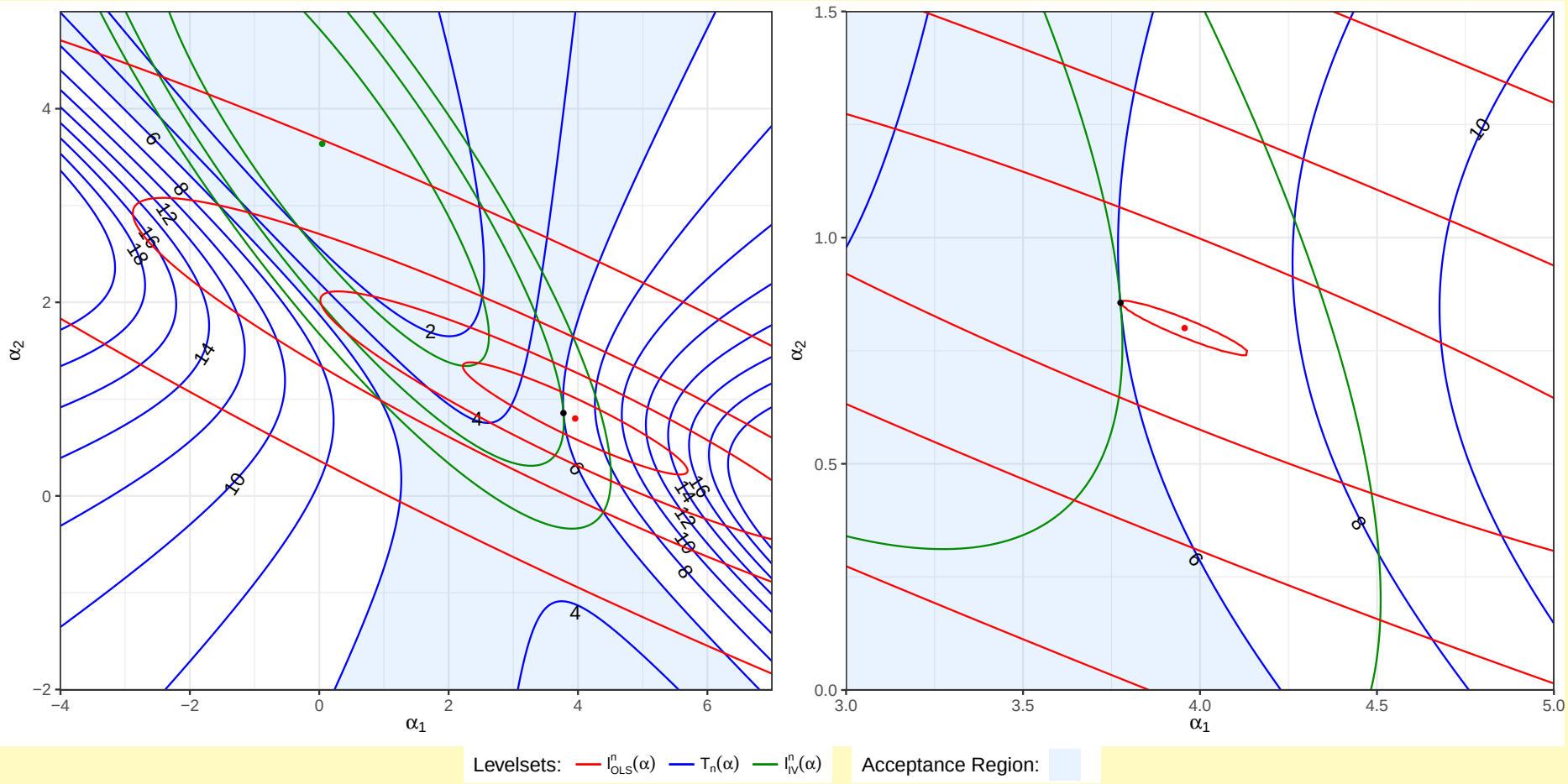
How can we choose the gamma? Idea: PULSE (p-uncorr. least sq. est.)

Choose γ, such that `cor.test(`$\mathbf{A}, \mathbf{Y} - \mathbf{X}\hat{b}_n^\gamma$`).pvalue == 0.05`

