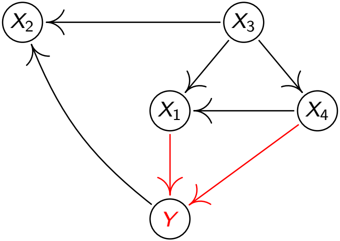
Causality



Jonas Peters
University of Copenhagen
LxMLS, Virtual, 21.7.–29.7.2020

Die Junge Akademie
CARLSBERGFONDET
VILLUM FONDEN
UNIVERSITY OF COPENHAGEN

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

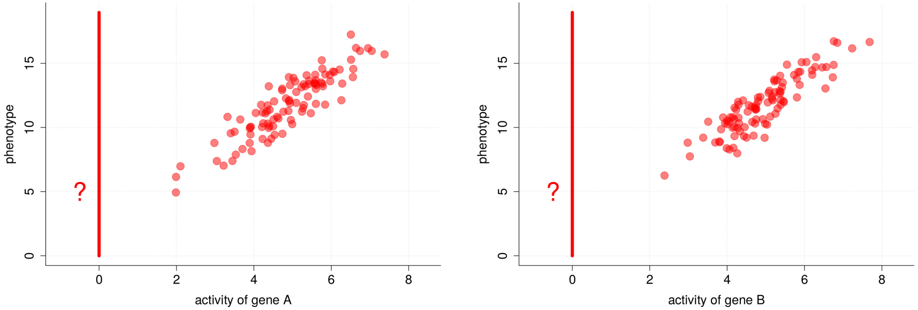
  Hernan & Robins: Causal Inference, Chapman & Hall/CRC 2019,

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

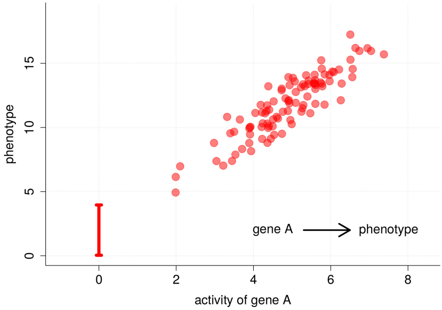
  Pearl: Causality, Cambridge Univ. Press 2009,

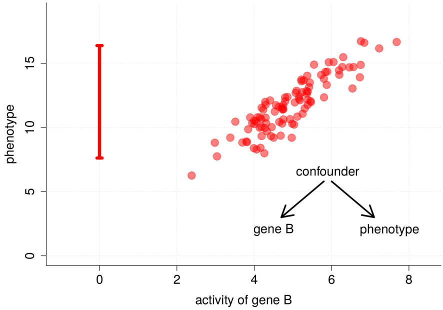
  ... and others.

Consider the following problem.



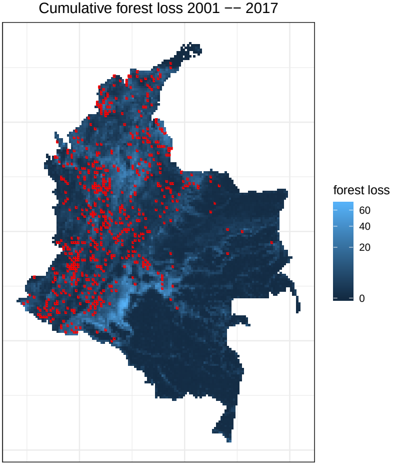
Causality matters!





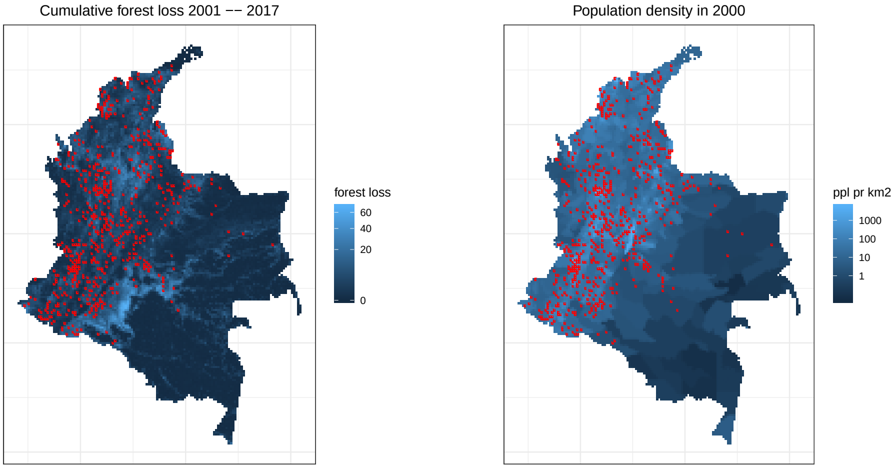
Example: Forest Loss

Colombia



Example: Forest Loss

Colombia



Christiansen et al. 2019

Example: kidney stones

	Treatment A	Treatment B
	$\frac{273}{350} = 0.78$	$\frac{289}{350} = 0.83$
	$\frac{562}{700} = 0.80$	

Charig et al.: *Comparison of treatment of renal calculi by open surgery, (...)* , British Medical Journal, 1986

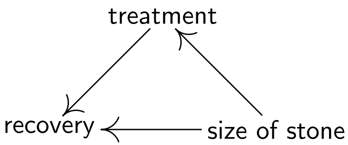
Example: kidney stones

	Treatment A	Treatment B
Small Stones ($\frac{357}{700} = 0.51$)	$\frac{81}{87} = 0.93$	$\frac{234}{270} = 0.87$
Large Stones ($\frac{343}{700} = 0.49$)	$\frac{192}{263} = 0.73$	$\frac{55}{80} = 0.69$
	$\frac{273}{350} = 0.78$	$\frac{289}{350} = 0.83$
	$\frac{562}{700} = 0.80$	

Charig et al.: *Comparison of treatment of renal calculi by open surgery, (...)* , British Medical Journal, 1986

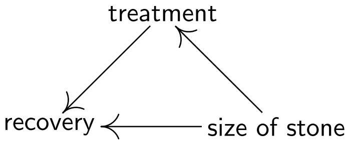
Example: kidney stones

underlying ground truth:



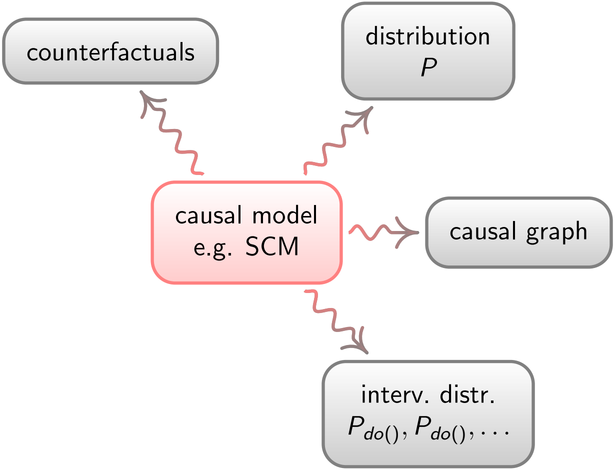
Example: kidney stones

underlying ground truth:



What is the expected recovery if all get treatment A?

What is a causal model?



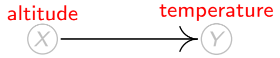
Part I: Causal Models

Example: Two variables

SCMs model observational distributions.

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

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

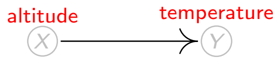


Example: Two variables

SCMs model observational distributions.

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

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



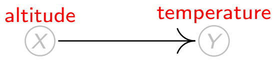
$$P: \qquad (X,Y) \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & -6 \\ -6 & 37 \end{pmatrix}\right)$$

Example: Two variables

SCMs model interventions, too.

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

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

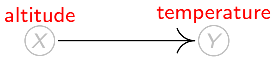


Example: Two variables

SCMs model interventions, too.

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

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



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

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.

$$X := N_X$$

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

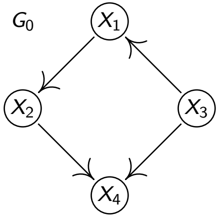
$$N_X, N_Y \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$$

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

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

$$X_1 := X_3 + N_1$$
$$X_2 := 2X_1 + N_2$$
$$X_3 := N_3$$
$$X_4 := -X_2 - X_3 + N_4$$

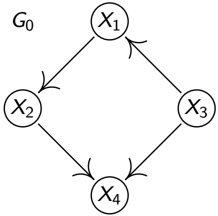
- N_i jointly independent $\mathcal{N}(0,1)$
- G_0 has no cycles



SCMs model observational distributions over $X_1, \ldots, X_d$. 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

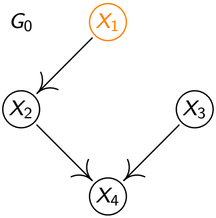


$$
\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix} = \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & 4 & 1 & -5 \\ 4 & 9 & 2 & -11 \\ 1 & 2 & 1 & -3 \\ -5 & -11 & -3 & 15 \end{pmatrix} \right)
$$

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

$$
\begin{aligned}
X_1 &:= 0 \\
X_2 &:= f_2(X_1, N_2) \\
X_3 &:= f_3(N_3) \\
X_4 &:= f_4(X_2, X_3, N_4)
\end{aligned}
$$

- N_i jointly independent
- G_0 has no cycles

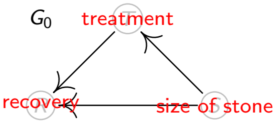


Example: kidney stones

Given: graph and P, i.e., only the structure, not the functions.

$$T := f_1(S, N_1)$$
$$R := f_2(T, S, N_2)$$
$$S := f_3(N_3)$$

- N_i jointly independent
- G_0 has no cycles

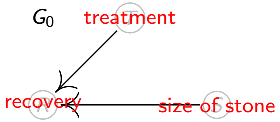


Example: kidney stones

Given: graph and P. We want to compute $P_{\mathrm{do}(T:=A)}$.

$$T := f_1(S, N_1) \quad T := A$$
$$R := f_2(T, S, N_2)$$
$$S := f_3(N_3)$$

- N_i jointly independent
- G_0 has no cycles



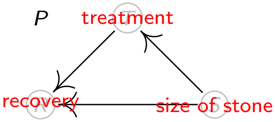
IMPORTANT: modularity, autonomy: Aldrich 1989, Pearl 2009, Schölkopf et al. 2012, ...

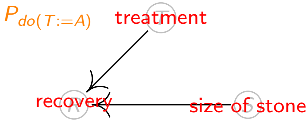
If you intervene only on X_j, you intervene only on X_j (MUTE).

Example: kidney stones

	Treatment A	Treatment B
Small Stones ($\frac{357}{700} = 0.51$)	$\frac{81}{87} = 0.93$	$\frac{234}{270} = 0.87$
Large Stones ($\frac{343}{700} = 0.49$)	$\frac{192}{263} = 0.73$	$\frac{55}{80} = 0.69$
	$\frac{273}{350} = 0.78$	$\frac{289}{350} = 0.83$
	$\frac{562}{700} = 0.80$	

Charig et al.: *Comparison of treatment of renal calculi by open surgery, (...)* , British Medical Journal, 1986





wanted: $P_{do(T:=A)}(R = 1)$

use: $P(R \mid S, T) = P_{do(T:=A)}(R \mid S, T)$

Example: kidney stones

$$
\begin{aligned}
E_{do(T:=A)}R &= P_{do(T:=A)}(R=1) \\
&= \sum_s P_{do(T:=A)}(R=1, S=s, T=A) \\
&= \sum_s P_{do(T:=A)}(R=1 \mid S=s, T=A) P_{do(T:=A)}(S=s, T=A) \\
&= \sum_s P_{do(T:=A)}(R=1 \mid S=s, T=A) P_{do(T:=A)}(S=s) \\
&= \sum_s P(R=1 \mid S=s, T=A) P(S=s) \\
&= 0.832 \\
&> 0.782 \\
&= \dots \\
&= P_{do(T:=B)}(R=1) = E_{do(T:=B)}R
\end{aligned}
$$

Example: kidney stones

$$
\begin{aligned}
E_{do(T:=A)}R &= P_{do(T:=A)}(R=1)\\
&= \sum_s P_{do(T:=A)}(R=1, S=s, T=A)\\
&= \sum_s P_{do(T:=A)}(R=1 \,|\, S=s, T=A) P_{do(T:=A)}(S=s, T=A)\\
&= \sum_s P_{do(T:=A)}(R=1 \,|\, S=s, T=A) P_{do(T:=A)}(S=s)\\
&= \sum_s P(R=1 \,|\, S=s, T=A) P(S=s)\\
&= 0.832 \qquad {\color{red}\neq P(R=1 \,|\, T=A)}\\
&> 0.782\\
&= \dots\\
&= P_{do(T:=B)}(R=1) = E_{do(T:=B)}R
\end{aligned}
$$

This idea holds more generally.

Definition

Given an SCM over (X, Y, W). We call $\mathrm{Z} \subseteq \mathrm{W}$ a valid adjustment set for (X, Y) if

$$p_{do(X:=x)}(y) = \sum_{\mathrm{z}} p(y|x, \mathrm{z})p(\mathrm{z}) \neq p(y|x)$$

This idea holds more generally.

Definition

Given an SCM over (X, Y, W). We call $\mathrm{Z} \subseteq \mathrm{W}$ a valid adjustment set for (X, Y) if

$$p_{do(X:=x)}(y) = \sum_{\mathrm{z}} p(y|x, \mathrm{z})p(\mathrm{z}) \neq p(y|x)$$

Proposition (Parent Adjustment)

Assume $Y \notin PA(X)$. Then

$$PA(X) \textit{ is a valid adjustment set for } (X, Y).$$

This idea holds more generally.

