Causality

Jonas Peters
MPI for Intelligent Systems, Tübingen

MLSS, Cádiz

18th May 2016

MAX-PLANCK-GESELLSCHAFT

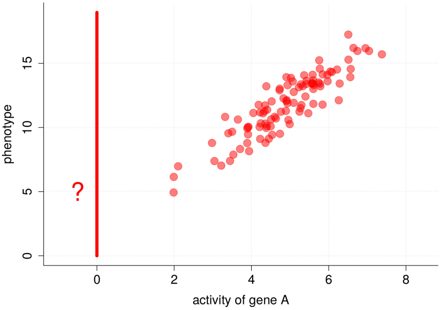
is based on work by ...

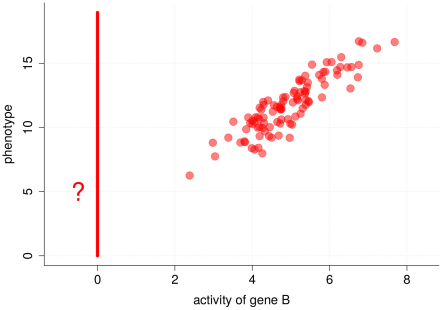
- UCLA: Judea Pearl
- CMU: Peter Spirtes, Clark Glymour, Richard Scheines
- Harvard University: Donald Rubin, Jamie Robins
- ETH Zürich: Peter Bühlmann, Nicolai Meinshausen
- Max-Planck-Institute Tübingen: Dominik Janzing, Bernhard Schölkopf
- University of Amsterdam: Joris Mooij
- Patrik Hoyer
- ... and many others

Step 1: Consider the following problem.

表现

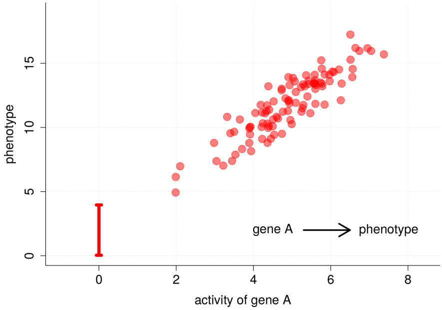
两个不同的基因都和某个症状有极强的相关性

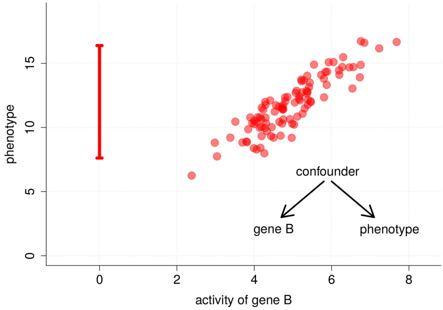




这里提出的问题就是如果我们抑制这其中一个基因的表达，结果会是什么

Step 2: Causality matters!