Idea 2: Anchor Regression

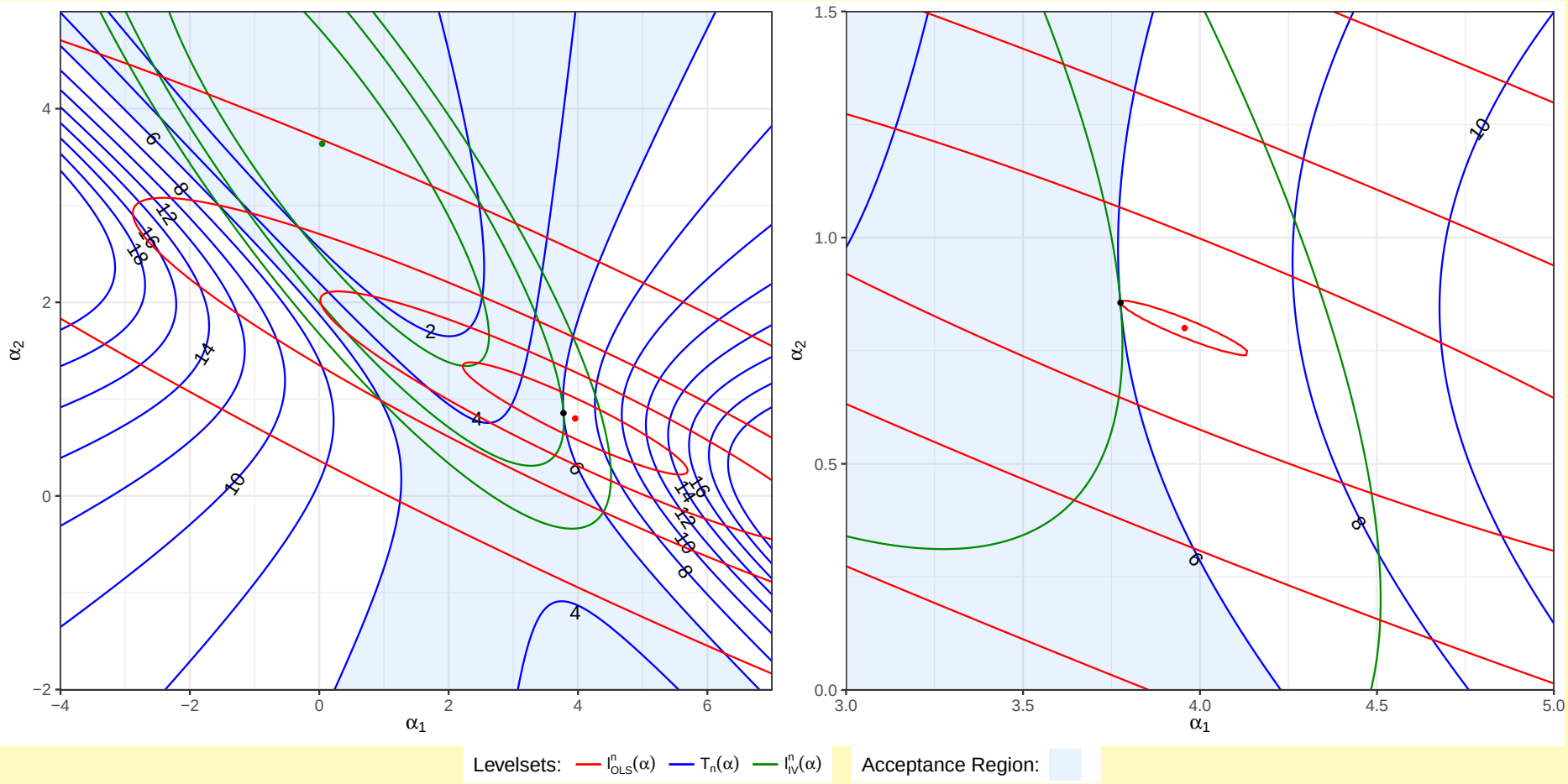


Idea 2: Anchor Regression



Prop: Optimization problem can be solved efficiently.

Idea 2: Anchor Regression



Prop: Optimization problem can be solved efficiently.

Prop: Assume identifiable IV setting. Then PULSE $\to_P$ causal parameter.

Jakobsen and JP: Distributional Robustness of K-class Estimators and the PULSE (arXiv:2005.03353)

So far: linear models.

So far: linear models.

1. Can such a trade-off ever be useful in practice?
2. Do the ideas extend to nonlinear models?

So far: linear models.

1. Can such a trade-off ever be useful in practice?
2. Do the ideas extend to nonlinear models?

Idea 3: CausalKinetiX

Example: Maillard reaction Glu, Mel, C5, ForAc, Triose, Cn, AcAc, Amad, lysR, Fru, AMP

$$\frac{d}{dt}[\mathrm{Glu}]_t = -\theta_1[\mathrm{Glu}]_t + \theta_2[\mathrm{Fru}]_t - \theta_3[\mathrm{lysR}]_t[\mathrm{Glu}]_t$$
$$\frac{d}{dt}[\mathrm{Mel}]_t = \theta_4[\mathrm{AMP}]_t \qquad \dots$$

Idea 3: CausalKinetiX

Example: Maillard reaction Glu, Mel, C5, ForAc, Triose, Cn, AcAc, Amad, lysR, Fru, AMP

http://www.srfcdn.ch/radio/modules/dynimages/624/srf-1/2015/01/diverses/264377_150114_raclette_key.jpg

Idea 3: CausalKinetiX

Example: Maillard reaction Glu, Mel, C5, ForAc, Triose, Cn, AcAc, Amad, lysR, Fru, AMP

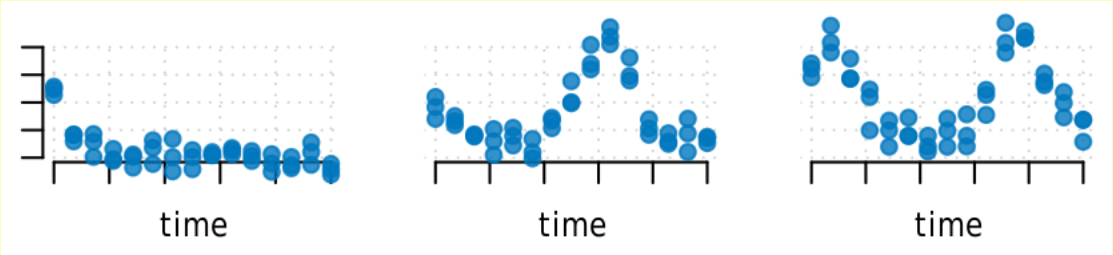
$$\frac{d}{dt}[\mathrm{Glu}]_t = -\theta_1[\mathrm{Glu}]_t + \theta_2[\mathrm{Fru}]_t - \theta_3[\mathrm{lysR}]_t[\mathrm{Glu}]_t$$
$$\frac{d}{dt}[\mathrm{Mel}]_t = \theta_4[\mathrm{AMP}]_t \qquad \dots$$

Idea 3: CausalKinetiX

Example: Maillard reaction Glu, Mel, C5, ForAc, Triose, Cn, AcAc, Amad, lysR, Fru, AMP

$$\frac{d}{dt}[\mathrm{Glu}]_t = -\theta_1[\mathrm{Glu}]_t + \theta_2[\mathrm{Fru}]_t - \theta_3[\mathrm{lysR}]_t[\mathrm{Glu}]_t$$

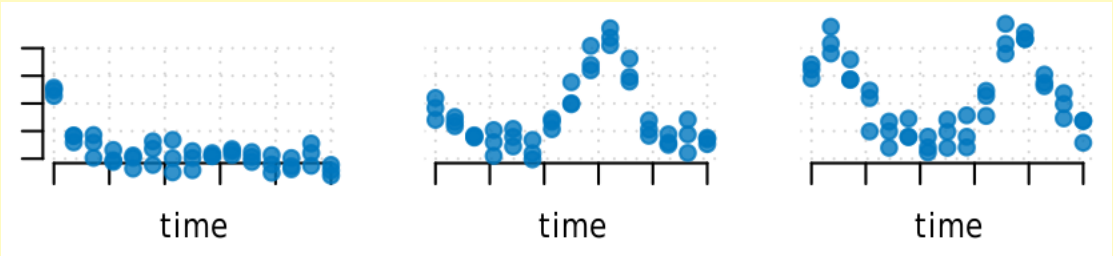
$$\frac{d}{dt}[\mathrm{Mel}]_t = \theta_4[\mathrm{AMP}]_t \qquad \dots$$



Idea 3: CausalKinetiX

Example: Maillard reaction Glu, Mel, C5, ForAc, Triose, Cn, AcAc, Amad, lysR, Fru, AMP

$$\frac{d}{dt}[\mathrm{Glu}]_t = -\theta_1[\mathrm{Glu}]_t + \theta_2[\mathrm{Fru}]_t - \theta_3[\mathrm{lysR}]_t[\mathrm{Glu}]_t$$
$$\frac{d}{dt}[\mathrm{Mel}]_t = \theta_4[\mathrm{AMP}]_t \qquad \dots$$



If the network structure is unknown, can we recover it from data?

Brands et al. (2002): "Kinetic model. of reactions in heated monosaccharide-casein systems.", J. of agr. and food chemistry.