Definition

Given an SCM over (X, Y, W). We call $\mathrm{Z} \subseteq \mathrm{W}$ a valid adjustment set for (X, Y) if

$$p_{do(X:=x)}(y) = \sum_{\mathrm{z}} p(y|x, \mathrm{z}) p(\mathrm{z}) \neq p(y|x)$$

Proposition (Parent Adjustment)

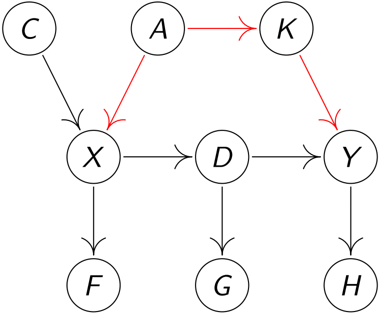
Assume $Y \notin PA(X)$. Then

$$PA(X) \textit{ is a valid adjustment set for } (X, Y).$$

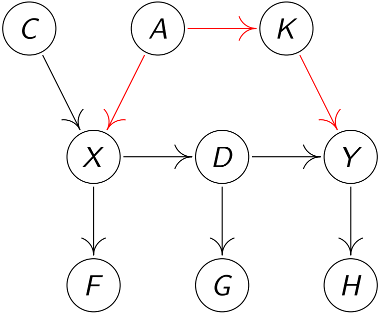
In particular, if $\emptyset$ is valid adjustment set, then

$$p_{do(X:=x)}(y) = p(y|x).$$

Adjusting in Linear Gaussian Models



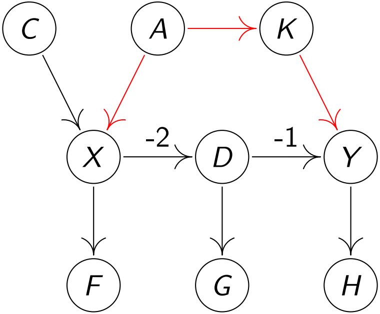
$X \leftarrow A \rightarrow K \rightarrow Y$ is a "backdoor path" from X to Y.



$X \leftarrow A \rightarrow K \rightarrow Y$ is a "backdoor path" from X to Y.

$$\mathrm{Z} = \{C, A\},$$
$$\mathrm{Z} = \{K\},$$
$$\mathrm{Z} = \{F, C, K\}$$

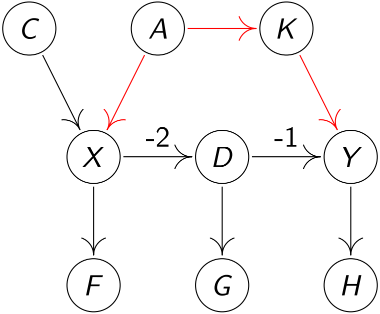
are valid adjustment sets for (X, Y) (no proof).



$X \leftarrow A \rightarrow K \rightarrow Y$ is a "backdoor path" from X to Y.

$$\mathrm{Z} = \{C, A\},$$
$$\mathrm{Z} = \{K\},$$
$$\mathrm{Z} = \{F, C, K\}$$

are valid adjustment sets for (X, Y) (no proof).



$X \leftarrow A \rightarrow K \rightarrow Y$ is a "backdoor path" from X to Y.

$$\begin{aligned} Z &= \{C, A\}, \\ Z &= \{K\}, \\ Z &= \{F, C, K\} \end{aligned}$$

are valid adjustment sets for (X, Y) (no proof).
Thus (no proof): `lm(Y ~ X + K)` yields consistent estimator for

$$\frac{\partial}{\partial x} E_{do(X:=x)} Y = (-2) \cdot (-1) = 2.$$

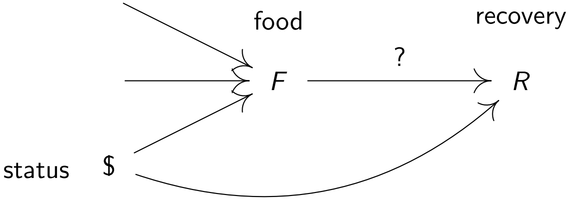
```
n <- 500

# generate a sample from the distr. ent. by the SCM
set.seed(1)
C <- rnorm(n)
A <- 0.8*rnorm(n)
K <- A + 0.1*rnorm(n)
X <- C - 2*A + 0.2*rnorm(n)
F <- 3*X + 0.8*rnorm(n)
D <- -2*X + 0.5*rnorm(n)
G <- D + 0.5*rnorm(n)
Y <- 2*K - D + 0.2*rnorm(n)
H <- 0.5*Y + 0.1*rnorm(n)

lm(Y~X)$coefficients
lm(Y~X+K)$coefficients
lm(Y~X+F+C+K)$coefficients
lm(Y~X+F+C+K+H)$coefficients
```

```r
n <- 500

# generate a sample from the distr. ent. by the SCM
set.seed(1)
C <- rnorm(n)
A <- 0.8*rnorm(n)
K <- A + 0.1*rnorm(n)
X <- C - 2*A + 0.2*rnorm(n)
F <- 3*X + 0.8*rnorm(n)
D <- -2*X + 0.5*rnorm(n)
G <- D + 0.5*rnorm(n)
Y <- 2*K - D + 0.2*rnorm(n)
H <- 0.5*Y + 0.1*rnorm(n)

lm(Y~X)$coefficients
lm(Y~X+K)$coefficients
lm(Y~X+F+C+K)$coefficients
lm(Y~X+F+C+K+H)$coefficients
```

Do not simply throw in as many variables as possible.

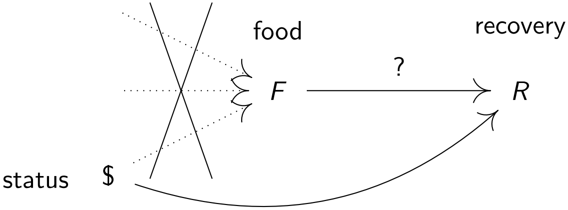
James Lind (1716–94):

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



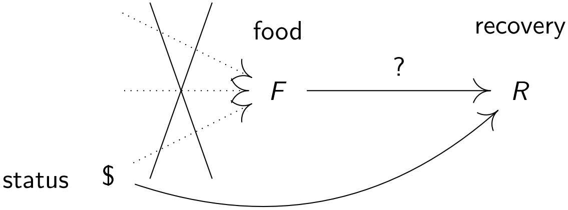
James Lind (1716–94):

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



James Lind (1716–94):

Randomize! F and R dependent $\implies$ 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;"

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

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.

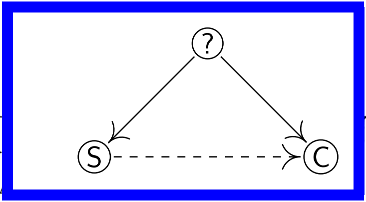
Possible Causes of the Increase

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

Example: smoking

BRITISH M RNAL

LONDON SA



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.

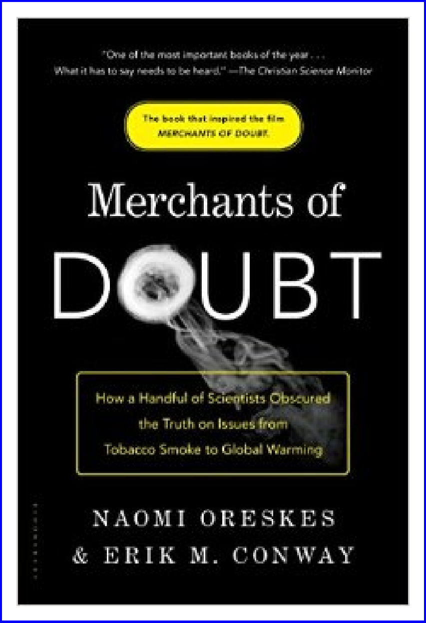
Professor of Medical Statistics, London School of Hygiene and Tropical Medicine ; Honorary Director of the Statistical Research Unit of the Medical Research Council

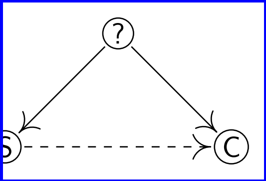
In England and Wales the phenomenal increase in the number of deaths attributed to cancer of the lung provides one of the most striking changes in the pattern of mortality recorded by the Registrar-General. For example, in the quarter of a century between 1922 and 1947 the annual number of deaths recorded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

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.

Jnit of the Medical Research Council

ND

HILL, Ph.D., D.Sc.

id Tropical Medicine; Honorary Director of the Statistical
ledical 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

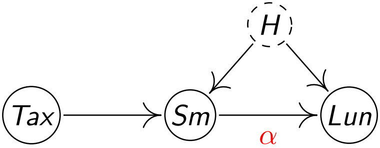
{**probabilistically** / **interventionally**} **equivalent**

if they entail the same

{observational / observational & interventional}

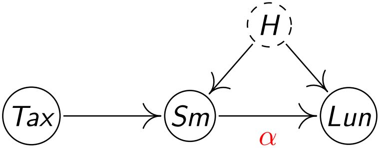
distributions. Here, it suffices to consider interventions that set a variable X_j to a fully supported $\tilde{N}_j$ ("randomized experiments").

Consider this graph



$$Lun = {\color{red}\alpha} Sm + \beta H + N$$

Consider this graph



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

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

$$
\begin{tikzcd}
 & & H \arrow[dl] \arrow[dr] & \\
Tax \arrow[r] & Sm \arrow[rr, "\alpha"'] & & Lun
\end{tikzcd}
$$

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

Summary Part I:

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

Summary Part I:

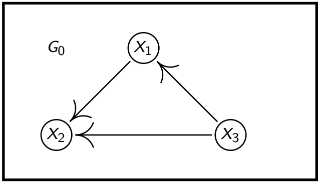
- What if interested in iid prediction, i.e., observational data? Don't worry (too much) about causality!
- But often, we are interested in a system's behaviour under intervention.

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.

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

- N_i jointly independent
- G_0 has no cycles

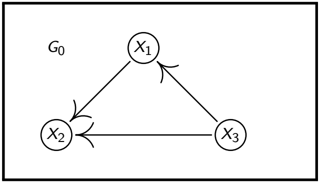


Summary Part I:

- 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.

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

- N_i jointly independent
- G_0 has no cycles



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

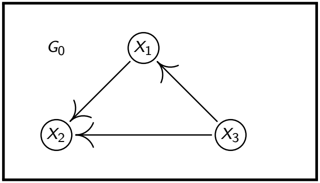
  ComputeInterventions.ipynb (skip Exercise 1 and maybe Exercise on d-sep.)

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.

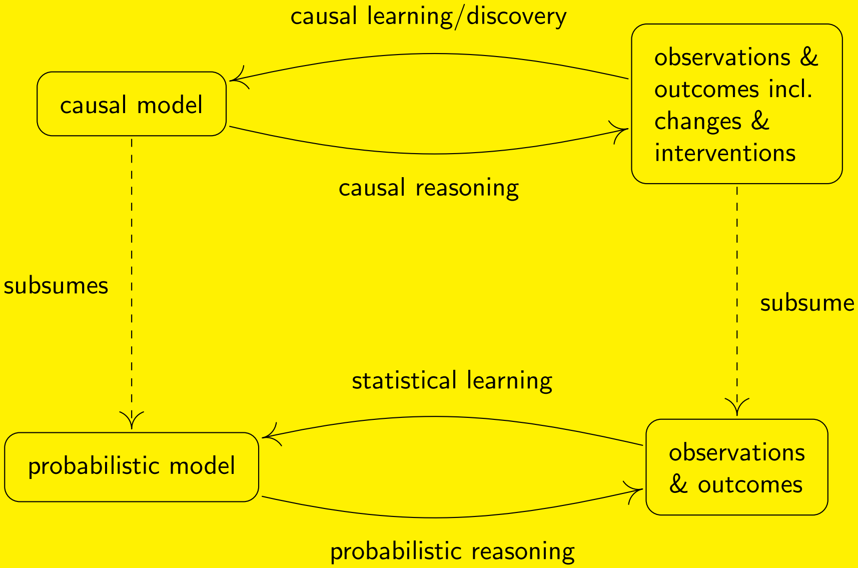
$$
\begin{aligned}
X_1 &:= f_1(X_3, N_1)\\
X_2 &:= f_2(X_1, X_3, N_2)\\
X_3 &:= f_3(N_3)
\end{aligned}
$$

- N_i jointly independent
- G_0 has no cycles



- adjusting: graph + observational distribution $\rightsquigarrow$ interventions
  ComputeInterventions.ipynb (skip Exercise 1 and maybe Exercise on d-sep.)
- instrumental variables: may help if there are hidden variables
  InstrumentalVariables.ipynb (skip Exercise 1)

Part II: Structure Learning or Causal Discovery


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

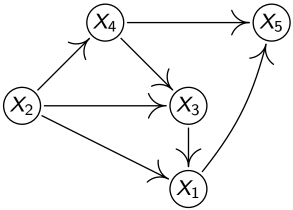
The Problem of Causal Discovery:

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

$\xrightarrow{?}$

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

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
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$



Correlation (Dependence) does not imply causation

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

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

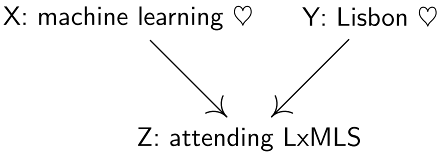
Reichenbach's common cause principle.
Assume that $X \not\perp\!\!\!\perp Y$. Then

- 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:

$$\begin{tikzcd} \text{X: machine learning } \heartsuit \arrow[dr] & & \text{Y: Lisbon } \heartsuit \arrow[dl] \\ & \text{Z: attending LxMLS} & \end{tikzcd}$$

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

Definition

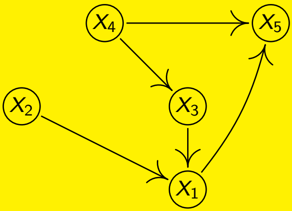
P is Markov w.r.t. G if

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

Definition: graphs

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

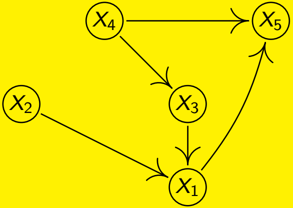
- parents, children, descendants, ancestors, ...
- paths, directed paths
- immoralities (or v-structures)
- d-separation (see next)
- ...