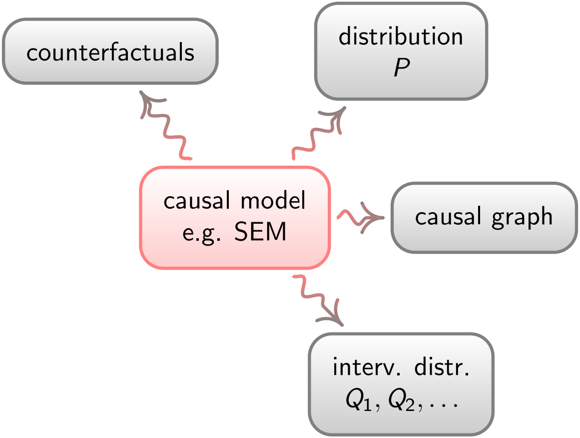




如果我们可以确认A是导致phenotype的原因，那么我们就有更高的确信度，
认为如果我们一直A的表达，预测的phenotyoe会很低

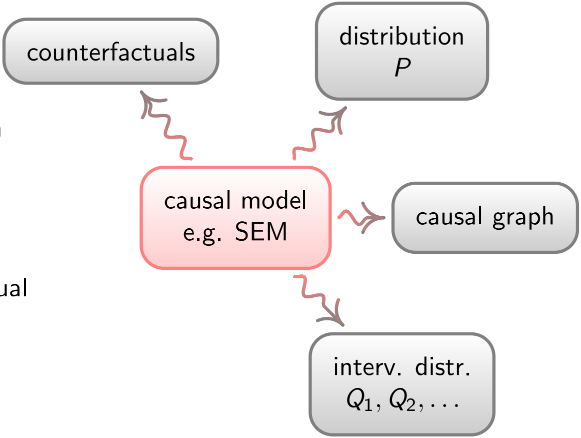
但是对于gene B来说，如果你不能确认它是原因的话，
那么仅仅通过相关性是预测不出干预效果的，因为confounder的作用

Step 3: What is a causal model?

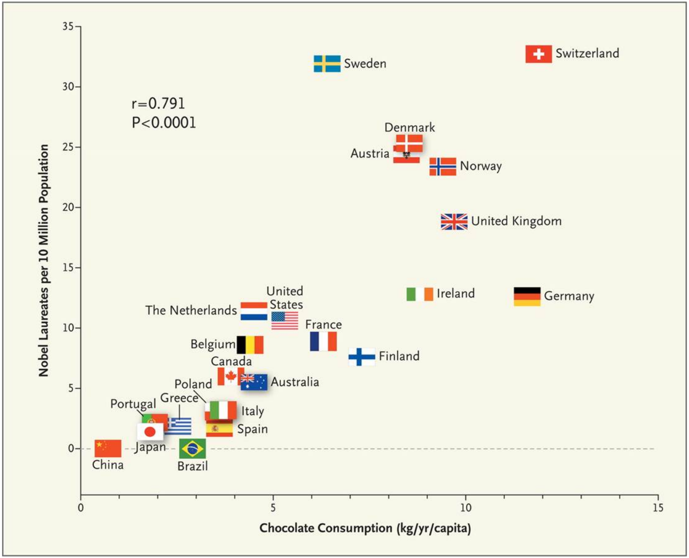


Step 4: What questions are being asked?

- How to compute interventions?
- What if there are hidden variables?
- What are nice graphical representations?
- Can we test counterfactual statements?
- Can we infer the graph structure?

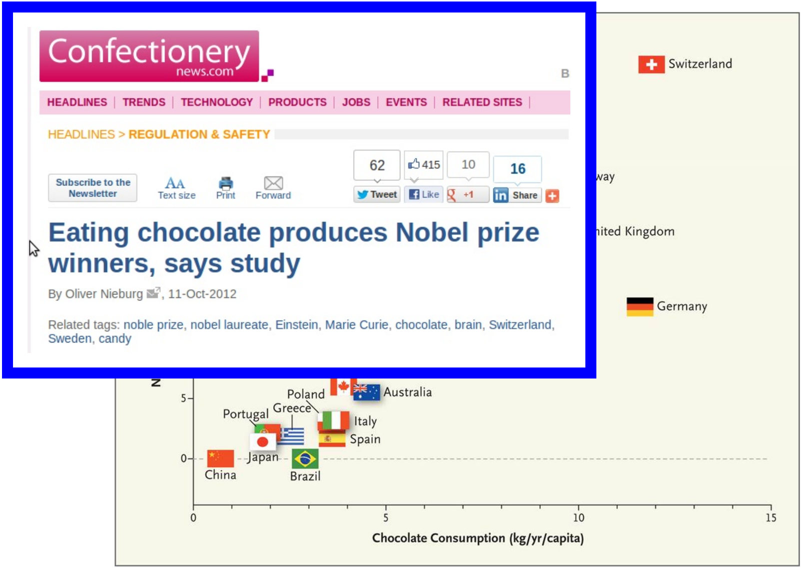


Example: chocolate



F. H. Messerli: *Chocolate Consumption, Cognitive Function, and Nobel Laureates*, N Engl J Med 2012

Example: chocolate



Confectionery news.com

HEADLINES | TRENDS | TECHNOLOGY | PRODUCTS | JOBS | EVENTS | RELATED SITES |

HEADLINES > REGULATION & SAFETY

Subscribe to the Newsletter | Text size | Print | Forward | 62 Tweet | 415 Like | 10 +1 | 16 Share