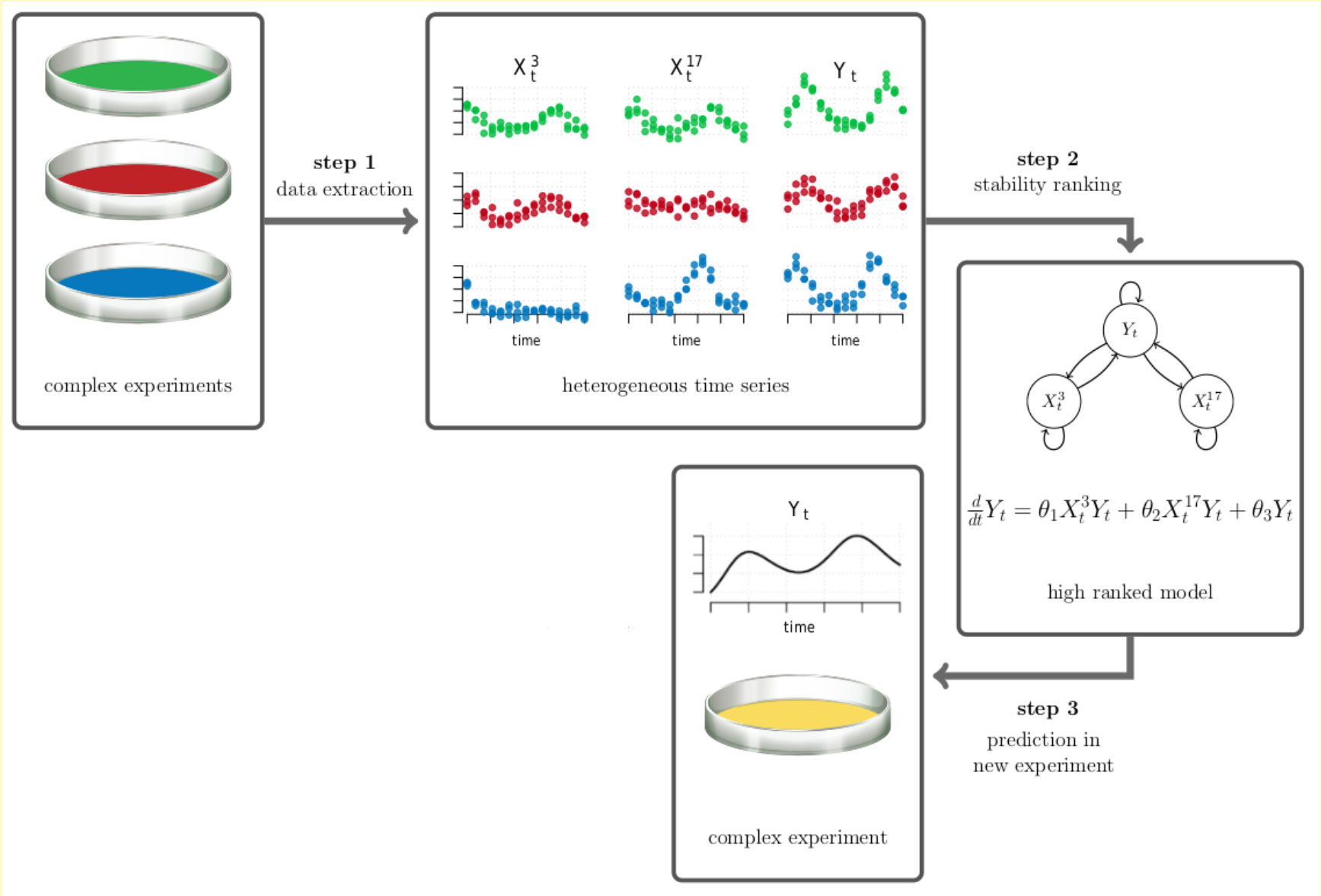
Idea 3: CausalKinetiX

King et al, Nature 2004, Bongard et al, PNAS 2007, Calderhead et al, NIPS 2009, Schmidt et al, Science 2009, Oates, Mukherjee, AoAS 2012, Babtie et al, PNAS 2014, Hill et al, Nature Methods 2016, Chen et al, JASA 2016, Rudy et al, Science Advances 2017, Brunten et al, Nature Comm 2017, Mikkelsen, Hansen, arXiv 1710.09308, many more...

Often: nonlinear least squares

$$\operatorname*{argmin}_{\theta} \sum_{t \in \{t_1, \ldots, t_m\}} \|\tilde{\mathbf{X}}_t - \mathbf{X}_t\|_2^2, \text{ where } \dot{\mathbf{X}}_t = g_\theta(\mathbf{X}_t, t), \mathbf{X_0} = \mathbf{x_0}$$

Idea 3: CausalKinetiX



N. Pfister, S. Bauer, JP: *Learning stable structures in kinetic systems: benefits of a causal approach*, PNAS 2019

Idea 3: CausalKinetiX

Real data: $(Y, X^1, \ldots, X^{411})$, 11 time points, 5 exp., 3 rep.;

Idea 3: CausalKinetiX

Real data: $(Y, X^1, \ldots, X^{411})$, 11 time points, 5 exp., 3 rep.; $Z_t := 2 - Y_t$

Idea 3: CausalKinetiX

Real data: $(Y, X^1, \ldots, X^{411})$, 11 time points, 5 exp., 3 rep.; $Z_t := 2 - Y_t$