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!!

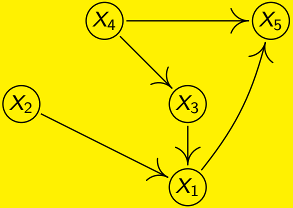


X_2 and X_5 are d-sep. by $\{X_1, X_4\}$
X_4 and X_1 are d-sep. by $\{X_2, X_3\}$
X_2 and X_4 are d-sep. by $\{\}$
X_4 and X_1 are NOT d-sep. by $\{X_3, X_5\}$

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!!



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

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

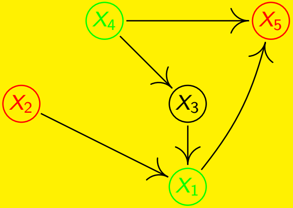
$\circ \cdots \to \circ \leftarrow \cdots \circ$ blocks a path.

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.

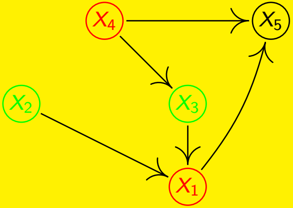
$\circ \cdots \rightarrow \circ \leftarrow \cdots \circ$ blocks a path.

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.

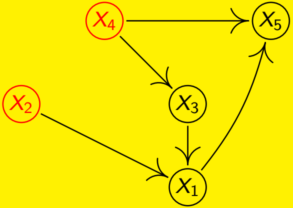
$\circ \cdots \rightarrow \circ \leftarrow \cdots \circ$ blocks a path.

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.

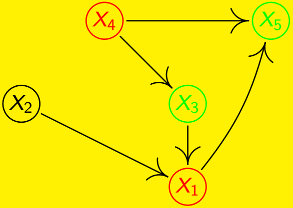
$\circ \cdots \rightarrow \circ \leftarrow \cdots \circ$ blocks a path.

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:

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

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

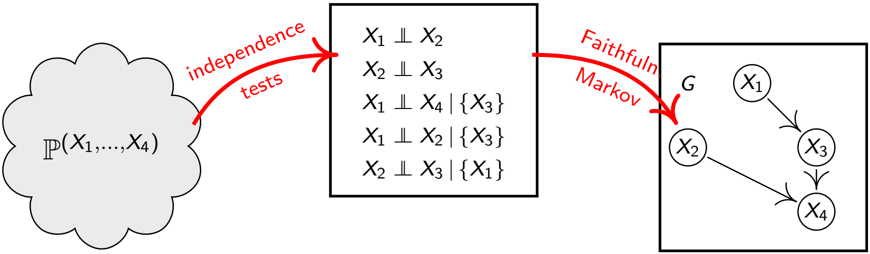
(plus symmetric statements). What is G?

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

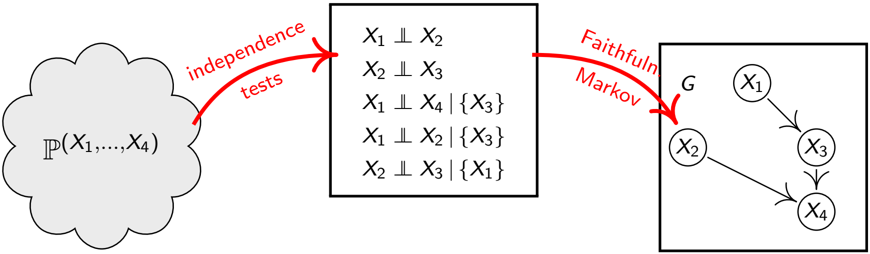
$$\begin{aligned}(Y,Z) &\perp\!\!\!\perp W \,|\, \emptyset\\ W &\perp\!\!\!\perp Y \,|\, (X,Z)\\ (W,X) &\perp\!\!\!\perp Y \,|\, Z\end{aligned}$$

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

Idea 1: independence-based methods



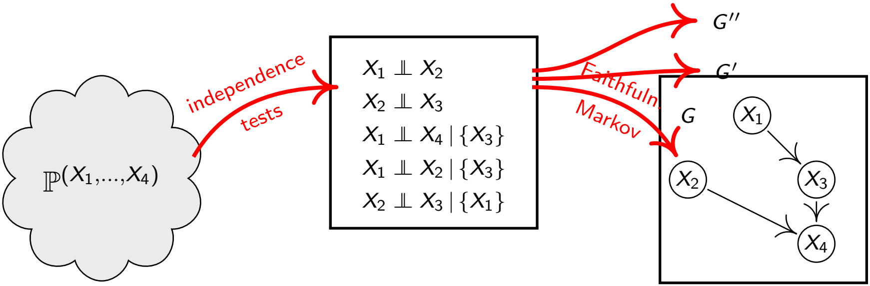
Idea 1: independence-based methods



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.

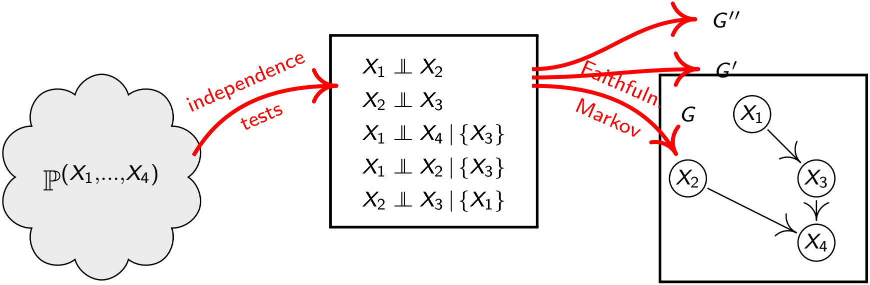
Idea 1: independence-based methods



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



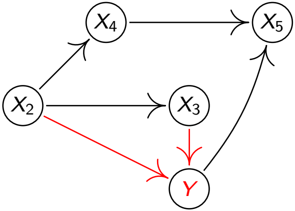
Method: ~~IC (Pearl 2009);~~ PC, FCI (Spirtes et al., 2000)

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

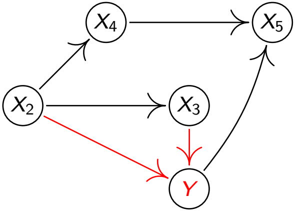
Y	X_2	X_3	X_4	X_5
3.4	−0.3	5.8	−2.1	2.2
1.7	−0.2	7.0	−1.2	0.4
−2.4	−0.1	4.3	−0.7	3.5
2.3	−0.3	5.5	−1.1	−4.4
3.5	−0.2	3.9	−0.9	−3.9
⋮	⋮	⋮	⋮	⋮



The Problem of Causal Discovery:

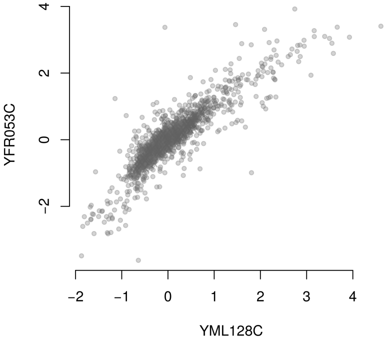
observed data $\xrightarrow{?}$ causal model

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
⋮	⋮	⋮	⋮	⋮



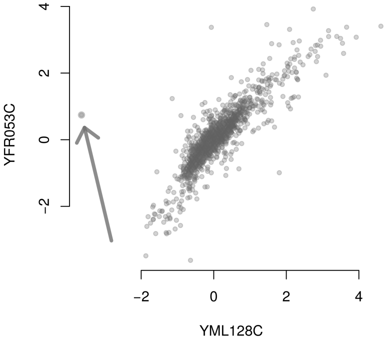
Here: Find the direct causes of Y!

Choose the predictor with the strongest correlation...

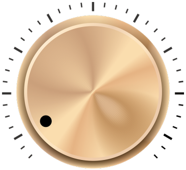


data from: Kemmeren et al. 2014

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

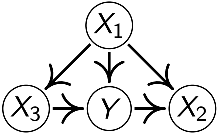


data from: Kemmeren et al. 2014

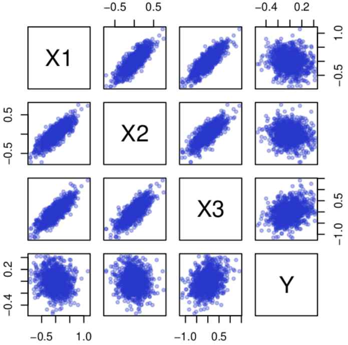
invariance  prediction invariance  prediction

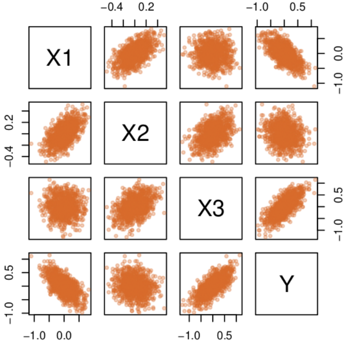
Invariant Causal Prediction

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 ***
```

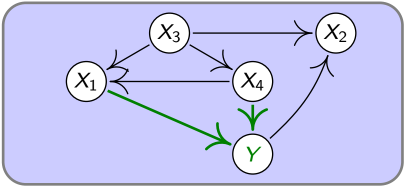
ICP (R-package InvariantCausalPrediction)

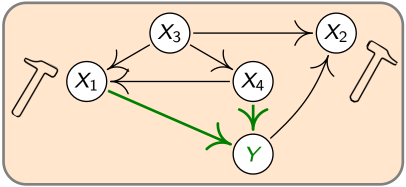
```
> ExpInd
```

[1]11111111111111111111111111111111...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
```

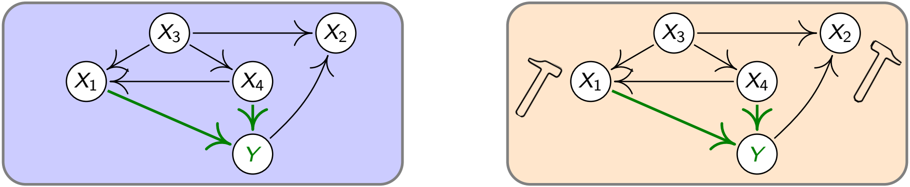
Fundamental assumption: $X_1, X_4 \to Y$ is invariant under interventions.





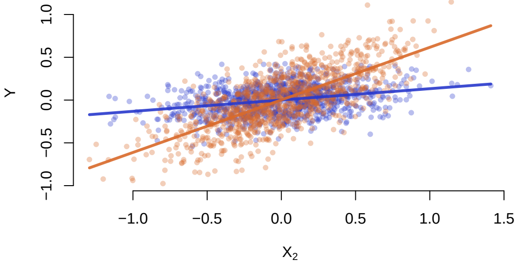
cf. modularity, autonomy, Haavelmo 1944, Aldrich 1989, Pearl 2009, ...

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

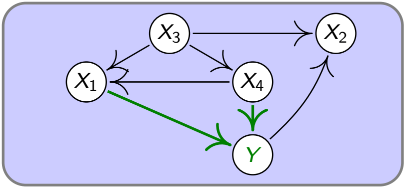


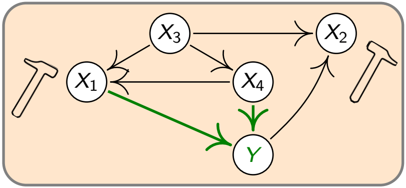
cf. modularity, autonomy, Haavelmo 1944, Aldrich 1989, Pearl 2009, . . .

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



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





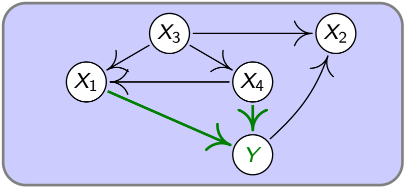
cf. modularity, autonomy, Haavelmo 1944, Aldrich 1989, Pearl 2009, ...

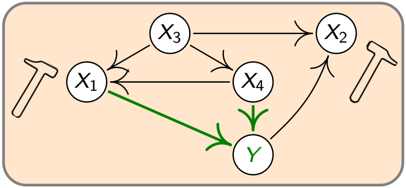
Key idea: Use ⬭ and ⬭ data and search for invariant models.

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: $X_1, X_4 \rightarrow Y$ 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)

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 $i = 1, \ldots, n$: $\qquad Y_i = X_{S,i} \cdot \gamma + \varepsilon_i$.
- $\varepsilon_1, \ldots, \varepsilon_n$ 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 $i = 1, \ldots, n$: $\qquad Y_i = X_{S,i} \cdot \gamma + \varepsilon_i$.
- $\varepsilon_1, \ldots, \varepsilon_n$ 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 $i = 1, \ldots, n$: $\qquad Y_i = X_{S,i} \cdot \gamma + \varepsilon_i$.
- $\varepsilon_1, \ldots, \varepsilon_n$ 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.

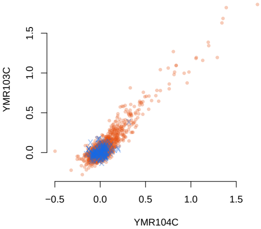
Relation to causality:
Environments: different interventions (not on Y). Then, $H_{0,PA(Y)}$ holds.
cf. modularity, autonomy, Haavelmo 1944, Aldrich 1989, Pearl 2009, Schölkopf et al. 2012, ...

Theorem (PBM 2016)

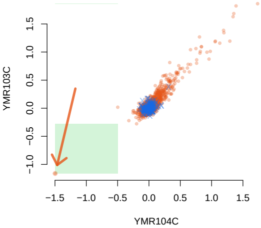
Assume H_{0,S^*} satisfied for some S^*. For any test level α we obtain

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

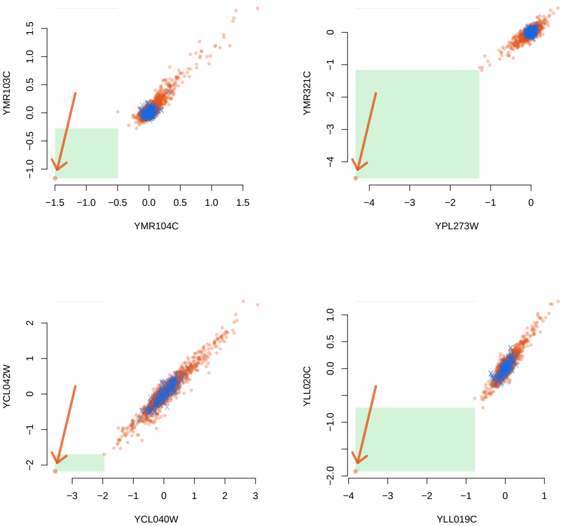
Predictors that are inferred to be causal by ICP...