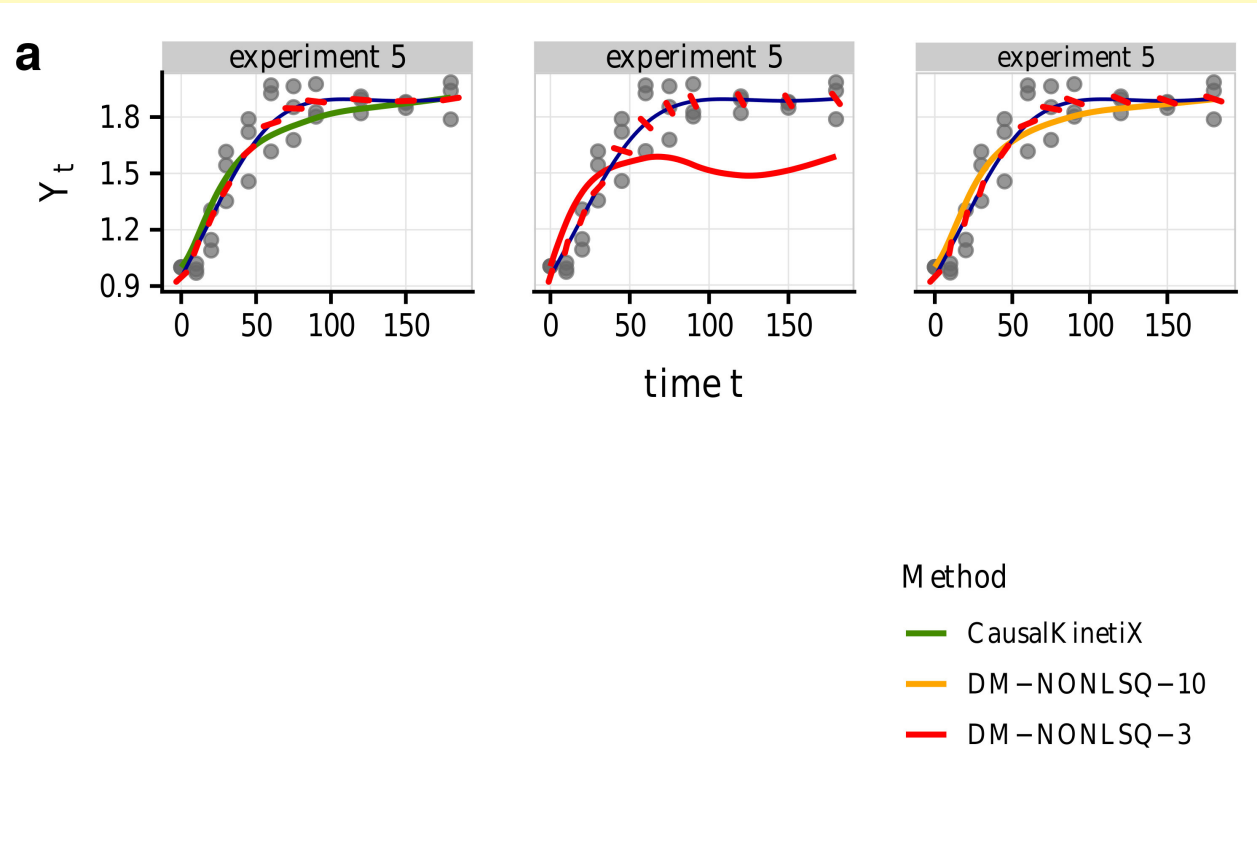
top ranked model $\dot{Y}_t = \theta_1 Z_t X_t^{56} X_t^{122} + \theta_2 Z_t X_t^{128} X_t^{168} - \theta_3 Y_t X_t^{33} X_t^{138}$

Idea 3: CausalKinetiX

Real data: $(Y, X^1, \ldots, X^{411})$, 11 time points, 5 exp., 3 rep.; $Z_t := 2 - Y_t$

top ranked model $\dot{Y}_t = \theta_1 Z_t X_t^{56} X_t^{122} + \theta_2 Z_t X_t^{128} X_t^{168} - \theta_3 Y_t X_t^{33} X_t^{138}$

In-sample plot

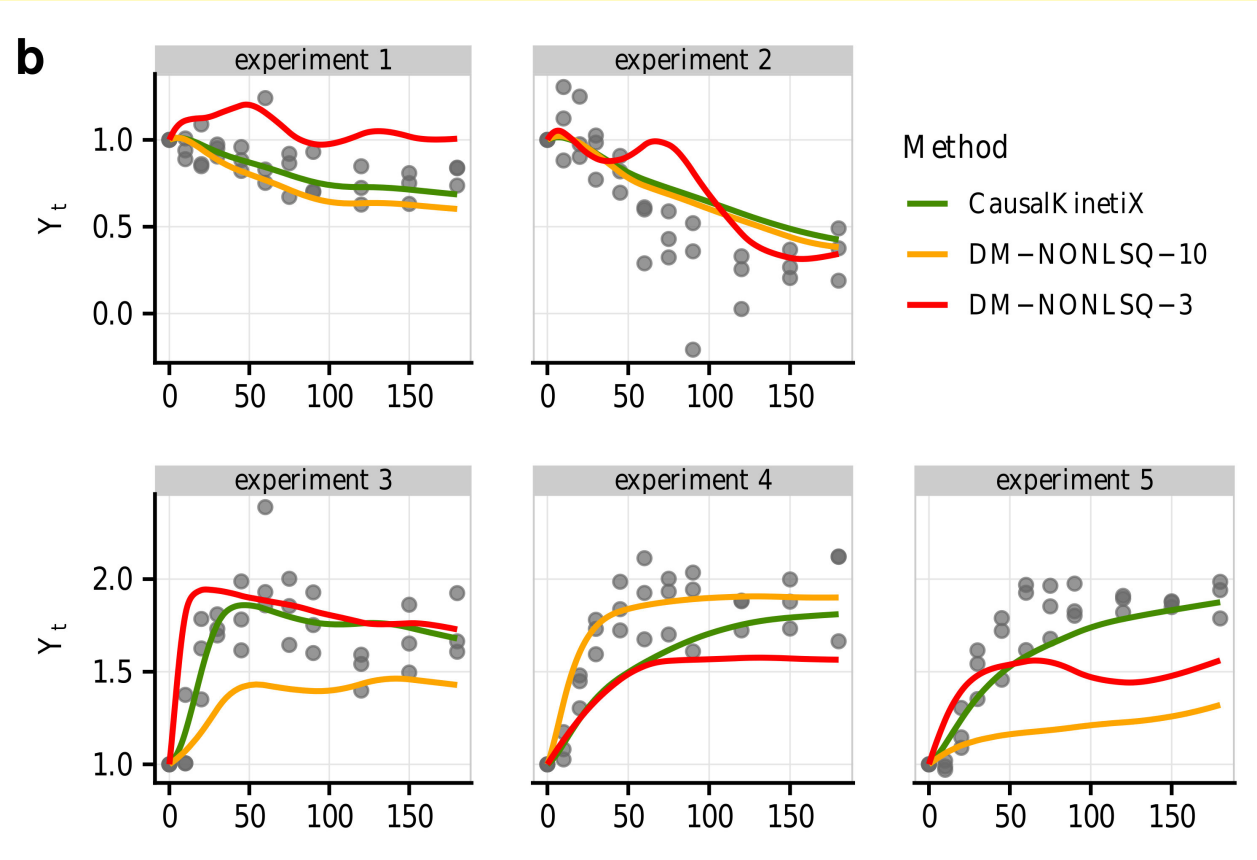


Idea 3: CausalKinetiX

Real data: $(Y, X^1, \ldots, X^{411})$, 11 time points, 5 exp., 3 rep.; $Z_t := 2 - Y_t$

top ranked model $\dot{Y}_t = \theta_1 Z_t X_t^{56} X_t^{122} + \theta_2 Z_t X_t^{128} X_t^{168} - \theta_3 Y_t X_t^{33} X_t^{138}$

Out-of-sample plot



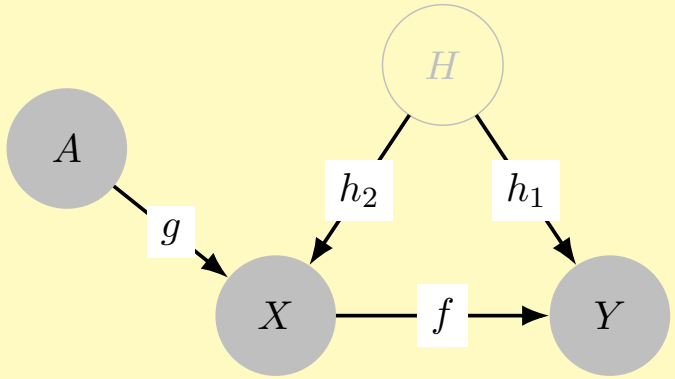
So far: linear models.

1. Can such a trade-off ever be useful in practice?
2. Do the ideas extend to nonlinear models?

Idea 4: Distribution Generalization

Specify an SCM by a model $M = (f, g, h_1, h_2, Q) \in \mathcal{M}$:

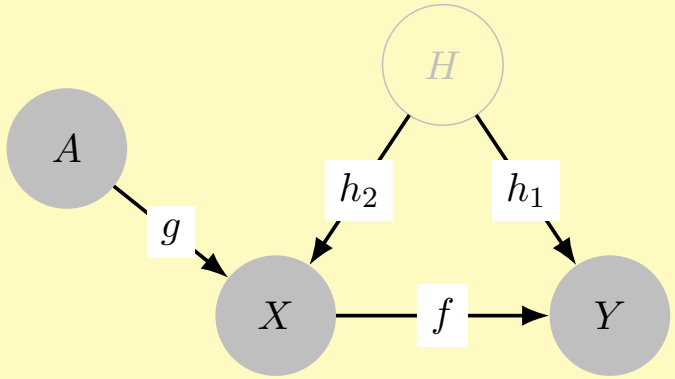
$$
\begin{aligned}
A &:= \epsilon_A \\
H &:= \epsilon_H \\
X &:= g(A) + h_2(H, \epsilon_X) \\
Y &:= f(X) + h_1(H, \epsilon_Y) \\
&(\epsilon_X, \epsilon_Y, \epsilon_A, \epsilon_H) \sim Q \text{ jointly indep..}
\end{aligned}
$$



Idea 4: Distribution Generalization

Specify an SCM by a model $M = (f, g, h_1, h_2, Q) \in \mathcal{M}$:

$$
\begin{aligned}
A &:= \epsilon_A \\
H &:= \epsilon_H \\
X &:= g(A) + h_2(H, \epsilon_X) \\
Y &:= f(X) + h_1(H, \epsilon_Y) \\
&(\epsilon_X, \epsilon_Y, \epsilon_A, \epsilon_H) \sim Q \text{ jointly indep..}
\end{aligned}
$$

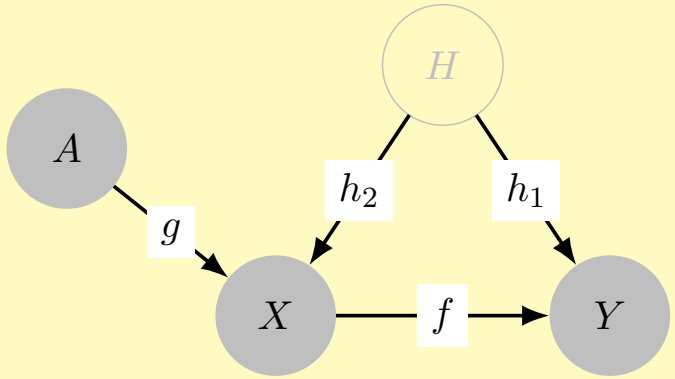


Goal: $\operatorname{argmin}_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} E_{M(i)}[(Y - f_\diamond(X))^2]$.

Idea 4: Distribution Generalization

Specify an SCM by a model $M = (f, g, h_1, h_2, Q) \in \mathcal{M}$:

$$A := \epsilon_A$$
$$H := \epsilon_H$$
$$X := g(A) + h_2(H, \epsilon_X)$$
$$Y := f(X) + h_1(H, \epsilon_Y)$$
$$(\epsilon_X, \epsilon_Y, \epsilon_A, \epsilon_H) \sim Q \text{ jointly indep..}$$



Goal: $\operatorname{argmin}_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} E_{M(i)}[(Y - f_\diamond(X))^2]$.

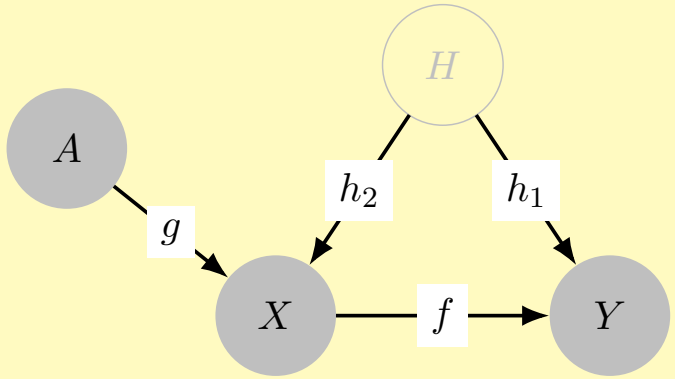
Definition

$(P_M, \mathcal{M})$ is said to generalize to $\mathcal{I}$ 'if' there exists a function $f^* \in \mathcal{F}$ s.t.

Idea 4: Distribution Generalization

Specify an SCM by a model $M = (f, g, h_1, h_2, Q) \in \mathcal{M}$:

$$
\begin{aligned}
A &:= \epsilon_A \\
H &:= \epsilon_H \\
X &:= g(A) + h_2(H, \epsilon_X) \\
Y &:= f(X) + h_1(H, \epsilon_Y) \\
(\epsilon_X, \epsilon_Y, \epsilon_A, \epsilon_H) &\sim Q \text{ jointly indep..}
\end{aligned}
$$



Goal: $\operatorname{argmin}_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} E_{M(i)}[(Y - f_\diamond(X))^2]$.

Definition

$(P_M, \mathcal{M})$ is said to generalize to $\mathcal{I}$ 'if' there exists a function $f^* \in \mathcal{F}$ s.t.

$$f^* = \operatorname*{argmin}_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} E_{\tilde{M}(i)}\left[(Y - f_\diamond(X))^2\right]$$

for all models $\tilde{M} \in \mathcal{M}$ with $P_{\tilde{M}} = P_M$.

Idea 4: Distribution Generalization

Recall: SCM assignment for X: $X := g(A) + h_2(H, \epsilon_X)$
Recall: SCM assignment for Y: $Y := f(X) + h_1(H, \epsilon_Y)$

Idea 4: Distribution Generalization

Recall: SCM assignment for X: $X := g(A) + h_2(H, \epsilon_X)$
Recall: SCM assignment for Y: $Y := f(X) + h_1(H, \epsilon_Y)$

A1: Identif. of f: $\forall \tilde{M} = (\tilde{f}, \ldots)$ s.t. $P_{\tilde{M}} = P_M$, we have $\tilde{f} = f$ on $\mathrm{supp}(X)$.
A2: Extrapolation of f: If $f = \tilde{f}$ on $\mathrm{supp}(X)$, then $f = \tilde{f}$ on $\mathrm{supp}_{\mathcal{I}}(X)$.
A3: Same for g: $\forall \tilde{M} = (\tilde{f}, \tilde{g}, \ldots)$ s.t. $P_{\tilde{M}} = P_M$, $\tilde{g} = g$ on $\mathrm{supp}(A) \cup \mathrm{supp}_{\mathcal{I}}(A)$.