...and the corresponding interventions on the predictor

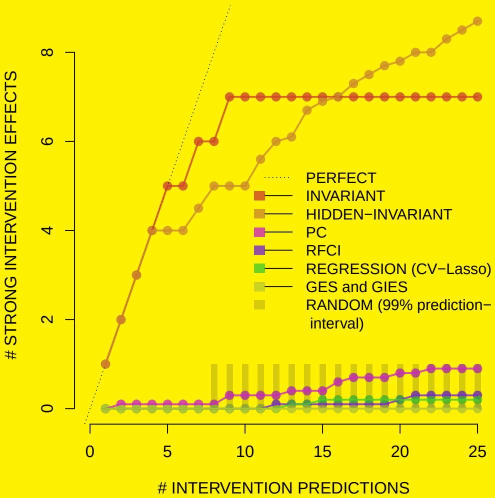


...and the corresponding interventions on the predictor



Kemmeren et al. 2014

Yeast data (Kemmeren et al., 2014)



Meinshausen et al., PNAS 2016

So far: invariance with respect to

anchor = environments

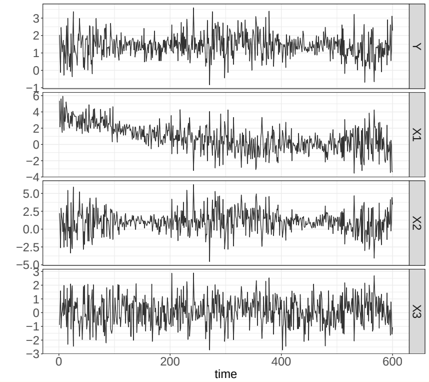
Also possible:

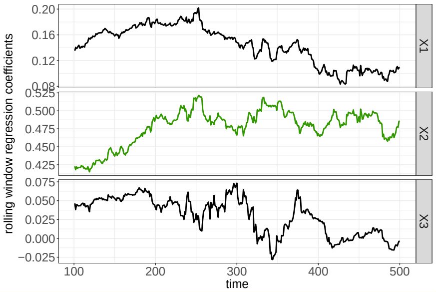
anchor = time

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
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

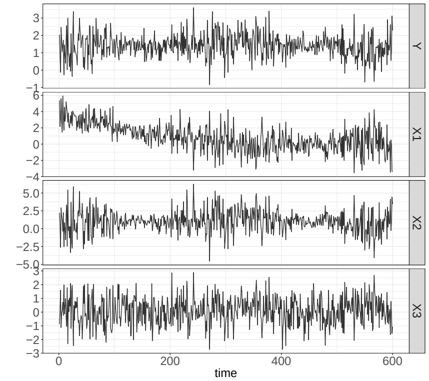
Regressing on (X_1, X_2, X_3):

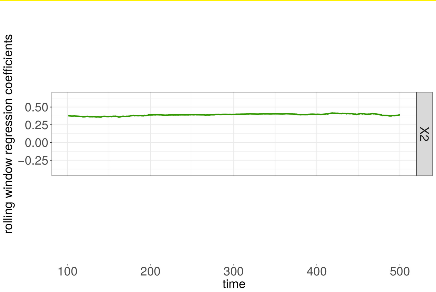




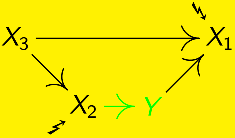
The coefficients change.

Regressing on X_1, X_2, and X_3:

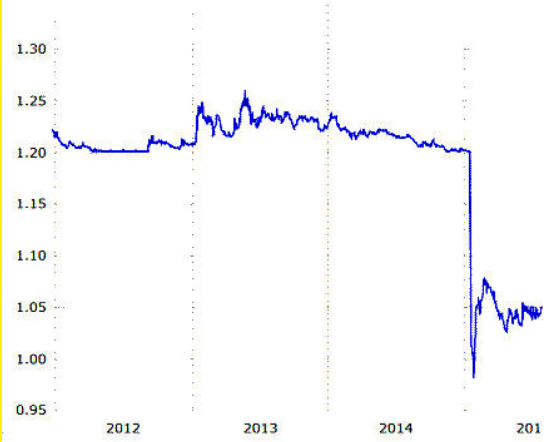




X_2 yields an invariant model. Ground truth:



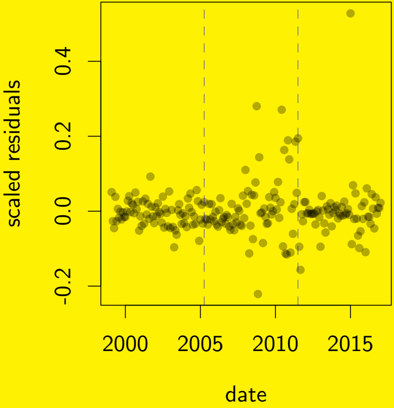
How much CHF do I need to pay for buying 1 EUR?



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



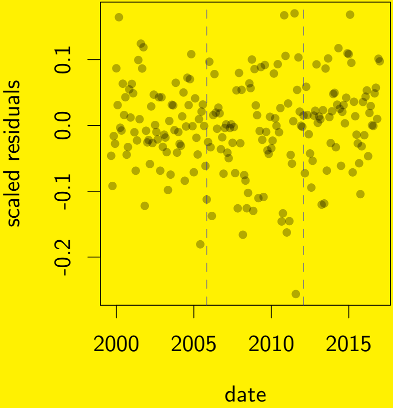


Figure: left plot (not invariant) and right plot (invariant)

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, 5

No

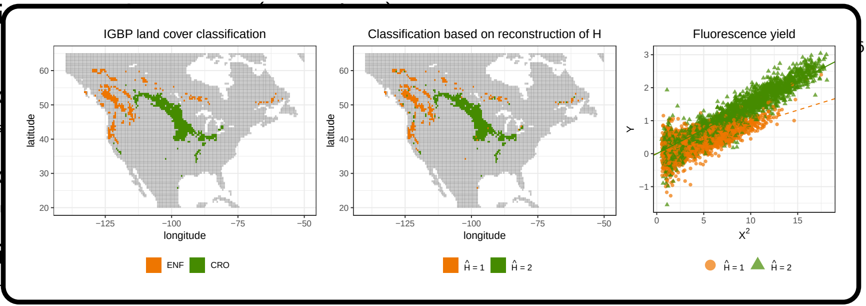
Pfis

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):

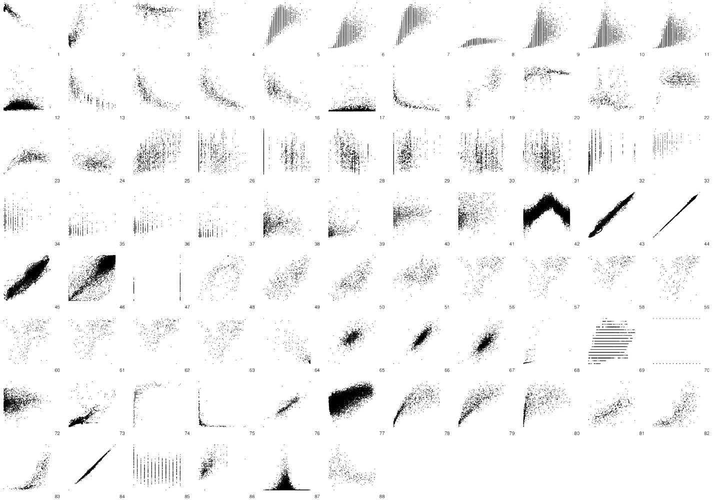
Christiansen, JP: *Invariance-based Causal Discovery in the Presence of Discrete Hidden Variables*, JMLR 2020

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

$$\boxed{\begin{array}{c} Y = f(X) + N_Y \\[2em] \text{with } N_Y, X \overset{ind}{\sim} \mathcal{N} \end{array}}$$

$$X \longrightarrow Y$$

Idea 3: restricted structural causal models

Consider a distribution entailed by

$$\boxed{\begin{array}{c} Y = f(X) + N_Y \\ \\ \text{with } N_Y, X \overset{ind}{\sim} \mathcal{N} \end{array}} \qquad X \longrightarrow Y$$

Then, if f is nonlinear, there is no

$$\boxed{\begin{array}{c} X = g(Y) + M_X \\ \\ \text{with } M_X, Y \overset{ind}{\sim} \mathcal{N} \end{array}} \qquad X \longleftarrow Y$$

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

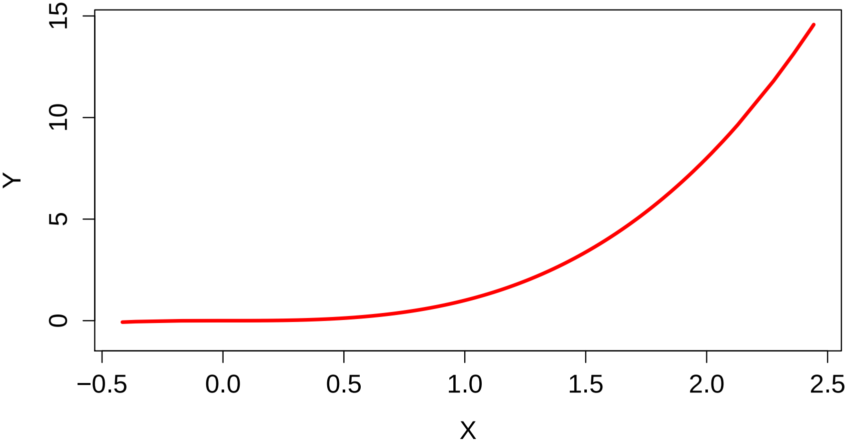
$$\boxed{\begin{array}{c} Y = X^3 + N_Y \\[1em] \text{with } N_Y, X \overset{ind}{\sim} \mathcal{N} \end{array}}$$

$$X \longrightarrow Y$$

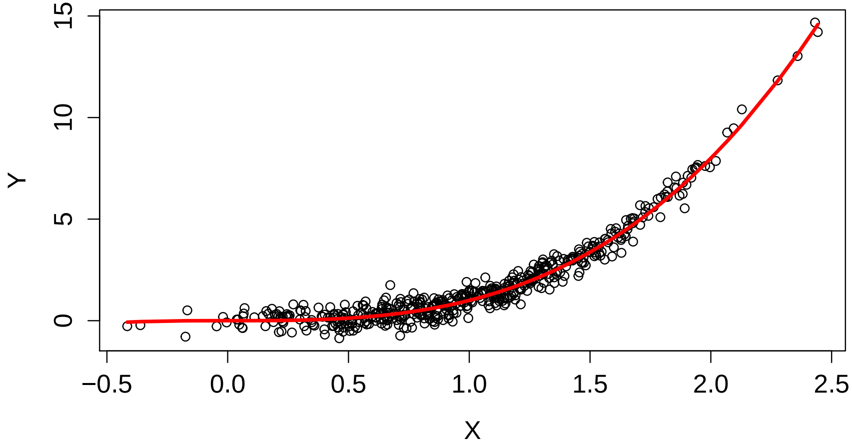
with

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

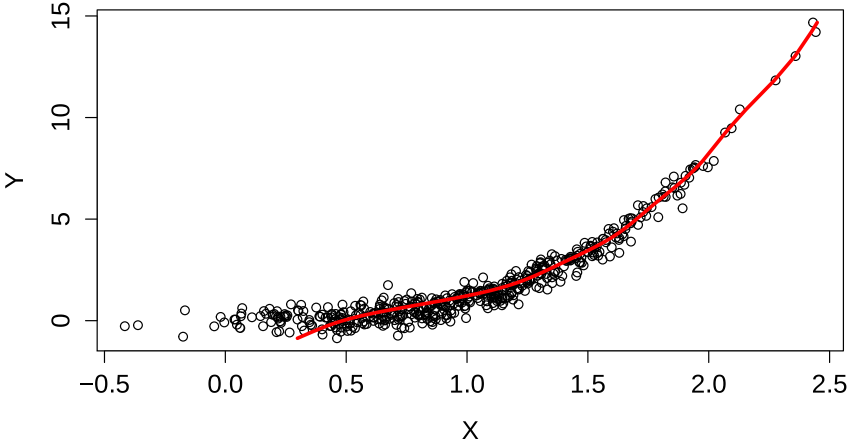
Idea 3: restricted structural causal models



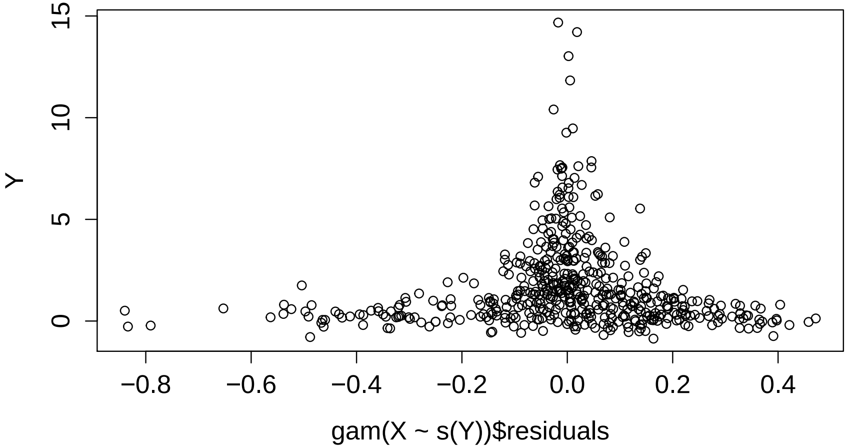
Idea 3: restricted structural causal models



Idea 3: restricted structural causal models



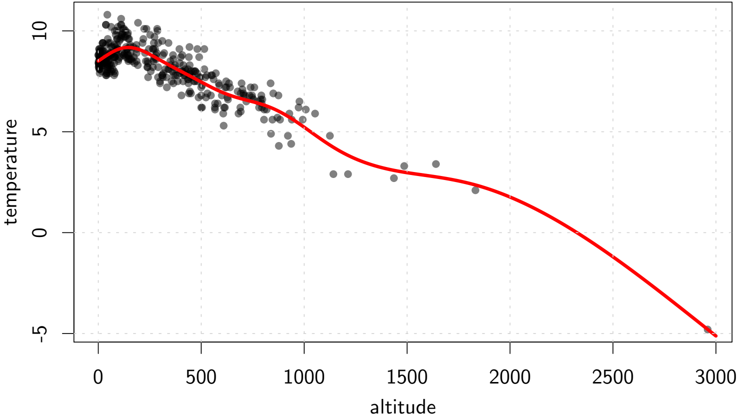
Idea 3: restricted structural causal models



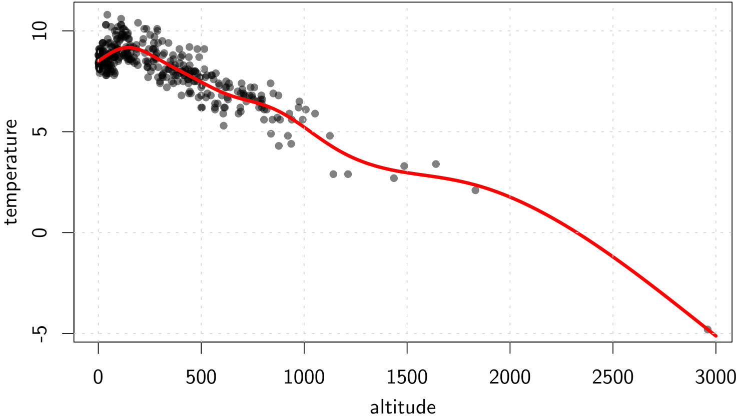
Idea 3: restricted structural causal models

Method... (jupyter notebook: RestrictedSCMs.ipynb)

Example: altitude and temperature



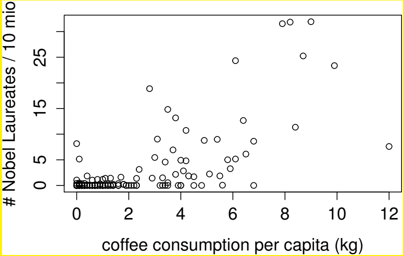
Example: altitude and temperature



p-value forward: 0.024

p-value backward: 0.0000000000019

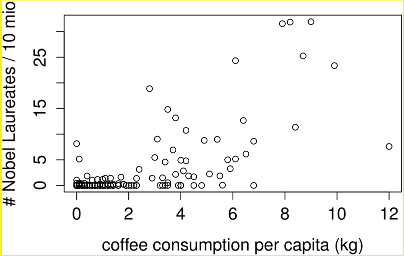
Example: coffee



Correlation: 0.698

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

Example: coffee

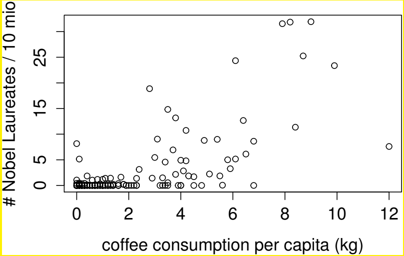


Correlation: 0.698
p-value: $< 2.2 \cdot 10^{-16}$

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

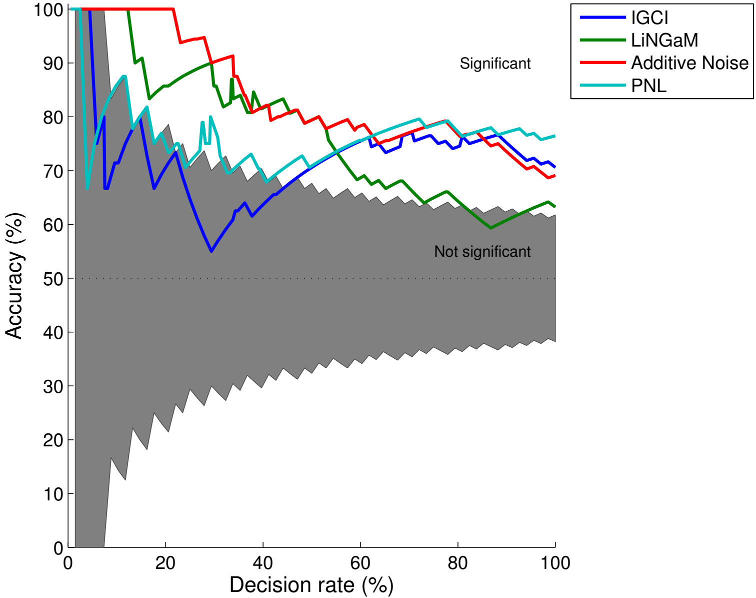


Correlation: 0.698
p-value: $< 2.2 \cdot 10^{-16}$

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:

identifiability for two variables $\rightsquigarrow$ identifiability for d variables

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

Idea 3: restricted structural causal models

Slightly surprising:

identifiability for two variables $\rightsquigarrow$ identifiability for d variables

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

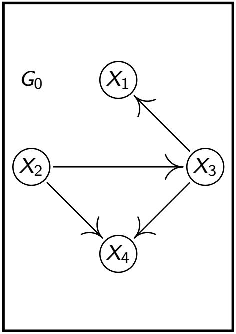
Most important counterexample: linear Gaussian.

Idea 3: restricted structural causal models

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

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

- N_i jointly independent
- G_0 has no cycles



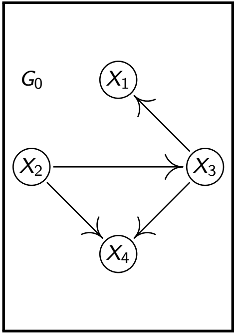
Structural equation model.
Can the DAG be recovered from $P(X_1, \ldots, X_4)$?

Idea 3: restricted structural causal models

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

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

- N_i jointly independent
- G_0 has no cycles



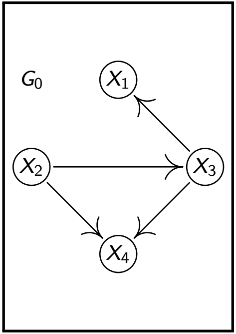
Structural equation model.
Can the DAG be recovered from $P(X_1, \dots, X_4)$? **No.**

Idea 3: restricted structural causal models

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

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

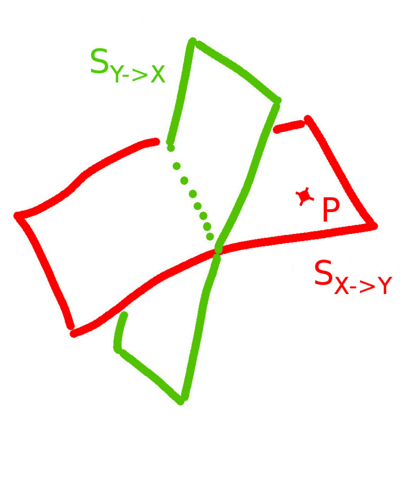
- $N_i \sim \mathcal{N}(0, \sigma_i^2)$ jointly independent
- G_0 has no cycles



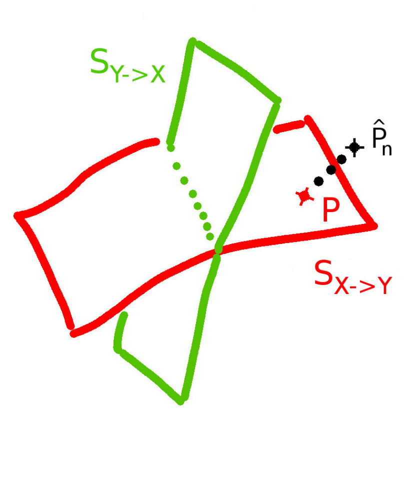
Additive noise model with Gaussian noise.
Can the DAG be recovered from $P(X_1, \dots, X_4)$? **Yes iff f_i nonlinear.**

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

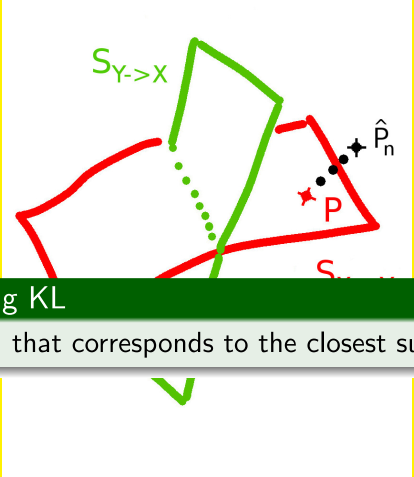
Idea 3: restricted structural causal models



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

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

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

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

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

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

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

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

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

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

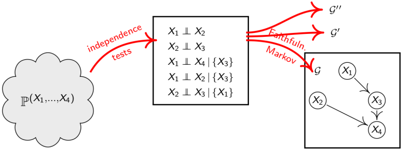
Idea 3: restricted structural causal models

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	99421195322159515895228914592354524516555026878588305014783
19	332771901227107591736177573311261125883583076258421902583546773505
20	2344880451051088988152559855229099188899081192234291298795803236068491263
21	34698768283588750028759328430181088222313944540438601719027559113446586077675521
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!)

InvariantCausalPrediction.ipynb

- Idea 3: additive noise (single environment)

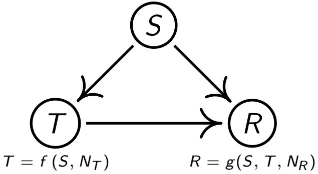
$$
\begin{aligned}
X_1 &= f_1(X_3) + N_1\\
X_2 &= N_2\\
X_3 &= f_3(X_2) + N_3\\
X_4 &= f_4(X_2, X_3) + N_4
\end{aligned}
$$

RestrictedSCMs.ipynb

Part III: Applications to Machine Learning (short)

Idea: RL

Recall the kidney stones:

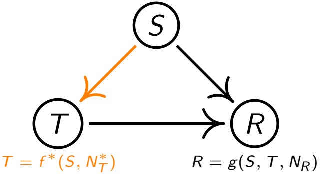


$$p(r, t, s) = p(r \mid t, s) \cdot \quad p(t \mid s) \quad \cdot p(s)$$

Question: What would happen if...?

Idea: RL

Recall the kidney stones:



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

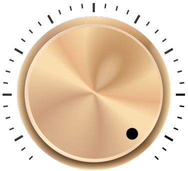
$$\downarrow$$

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

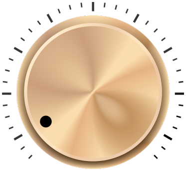
Question: What would happen if...?

What is $\sup_{p^*} \mathrm{E}_{p^*} R$?

Idea: anchor regression

invariance  prediction

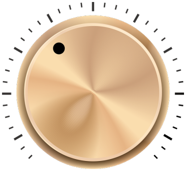
Idea: anchor regression

invariance  prediction

Idea: anchor regression

invariance  prediction

Idea: anchor regression

invariance  prediction

Find a trade-off between

- invariance with respect to 

AND

- predictive power

Idea: anchor regression

$Y \in \mathrm{R}^1$: target
$X \in \mathrm{R}^{1\times d}$: predictors
$A \in \mathrm{R}^{1\times q}$: anchors, $\mathrm{E}A^tA = Id$

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

Idea: anchor regression

$Y \in \mathrm{R}^1$: target
$X \in \mathrm{R}^{1 \times d}$: predictors
$A \in \mathrm{R}^{1 \times q}$: anchors, $\mathrm{E} A^t A = Id$

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

$\gamma \to 0$:	OLS
$\gamma \to \infty$:	IV solution (if identifiable)
$\gamma \to \infty$:	best invariant predictor (if not identifiable)

Idea: anchor regression

$Y \in \mathrm{R}^1$: target
$X \in \mathrm{R}^{1\times d}$: predictors
$A \in \mathrm{R}^{1\times q}$: anchors, $\mathrm{E}A^tA = Id$

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

$\gamma \to 0$: OLS
$\gamma \to \infty$: IV solution (if identifiable)
$\gamma \to \infty$: best invariant predictor (if not identifiable)

- Anchor regression minimizes worst case prediction error under shift interventions. Rothenhäusler, Bühlmann, Meinshausen, JP (arXiv:1801.06229)

Idea: anchor regression

$Y \in \mathrm{R}^1$: target
$X \in \mathrm{R}^{1\times d}$: predictors
$A \in \mathrm{R}^{1\times q}$: anchors, $\mathrm{E}A^tA = Id$

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

$\gamma \to 0$: OLS
$\gamma \to \infty$: IV solution (if identifiable)
$\gamma \to \infty$: best invariant predictor (if not identifiable)

- Anchor regression minimizes worst case prediction error under shift interventions. Rothenhäusler, Bühlmann, Meinshausen, JP (arXiv:1801.06229)
- The finite sample estimator is known as a k-class estimator for IV solution. Theil (1958), Nagar (1959), Jakobsen and JP (arXiv:2005.03353)

Idea: anchor regression

$$\begin{pmatrix} X \\ Y \\ H \end{pmatrix} \leftarrow \mathrm{B} \cdot \begin{pmatrix} X \\ Y \\ H \end{pmatrix} + \varepsilon + \mathrm{M}A,$$

$$\text{shifted:} \qquad \begin{pmatrix} X^v \\ Y^v \\ H^v \end{pmatrix} \leftarrow \mathrm{B} \cdot \begin{pmatrix} X^v \\ Y^v \\ H^v \end{pmatrix} + \varepsilon + v.$$