Idea 4: Distribution Generalization

Recall: SCM assignment for X: $X := g(A) + h_2(H, \epsilon_X)$
Recall: SCM assignment for Y: $Y := f(X) + h_1(H, \epsilon_Y)$

A1: Identif. of f: $\forall \tilde{M} = (\tilde{f}, \ldots)$ s.t. $P_{\tilde{M}} = P_M$, we have $\tilde{f} = f$ on $\mathrm{supp}(X)$.
A2: Extrapolation of f: If $f = \tilde{f}$ on $\mathrm{supp}(X)$, then $f = \tilde{f}$ on $\mathrm{supp}_{\mathcal{I}}(X)$.
A3: Same for g: $\forall \tilde{M} = (\tilde{f}, \tilde{g}, \ldots)$ s.t. $P_{\tilde{M}} = P_M$, $\tilde{g} = g$ on $\mathrm{supp}(A) \cup \mathrm{supp}_{\mathcal{I}}(A)$.

intervention	$\mathrm{supp}_{\mathcal{I}}(X)$	assumptions	distr. gen. possible?
on X (well-behaved)	$\subseteq \mathrm{supp}(X)$	A1	✓

Idea 4: Distribution Generalization

Recall: SCM assignment for X: $X := g(A) + h_2(H, \epsilon_X)$
Recall: SCM assignment for Y: $Y := f(X) + h_1(H, \epsilon_Y)$

A1: Identif. of f: $\forall \tilde{M} = (\tilde{f}, \ldots)$ s.t. $P_{\tilde{M}} = P_M$, we have $\tilde{f} = f$ on $\mathrm{supp}(X)$.
A2: Extrapolation of f: If $f = \tilde{f}$ on $\mathrm{supp}(X)$, then $f = \tilde{f}$ on $\mathrm{supp}_{\mathcal{I}}(X)$.
A3: Same for g: $\forall \tilde{M} = (\tilde{f}, \tilde{g}, \ldots)$ s.t. $P_{\tilde{M}} = P_M$, $\tilde{g} = g$ on $\mathrm{supp}(A) \cup \mathrm{supp}_{\mathcal{I}}(A)$.

intervention	$\mathrm{supp}_{\mathcal{I}}(X)$	assumptions	distr. gen. possible?
on X (well-behaved)	$\subseteq \mathrm{supp}(X)$	A1	✓
on X (well-behaved)	$\not\subseteq \mathrm{supp}(X)$	A1 and A2	✓

Idea 4: Distribution Generalization

Recall: SCM assignment for X: $X := g(A) + h_2(H, \epsilon_X)$
Recall: SCM assignment for Y: $Y := f(X) + h_1(H, \epsilon_Y)$

A1: Identif. of f: $\forall \tilde{M} = (\tilde{f}, \ldots)$ s.t. $P_{\tilde{M}} = P_M$, we have $\tilde{f} = f$ on $\mathrm{supp}(X)$.
A2: Extrapolation of f: If $f = \tilde{f}$ on $\mathrm{supp}(X)$, then $f = \tilde{f}$ on $\mathrm{supp}_{\mathcal{I}}(X)$.
A3: Same for g: $\forall \tilde{M} = (\tilde{f}, \tilde{g}, \ldots)$ s.t. $P_{\tilde{M}} = P_M$, $\tilde{g} = g$ on $\mathrm{supp}(A) \cup \mathrm{supp}_{\mathcal{I}}(A)$.

intervention	$\mathrm{supp}_{\mathcal{I}}(X)$	assumptions	distr. gen. possible?
on X (well-behaved)	$\subseteq \mathrm{supp}(X)$	A1	✓
on X (well-behaved)	$\not\subseteq \mathrm{supp}(X)$	A1 and A2	✓
on X (well-behaved)	$\not\subseteq \mathrm{supp}(X)$	A1	✗

Idea 4: Distribution Generalization

Recall: SCM assignment for X: $X := g(A) + h_2(H, \epsilon_X)$
Recall: SCM assignment for Y: $Y := f(X) + h_1(H, \epsilon_Y)$

A1: Identif. of f: $\forall \tilde{M} = (\tilde{f}, \ldots)$ s.t. $P_{\tilde{M}} = P_M$, we have $\tilde{f} = f$ on $\mathrm{supp}(X)$.
A2: Extrapolation of f: If $f = \tilde{f}$ on $\mathrm{supp}(X)$, then $f = \tilde{f}$ on $\mathrm{supp}_{\mathcal{I}}(X)$.
A3: Same for g: $\forall \tilde{M} = (\tilde{f}, \tilde{g}, \ldots)$ s.t. $P_{\tilde{M}} = P_M$, $\tilde{g} = g$ on $\mathrm{supp}(A) \cup \mathrm{supp}_{\mathcal{I}}(A)$.

intervention	$\mathrm{supp}_{\mathcal{I}}(X)$	assumptions	distr. gen. possible?
on X (well-behaved)	$\subseteq \mathrm{supp}(X)$	A1	✓
on X (well-behaved)	$\not\subseteq \mathrm{supp}(X)$	A1 and A2	✓
on X (well-behaved)	$\not\subseteq \mathrm{supp}(X)$	A1	✗
on A	$\subseteq \mathrm{supp}(X)$	A1 and A3	✓

Idea 4: Distribution Generalization

Recall: SCM assignment for X: $X := g(A) + h_2(H, \epsilon_X)$
Recall: SCM assignment for Y: $Y := f(X) + h_1(H, \epsilon_Y)$

A1: Identif. of f: $\forall \tilde{M} = (\tilde{f}, \ldots)$ s.t. $P_{\tilde{M}} = P_M$, we have $\tilde{f} = f$ on $\text{supp}(X)$.
A2: Extrapolation of f: If $f = \tilde{f}$ on $\text{supp}(X)$, then $f = \tilde{f}$ on $\text{supp}_{\mathcal{I}}(X)$.
A3: Same for g: $\forall \tilde{M} = (\tilde{f}, \tilde{g}, \ldots)$ s.t. $P_{\tilde{M}} = P_M$, $\tilde{g} = g$ on $\text{supp}(A) \cup \text{supp}_{\mathcal{I}}(A)$.

intervention	$\text{supp}_{\mathcal{I}}(X)$	assumptions	distr. gen. possible?
on X (well-behaved)	$\subseteq \text{supp}(X)$	A1	✓
on X (well-behaved)	$\not\subseteq \text{supp}(X)$	A1 and A2	✓
on X (well-behaved)	$\not\subseteq \text{supp}(X)$	A1	✗
on A	$\subseteq \text{supp}(X)$	A1 and A3	✓
on A	$\not\subseteq \text{supp}(X)$	A1, A2 and A3	✓