$Id - B$ invertible

Idea: anchor regression

$$\begin{pmatrix} X \\ Y \\ H \end{pmatrix} \leftarrow \mathrm{B} \cdot \begin{pmatrix} X \\ Y \\ H \end{pmatrix} + \varepsilon + \mathrm{M}A,$$

$$\text{shifted:} \qquad \begin{pmatrix} X^v \\ Y^v \\ H^v \end{pmatrix} \leftarrow \mathrm{B} \cdot \begin{pmatrix} X^v \\ Y^v \\ H^v \end{pmatrix} + \varepsilon + v.$$

$Id - B$ invertible

Theorem

For any $b \in \mathbb{R}^d$ we have

$$\underset{b}{\operatorname{argmin}} \ \mathrm{E}(Y - Xb)^2 + \gamma \|\mathrm{E}A^t(Y - Xb)\|_2^2 = \max_{v \in C^\gamma} \mathbb{E}[(Y^v - X^v b)^2],$$

where

$$C^\gamma := \{v = \mathrm{M}\delta \ \textit{such that} \ \|\delta\|_2 \leq \sqrt{\gamma}\}.$$

Idea: anchor regression

http://www.srfcdn.ch/radio/modules/dynimages/624/srf-1/2015/01/diverses/264377.150114_raclette_key.jpg

Idea: anchor regression

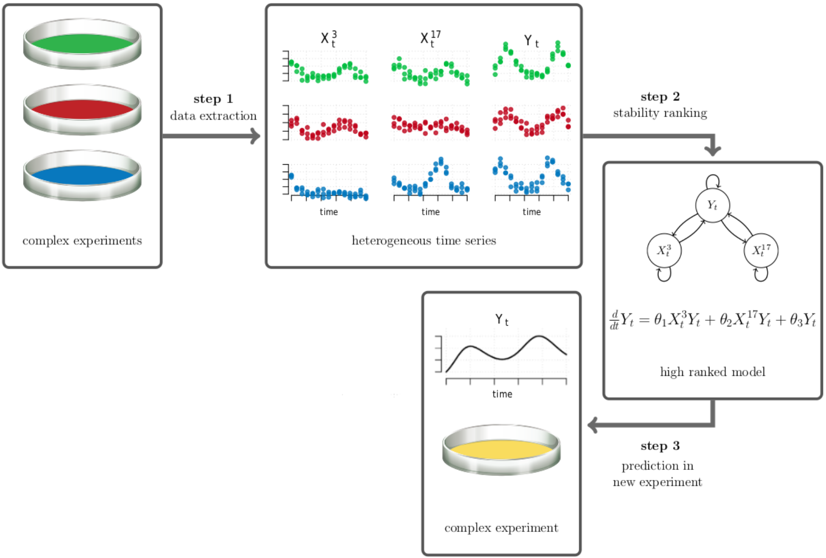
Example: Maillard reaction   Glu, Mel, C5, ForAc, Triose, Cn, AcAc, Amad, lysR, Fru, AMP

Idea: anchor regression

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

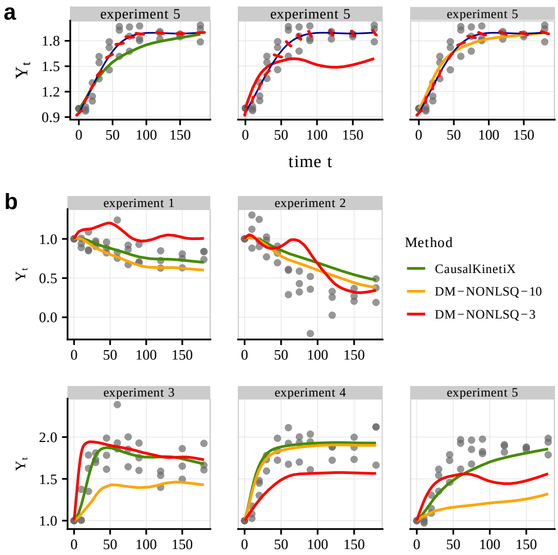
$$\frac{d}{dt}[\mathrm{Glu}]_t = -\theta_1[\mathrm{Glu}]_t + \theta_2[\mathrm{Fru}]_t - \theta_3[\mathrm{lysR}]_t[\mathrm{Glu}]_t$$
$$\frac{d}{dt}[\mathrm{Mel}]_t = \theta_4[\mathrm{AMP}]_t \qquad \ldots$$

Idea: anchor regression



N. Pfister, S. Bauer, JP: *Identifying Causal Structure in Large-Scale Kinetic Systems*, PNAS 2019

Idea: anchor regression



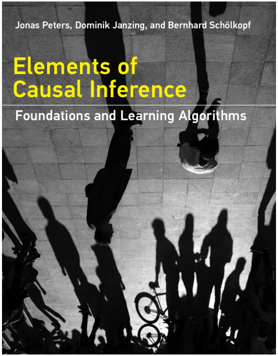
N. Pfister, S. Bauer, JP: *Identifying Causal Structure in Large-Scale Kinetic Systems*, PNAS 2019

Summary Part III:

- Idea 1: reformulate reinforcement learning, use causal structure
- Idea 2: semi-supervised learning from cause to effect does not work
- Idea 3: anchor regression

Summary Part III:

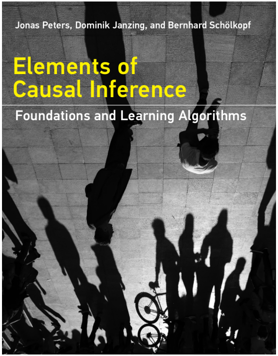
- Idea 1: reformulate reinforcement learning, use causal structure
- Idea 2: semi-supervised learning from cause to effect does not work
- Idea 3: anchor regression



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: semi-supervised learning from cause to effect does not work
- Idea 3: anchor regression



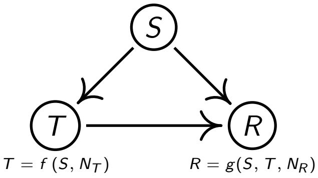
For an exhaustive list of references, download pdf of
JP, D. Janzing, B. Schölkopf: *Elements of Causal Inference: Foundations and Learning Algorithms*, MIT Press 2017.

— Tusind tak!

Part III: Applications to Machine Learning (long)

Idea 1: Blackjack

Recall the kidney stones:

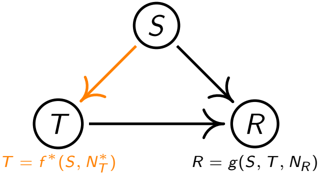


$$p(r, t, s) = p(r \mid t, s) \cdot \quad p(t \mid s) \quad \cdot p(s)$$

Question: What would happen if...?

Idea 1: Blackjack

Recall the kidney stones:



$$
\begin{aligned}
p(r, t, s) &= p(r \mid t, s) \cdot \quad p(t \mid s) \quad \cdot p(s) \\
&\qquad\qquad\qquad\quad \downarrow \\
p_3^*(r, t, s) &= p(r \mid t, s) \cdot \quad \underbrace{p^*(t \mid s)}_{p^*(t \mid s) = ?} \quad \cdot p(s)
\end{aligned}
$$

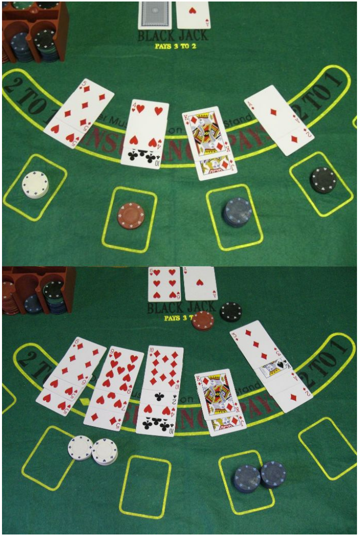
Question: What would happen if...?
What is $\sup_{p^*} \mathrm{E}_{p^*} R$?

Idea 1: Blackjack

(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: Blackjack



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

Idea 1: Blackjack
When can we learn?

Objects of Interest:

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

Idea 1: Blackjack

When can we learn?

Objects of Interest:

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

Questions:

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

Idea 1: Blackjack

When can we learn?

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 $\mathrm{E}_{p^*}\ell$?

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: Blackjack

Computation: Means

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

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

Idea 1: Blackjack

Computation: Means

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

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

Idea 1: Blackjack

Computation: Means

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

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

Estimate η by

$$\hat{\eta} = \frac{1}{N}\sum_{i=1}^{N} \ell(X_i, Y_i, Z_i)\; \underbrace{\frac{p^*(Y_i \mid X_i)}{p(Y_i \mid X_i)}}_{w_i} = \frac{1}{N}\sum_{i=1}^{N} M_i, \qquad \mathrm{E}_p\hat{\eta} = \eta$$

Idea 1: Blackjack

Computation: Means

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

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

Estimate η by

$$\hat{\eta} = \frac{1}{N}\sum_{i=1}^{N} \ell(X_i, Y_i, Z_i)\; \underbrace{\frac{p^*(Y_i \,|\, X_i)}{p(Y_i \,|\, X_i)}}_{w_i} = \frac{1}{N}\sum_{i=1}^{N} M_i, \qquad \mathrm{E}_p\hat{\eta} = \eta$$

Confidence intervals available!

Idea 1: Blackjack

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

Which p^* is best?

Idea 1: Blackjack

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

Which p^* is best? Parameterize and estimate

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

Idea 1: Blackjack

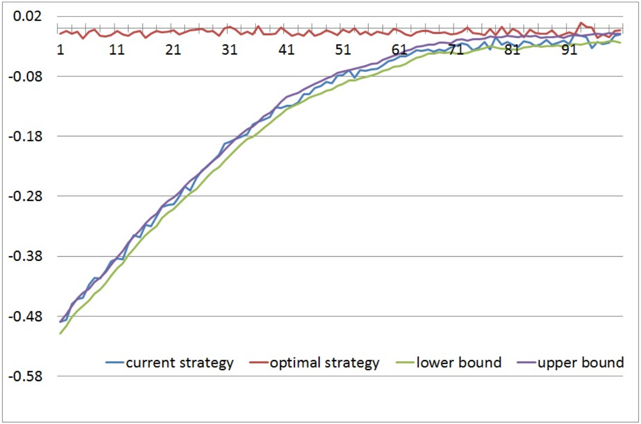
$$p(y \mid x) \to p^*(y \mid x)$$

Which p^* is best? Parameterize and estimate

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

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

Idea 1: Blackjack

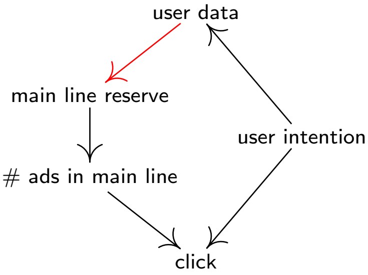


Idea 1: Blackjack

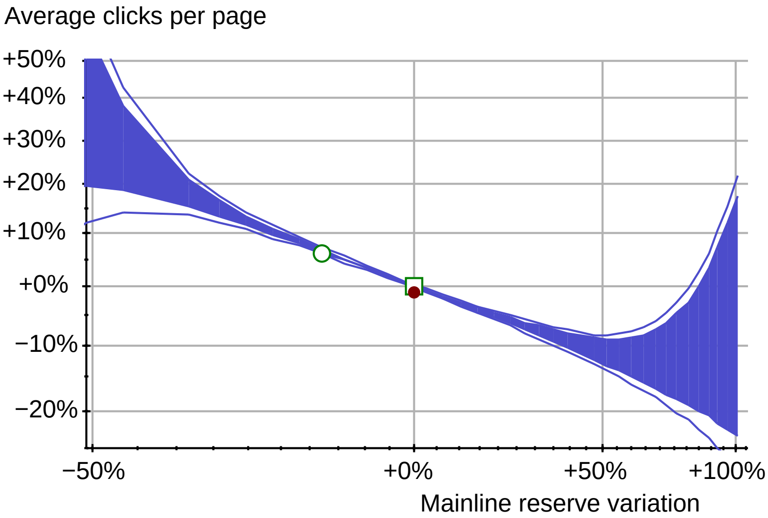
What can we do with $100,000$ samples?

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

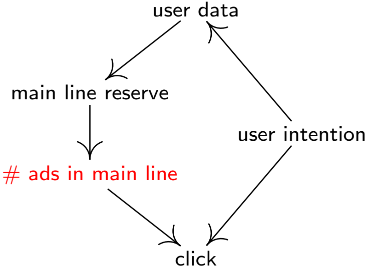
Idea 1: advertisement



Idea 1: advertisement

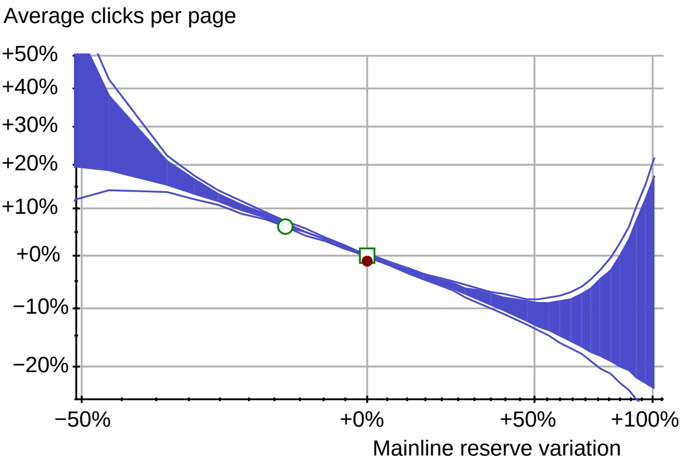


Idea 1: advertisement



Idea 1: advertisement

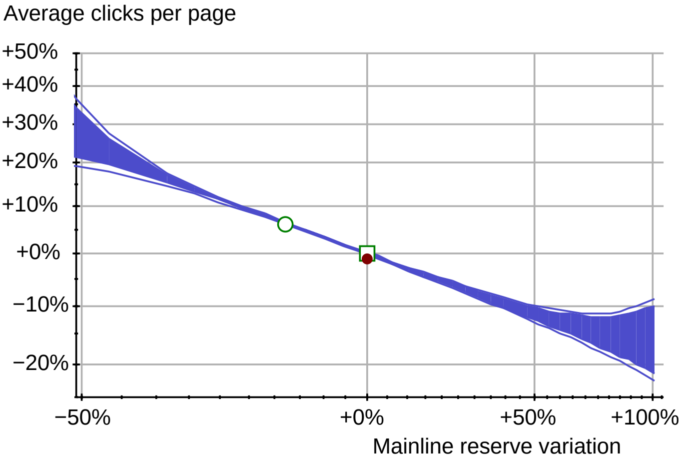
Old:



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

Idea 1: advertisement

Using discrete variable (ads shown in mainline):



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

Idea 2: semi-supervised learning

Consider a Markov factorization w.r.t. causal DAG:

$$p(x_1, \ldots, x_d) = \prod_{i=1}^{d} p(x_i \,|\, x_{pa(i)})$$

Idea 2: semi-supervised learning

Consider a Markov factorization w.r.t. causal DAG:

$$p(x_1, \ldots, x_d) = \prod_{i=1}^{d} p(x_i \mid x_{pa(i)})$$

Modularity suggests:

$$p(x_1 \mid x_{pa(1)}), \ldots, p(x_d \mid x_{pa(d)}) \text{ are “independent”}$$

Idea 2: semi-supervised learning

Consider a Markov factorization w.r.t. causal DAG:

$$p(x_1, \dots, x_d) = \prod_{i=1}^{d} p(x_i \mid x_{pa(i)})$$

Modularity suggests:

$$p(x_1 \mid x_{pa(1)}), \dots, p(x_d \mid x_{pa(d)}) \text{ are "independent"}$$

Special case:

$$p(cause), p(effect \mid cause) \text{ are "independent"}$$

Idea 2: semi-supervised learning

Consider a Markov factorization w.r.t. causal DAG:

$$p(x_1, \ldots, x_d) = \prod_{i=1}^{d} p(x_i \,|\, x_{pa(i)})$$

Modularity suggests:

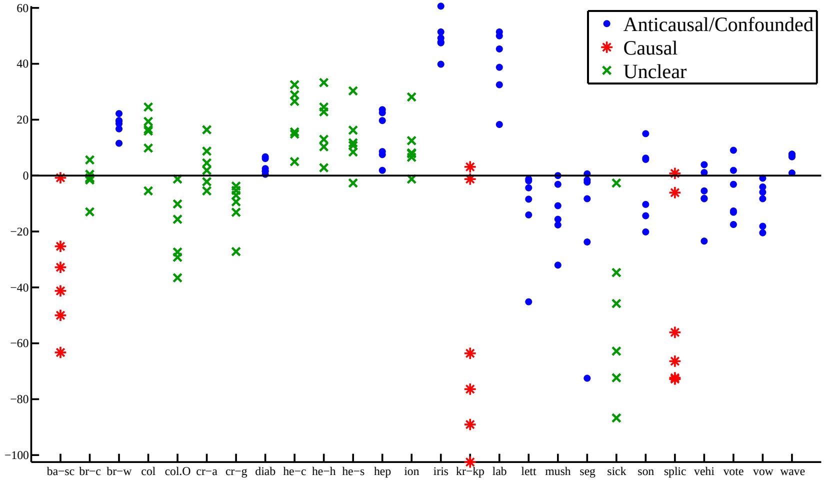
$$p(x_1 \,|\, x_{pa(1)}), \ldots, p(x_d \,|\, x_{pa(d)}) \text{ are "independent"}$$

Special case:

$$p(cause), p(effect \,|\, cause) \text{ are "independent"}$$

But then: Semi-supervised Learning does not work from cause to effect.

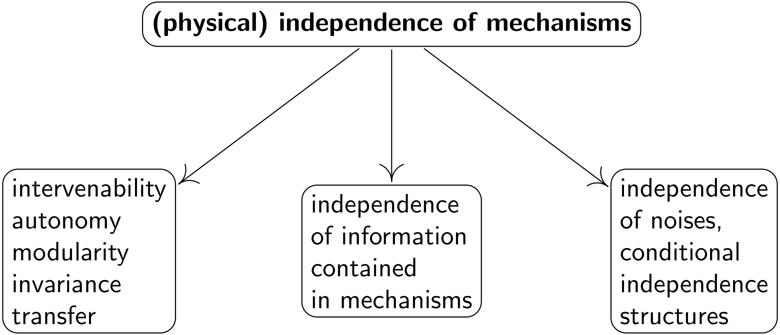
Idea 2: semi-supervised learning



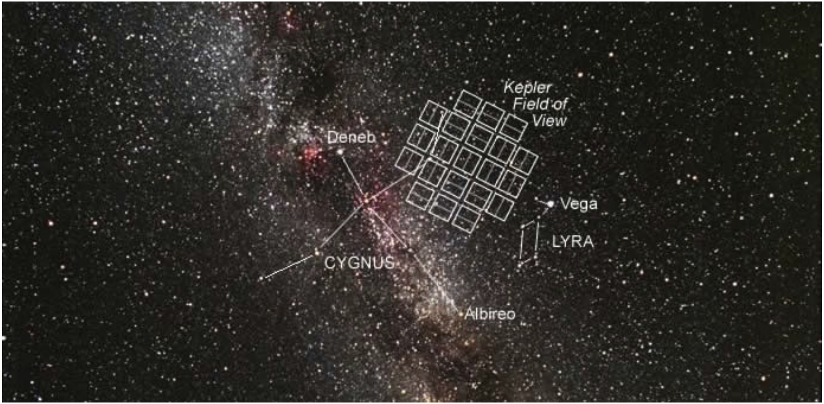
comparison to base classifier

Schölkopf et al.: *On causal and anticausal learning*, ICML 2012

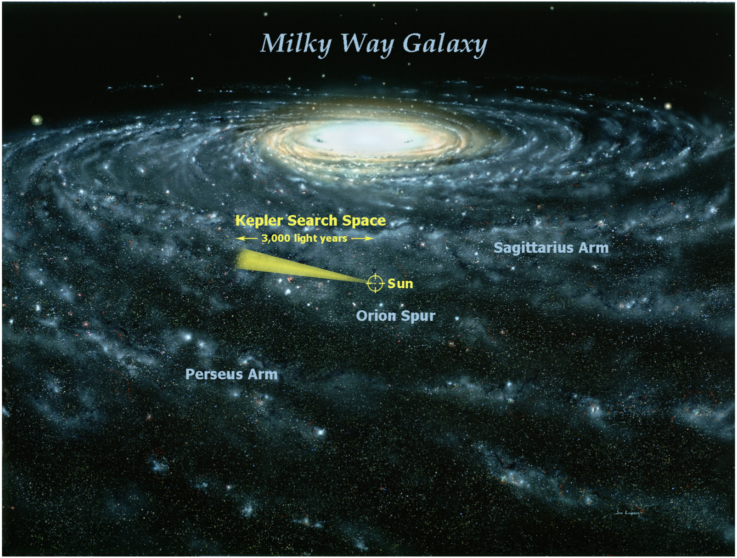
Idea 2: semi-supervised learning



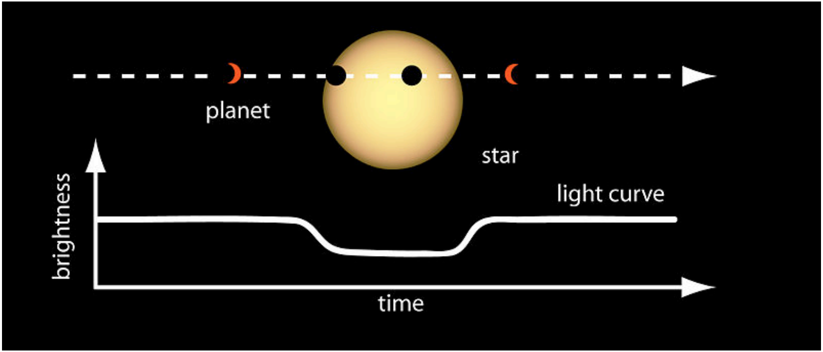
Idea 3: half-sibling regression



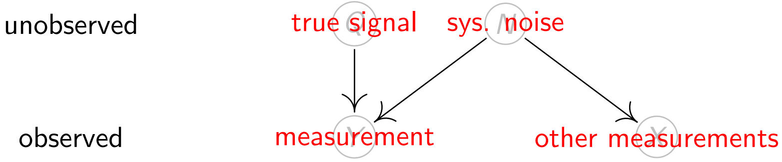
Idea 3: half-sibling regression



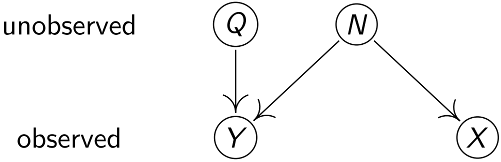
Idea 3: half-sibling regression



Idea 3: half-sibling regression

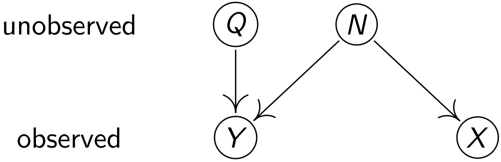


Idea 3: half-sibling regression



Assume $Y = f(N) + Q$.

Idea 3: half-sibling regression

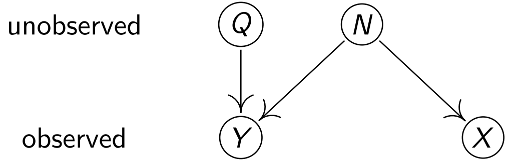


Assume $Y = f(N) + Q$.

Proposed idea:

Remove everything from Y explained by X.

Idea 3: half-sibling regression

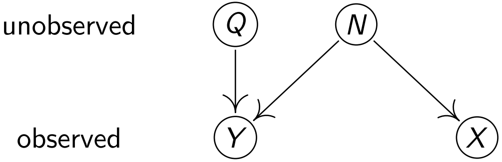


Assume $Y = f(N) + Q$.

Proposed idea:
Remove everything from Y explained by X.

Or: $\hat{Q} := Y - \mathrm{E}[Y \mid X]$.

Idea 3: half-sibling regression



Assume $Y = f(N) + Q$.

Proposed idea:

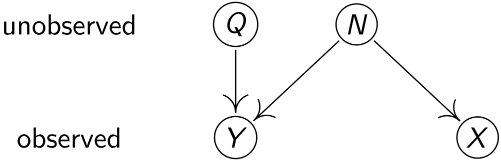
Remove everything from Y explained by X. Or: $\hat{Q} := Y - \mathrm{E}[Y \,|\, X]$.

Proposition

Convergence against "correct" signal Q (up to reparameterization) if

- *perfect reconstruction:* $\exists \psi$ *such that* $f(N) = \psi(X)$

Idea 3: half-sibling regression



Assume $Y = f(N) + Q$.

Proposed idea:

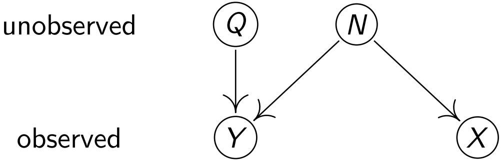
Remove everything from Y explained by X. Or: $\hat{Q} := Y - \mathrm{E}[Y \mid X]$.

Proposition

Convergence against "correct" signal Q (up to reparameterization) if

- *perfect reconstruction:* $\exists \psi$ *such that* $f(N) = \psi(X)$
- *low noise:* $X = g(N) + s \cdot R$ *and* $s \to 0$

Idea 3: half-sibling regression



Assume $Y = f(N) + Q$.

Proposed idea:

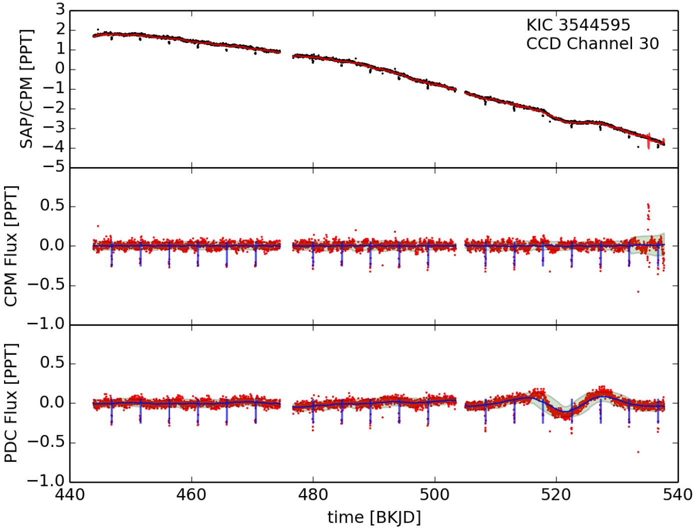
Remove everything from Y explained by X. Or: $\hat{Q} := Y - \mathrm{E}[Y \mid X]$.