Idea 4: Distribution Generalization

Recall: SCM assignment for X: $X := g(A) + h_2(H, \epsilon_X)$
Recall: SCM assignment for Y: $Y := f(X) + h_1(H, \epsilon_Y)$

A1: Identif. of f: $\forall \tilde{M} = (\tilde{f}, \ldots)$ s.t. $P_{\tilde{M}} = P_M$, we have $\tilde{f} = f$ on $\mathrm{supp}(X)$.
A2: Extrapolation of f: If $f = \tilde{f}$ on $\mathrm{supp}(X)$, then $f = \tilde{f}$ on $\mathrm{supp}_{\mathcal{I}}(X)$.
A3: Same for g: $\forall \tilde{M} = (\tilde{f}, \tilde{g}, \ldots)$ s.t. $P_{\tilde{M}} = P_M$, $\tilde{g} = g$ on $\mathrm{supp}(A) \cup \mathrm{supp}_{\mathcal{I}}(A)$.

intervention	$\mathrm{supp}_{\mathcal{I}}(X)$	assumptions	distr. gen. possible?
on X (well-behaved)	$\subseteq \mathrm{supp}(X)$	A1	✓
on X (well-behaved)	$\not\subseteq \mathrm{supp}(X)$	A1 and A2	✓
on X (well-behaved)	$\not\subseteq \mathrm{supp}(X)$	A1	✗
on A	$\subseteq \mathrm{supp}(X)$	A1 and A3	✓
on A	$\not\subseteq \mathrm{supp}(X)$	A1, A2 and A3	✓
on A	$\not\subseteq \mathrm{supp}(X)$	A1 and A2	✗

R. Christiansen, N. Pfister, M. Jakobsen, N. Gnecco, JP: A causal framework for distribution generalization, arXiv:2006.07433
further methods: Theil 1958, Racine and Hayfield 2018

Idea 4: Distribution Generalization

Recall: SCM assignment for X: $X := g(A) + h_2(H, \epsilon_X)$
Recall: SCM assignment for Y: $Y := f(X) + h_1(H, \epsilon_Y)$

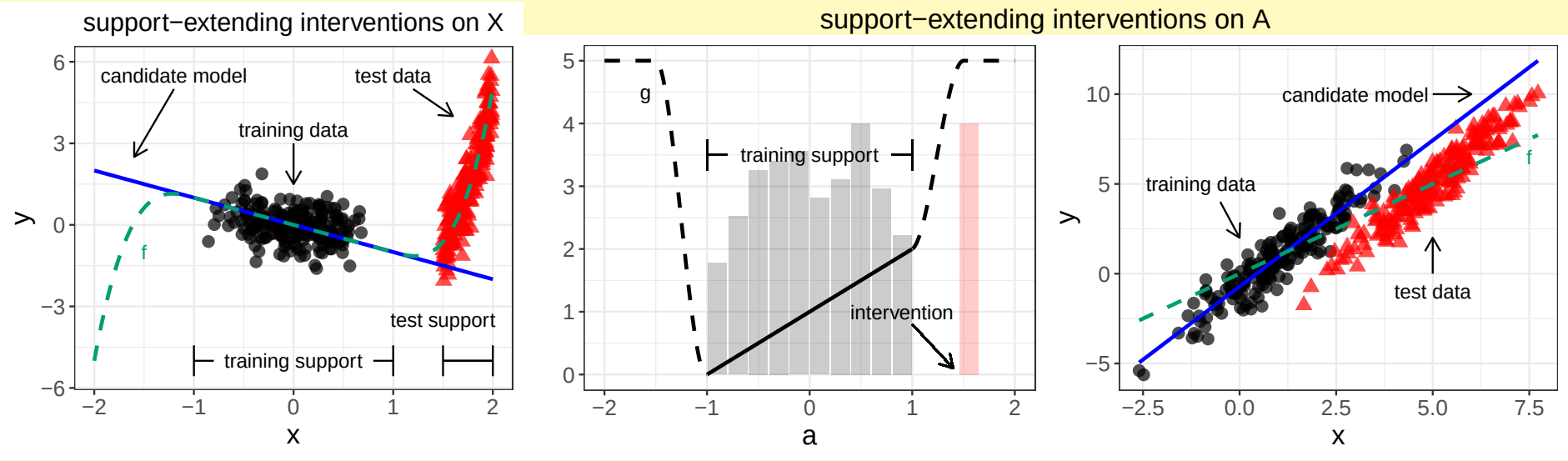
A1: Identif. of f: $\forall \tilde{M} = (\tilde{f}, \ldots)$ s.t. $P_{\tilde{M}} = P_M$, we have $\tilde{f} = f$ on $\mathrm{supp}(X)$.
A2: Extrapolation of f: If $f = \tilde{f}$ on $\mathrm{supp}(X)$, then $f = \tilde{f}$ on $\mathrm{supp}_{\mathcal{I}}(X)$.
A3: Same for g: $\forall \tilde{M} = (\tilde{f}, \tilde{g}, \ldots)$ s.t. $P_{\tilde{M}} = P_M$, $\tilde{g} = g$ on $\mathrm{supp}(A) \cup \mathrm{supp}_{\mathcal{I}}(A)$.

intervention	$\mathrm{supp}_{\mathcal{I}}(X)$	assumptions	distr. gen. possible?
on X (well-behaved)	$\subseteq \mathrm{supp}(X)$	A1	✓
on X (well-behaved)	$\not\subseteq \mathrm{supp}(X)$	A1 and A2	✓
on X (well-behaved)	$\not\subseteq \mathrm{supp}(X)$	A1	✗
on A	$\subseteq \mathrm{supp}(X)$	A1 and A3	✓
on A	$\not\subseteq \mathrm{supp}(X)$	A1, A2 and A3	✓
on A	$\not\subseteq \mathrm{supp}(X)$	A1 and A2	✗

R. Christiansen, N. Pfister, M. Jakobsen, N. Gnecco, JP: A causal framework for distribution generalization, arXiv:2006.07433
further methods: Theil 1958, Racine and Hayfield 2018

Idea 4: Distribution Generalization

Visualization of impossibility results:



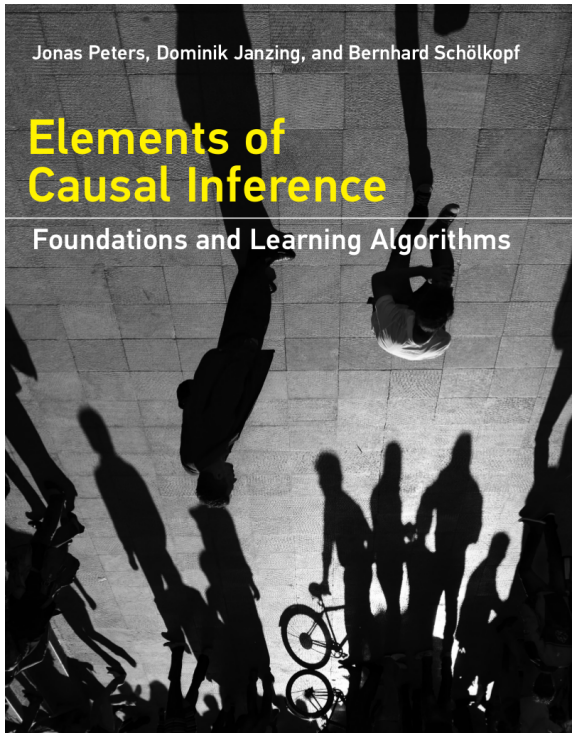
R. Christiansen, N. Pfister, M. Jakobsen, N. Gnecco, JP: A causal framework for distribution generalization, arXiv:2006.07433

Summary Part III:

- Idea 1: Reformulate reinforcement learning (use causal structure)
- Idea 2: Anchor regression (trade-off pred. and invariance)
- Idea 3: CausalKinetiX (dynamical systems)
- Idea 4: Distribution generalization (theoretical framework)

Summary Part III:

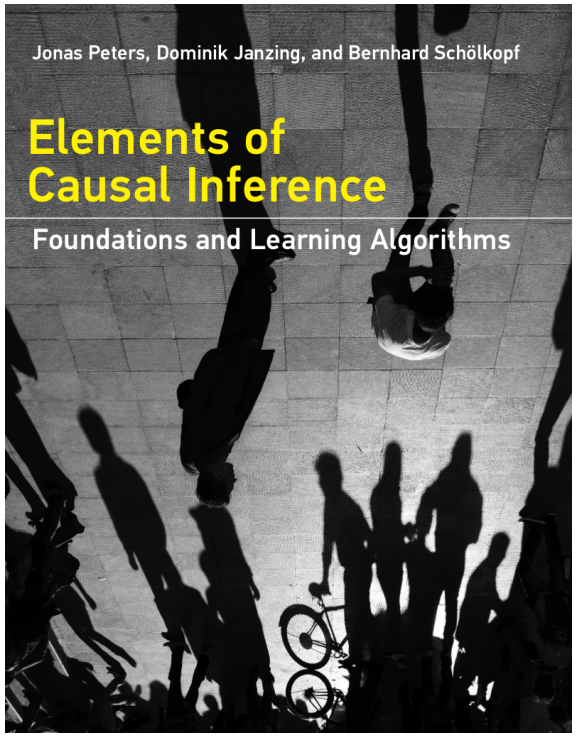
- Idea 1: Reformulate reinforcement learning (use causal structure)
- Idea 2: Anchor regression (trade-off pred. and invariance)
- Idea 3: CausalKinetiX (dynamical systems)
- Idea 4: Distribution generalization (theoretical framework)



For an exhaustive list of references, download pdf of
JP, D. Janzing, B. Schölkopf: *Elements of Causal Inference: Foundations and Learning Algorithms*, MIT Press 2017.

Summary Part III:

- Idea 1: Reformulate reinforcement learning (use causal structure)
- Idea 2: Anchor regression (trade-off pred. and invariance)
- Idea 3: CausalKinetiX (dynamical systems)
- Idea 4: Distribution generalization (theoretical framework)



For an exhaustive list of references, download pdf of
JP, D. Janzing, B. Schölkopf: *Elements of Causal Inference: Foundations and Learning Algorithms*, MIT Press 2017.

— Tusind tak!

How invariant is

$$\dot{Y}_t = \theta_1 X_t^8 \ ?$$

1. For each repetition i, smooth target trajectory. $\rightsquigarrow \hat{y}^{(i)}$
2. Obtain fitted values for target deriv. (other exp.). $\rightsquigarrow \xi_{t_1}^{(i)}, \ldots, \xi_{t_m}^{(i)}$
3. Smooth target trajectory, with constraints on derivatives. $\rightsquigarrow \hat{y}^{(i)}$

How invariant is

$$\dot{Y}_t = \theta_1 X_t^8 \ ?$$

1. For each repetition i, smooth target trajectory. $\rightsquigarrow \hat{y}^{(i)}$
2. Obtain fitted values for target deriv. (other exp.). $\rightsquigarrow \xi_{t_1}^{(i)}, \ldots, \xi_{t_m}^{(i)}$
3. Smooth target trajectory, with constraints on derivatives. $\rightsquigarrow \hat{y}^{(i)}$
4. Score for model ranking

$$\sum_{i=1}^{n} \left[\mathrm{RSS}^{(i)} - \mathrm{RSS}^{(i)} \right] / \left[\mathrm{RSS}^{(i)} \right],$$

where $\mathrm{RSS}^{(i)} := \sum_\ell \left(\hat{y}_{t_\ell}^{(i)} - Y_{t_\ell}^{(i)} \right)^2$.

How invariant is

$$\dot{Y}_t = \theta_1 X_t^8 \ ?$$

1. For each repetition i, smooth target trajectory. $\rightsquigarrow \hat{y}^{(i)}$
2. Obtain fitted values for target deriv. (other exp.). $\rightsquigarrow \xi_{t_1}^{(i)}, \ldots, \xi_{t_m}^{(i)}$
3. Smooth target trajectory, with constraints on derivatives. $\rightsquigarrow \hat{y}^{(i)}$
4. Score for model ranking

$$\sum_{i=1}^{n} \left[\mathrm{RSS}^{(i)} - \mathrm{RSS}^{(i)} \right] / \left[\mathrm{RSS}^{(i)} \right],$$

   where $\mathrm{RSS}^{(i)} := \sum_\ell \left(\hat{y}_{t_\ell}^{(i)} - Y_{t_\ell}^{(i)} \right)^2$.
5. Turn the score for models into a score for variables/complexes.

Top ranked variables:

rank	**held-out-experiment**				
	1	2	3	4	5
1	$\mathbf{X^{33}}$	$\mathbf{X^{33}}$	$\mathbf{X^{33}}$	$\mathbf{X^{33}}$	$\mathbf{X^{33}}$
2	$\mathbf{X^{56}}$	X^{38}	X^{73}	$\mathbf{X^{38}}$	$\mathbf{X^{56}}$
3	$\mathbf{X^{122}}$	X^{61}	$\mathbf{X^{122}}$	$\mathbf{X^{128}}$	$\mathbf{X^{122}}$
4	$\mathbf{X^{128}}$	$\mathbf{X^{128}}$	$\mathbf{X^{138}}$	$\mathbf{X^{168}}$	$\mathbf{X^{128}}$
5	$\mathbf{X^{138}}$	$\mathbf{X^{138}}$	$\mathbf{X^{168}}$	X^{246}	$\mathbf{X^{138}}$
6	$\mathbf{X^{168}}$	$\mathbf{X^{168}}$	X^{215}	X^{61}	$\mathbf{X^{168}}$