Proposition

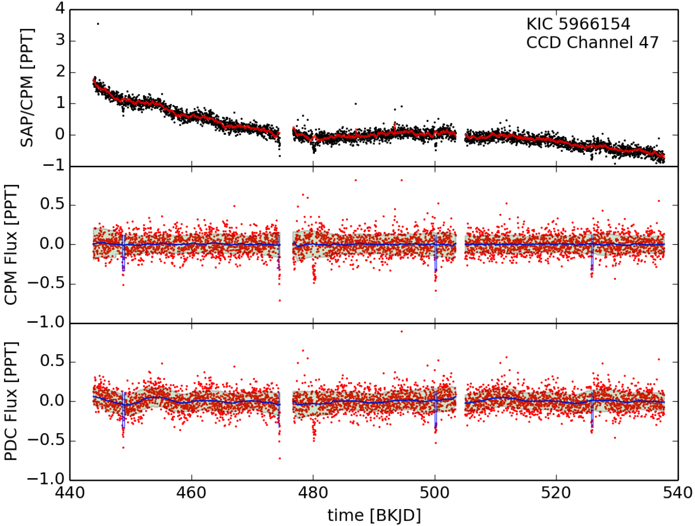
Convergence against "correct" signal Q (up to reparameterization) if

- *perfect reconstruction:* $\exists \psi$ *such that* $f(N) = \psi(X)$
- *low noise:* $X = g(N) + s \cdot R$ *and* $s \to 0$
- *many X's:* $X_i = g_i(N) + R_i,\ i = 1, \dots, \infty$

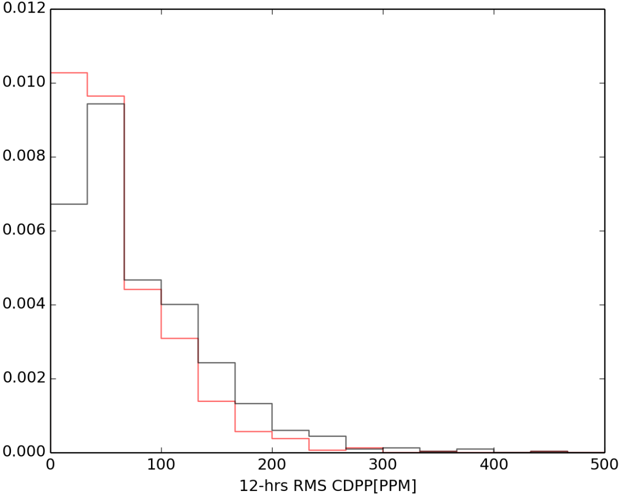
Idea 3: half-sibling regression



Idea 3: half-sibling regression

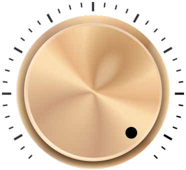


Idea 3: half-sibling regression

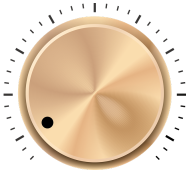


Schölkopf et al.: *Removing systematic errors for exoplanet search via latent causes*, ICML 2015

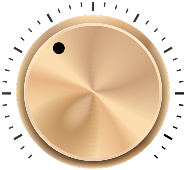
Idea 4: anchor regression

invariance  prediction

Idea 4: anchor regression

invariance  prediction

Idea 4: anchor regression

invariance  prediction

Idea 4: anchor regression

invariance  prediction

Find a trade-off between

- invariance with respect to 

AND
- predictive power

Idea 4: anchor regression

$Y \in \mathrm{R}^1$: target
$X \in \mathrm{R}^{1\times d}$: predictors
$A \in \mathrm{R}^{1\times q}$: anchors, $\mathrm{E}A^tA = Id$

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

Idea 4: anchor regression

$Y \in \mathrm{R}^1$: target
$X \in \mathrm{R}^{1\times d}$: predictors
$A \in \mathrm{R}^{1\times q}$: anchors, $\mathrm{E}A^tA = Id$

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

$\gamma \to 0$: OLS
$\gamma \to \infty$: IV solution (if identifiable)
$\gamma \to \infty$: best invariant predictor (if not identifiable)

Idea 4: anchor regression

$Y \in \mathrm{R}^1$: target
$X \in \mathrm{R}^{1\times d}$: predictors
$A \in \mathrm{R}^{1\times q}$: anchors, $\mathrm{E}A^t A = Id$

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

$\gamma \to 0$:	OLS
$\gamma \to \infty$:	IV solution (if identifiable)
$\gamma \to \infty$:	best invariant predictor (if not identifiable)

- Anchor regression minimizes worst case prediction error under shift interventions.

Rothenhäusler, Bühlmann, Meinshausen, JP (arXiv:1801.06229)

Idea 4: anchor regression

$Y \in \mathrm{R}^1$: target
$X \in \mathrm{R}^{1\times d}$: predictors
$A \in \mathrm{R}^{1\times q}$: anchors, $\mathrm{E}A^t A = Id$

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

$\gamma \to 0$:	OLS
$\gamma \to \infty$:	IV solution (if identifiable)
$\gamma \to \infty$:	best invariant predictor (if not identifiable)

- Anchor regression minimizes worst case prediction error under shift interventions. Rothenhäusler, Bühlmann, Meinshausen, JP (arXiv:1801.06229)
- The finite sample estimator is known as a k-class estimator for IV solution. Theil (1958), Nagar (1959), Jakobsen and JP (arXiv:2005.03353)

Idea 4: anchor regression

$$\begin{pmatrix} X \\ Y \\ H \end{pmatrix} \leftarrow \mathrm{B} \cdot \begin{pmatrix} X \\ Y \\ H \end{pmatrix} + \varepsilon + \mathrm{M}A,$$

$$\text{shifted:} \qquad \begin{pmatrix} X^v \\ Y^v \\ H^v \end{pmatrix} \leftarrow \mathrm{B} \cdot \begin{pmatrix} X^v \\ Y^v \\ H^v \end{pmatrix} + \varepsilon + v.$$

$Id - B$ invertible

Idea 4: anchor regression

$$\begin{pmatrix} X \\ Y \\ H \end{pmatrix} \leftarrow \mathrm{B} \cdot \begin{pmatrix} X \\ Y \\ H \end{pmatrix} + \varepsilon + \mathrm{M}A,$$

shifted: $$\begin{pmatrix} X^v \\ Y^v \\ H^v \end{pmatrix} \leftarrow \mathrm{B} \cdot \begin{pmatrix} X^v \\ Y^v \\ H^v \end{pmatrix} + \varepsilon + v.$$

$Id - B$ invertible

Theorem

For any $b \in \mathbb{R}^d$ we have

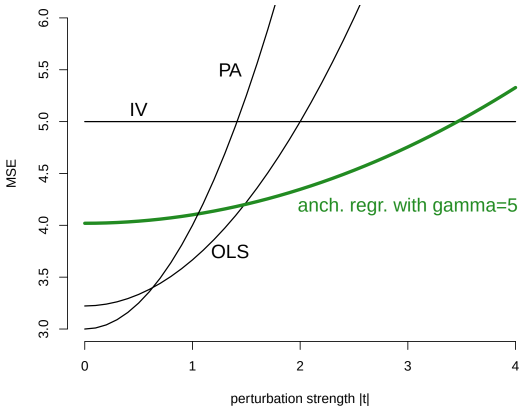
$$\underset{b}{\operatorname{argmin}}\ \mathrm{E}(Y - Xb)^2 + \gamma \|\mathrm{E}A^t(Y - Xb)\|_2^2 = \max_{v \in C^\gamma} \mathbb{E}[(Y^v - X^v b)^2],$$

where

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

Idea 4: anchor regression

MSE under a shift of X (γ fixed):



Idea 4: anchor regression

http://www.srfcdn.ch/radio/modules/dynimages/624/srf-1/2015/01/diverses/264377.150114_raclette_key.jpg

Idea 4: anchor regression

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

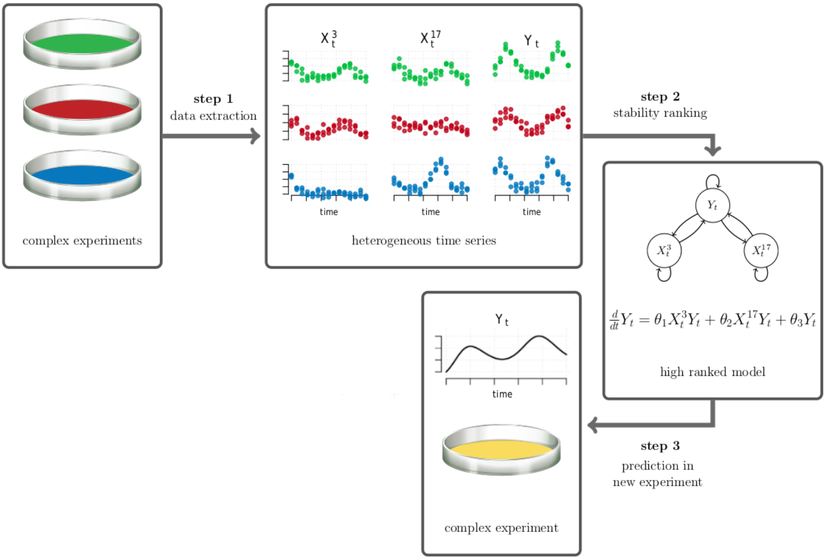
Idea 4: anchor regression

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

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

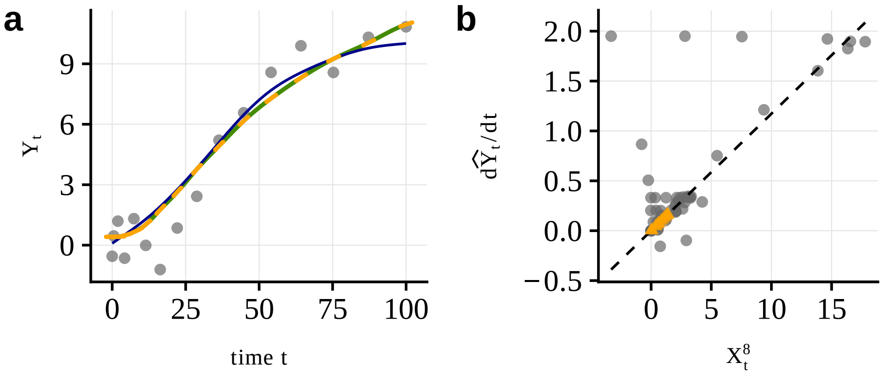
$$\frac{d}{dt}[\mathrm{Mel}]_t = \theta_4[\mathrm{AMP}]_t \qquad \ldots$$

Idea 4: anchor regression



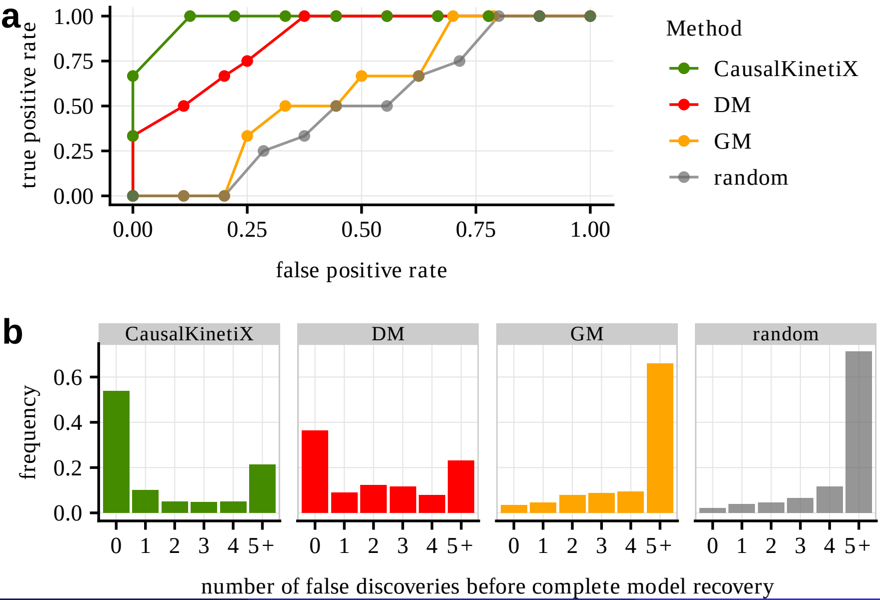
N. Pfister, S. Bauer, JP: *Identifying Causal Structure in Large-Scale Kinetic Systems*, PNAS 2019

Idea 4: anchor regression

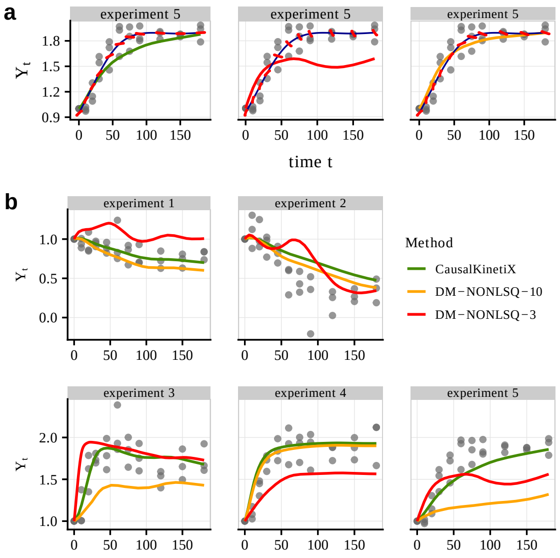


N. Pfister, S. Bauer, JP: *Identifying Causal Structure in Large-Scale Kinetic Systems*, PNAS 2019

Idea 4: anchor regression



Idea 4: anchor regression



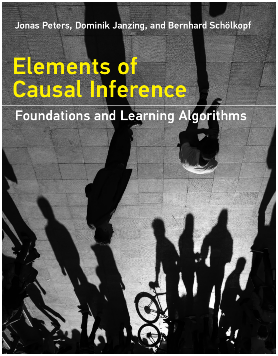
N. Pfister, S. Bauer, JP: *Identifying Causal Structure in Large-Scale Kinetic Systems*, PNAS 2019

Summary Part III:

- Idea 1: reformulate reinforcement learning, use causal structure
- Idea 2: semi-supervised learning from cause to effect does not work
- Idea 3: half-sibling regression
- Idea 4: anchor regression

Summary Part III:

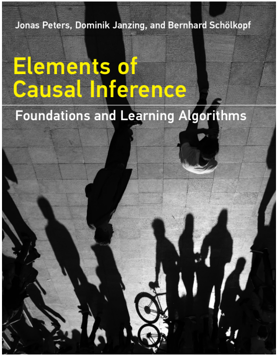
- Idea 1: reformulate reinforcement learning, use causal structure
- Idea 2: semi-supervised learning from cause to effect does not work
- Idea 3: half-sibling regression
- Idea 4: anchor regression



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: semi-supervised learning from cause to effect does not work
- Idea 3: half-sibling regression
- Idea 4: anchor regression



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!