Eating chocolate produces Nobel prize winners, says study

By Oliver Nieburg, 11-Oct-2012

Related tags: noble prize, nobel laureate, Einstein, Marie Curie, chocolate, brain, Switzerland, Sweden, candy

F. H. Messerli: *Chocolate Consumption, Cognitive Function, and Nobel Laureates*, N Engl J Med 2012

Example: chocolate

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

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 MEDICAL JOURNAL

TABLE VII.—*Estimate of Total Amount of Tobacco Ever Consumed by Smokers; Lung-carcinoma Patients and Control Patients with Diseases Other Than Cancer*

Disease Group	No. Who have Smoked Altogether					Probability Test
	365 Cigs.–	50,000 Cigs.–	150,000 Cigs.–	250,000 Cigs.–	500,000 Cigs.+	
Males:						
Lung-carcinoma patients (647)	19 (2·9%)	145 (22·4%)	183 (28·3%)	225 (34·8%)	75 (11·6%)	$\chi^2=30{\cdot}60$; n=4; P<0·001
Control patients with diseases other than cancer (622)..	36 (5·8%)	190 (30·5%)	182 (29·3%)	179 (28·9%)	35 (5·6%)	
Females:						
Lung-carcinoma patients (41)..	10 (24·4%)	19 (46·3%)	5 (12·2%)	7 (17·1%)	0 (0·0%)	$\chi^2=12{\cdot}97$; n=2; 0·001 < P < 0·01 (Women smoking 15 or more cigarettes a day grouped together)
Control patients with diseases other than cancer (28) ..	19 (67·9%)	5 (17·9%)	3 (10·7%)	1 (3·6%)	0 (0·0%)	

UNG

ouncil

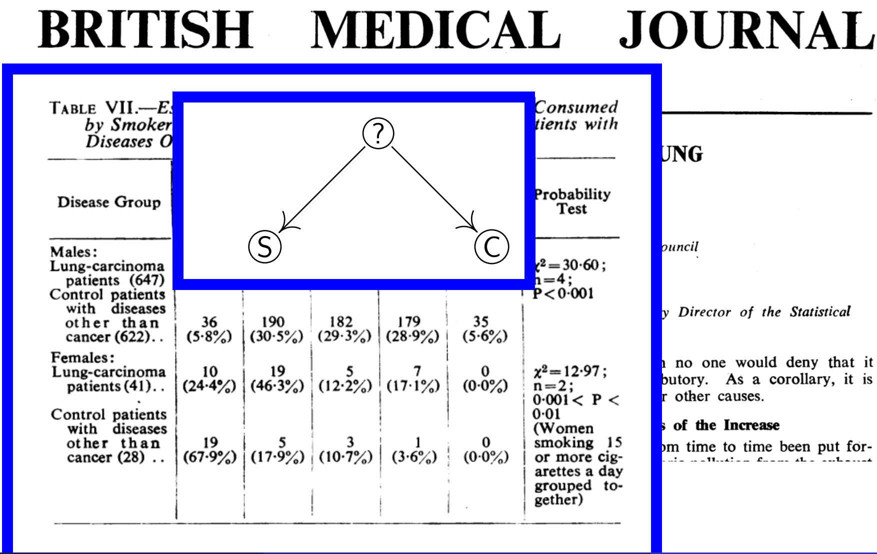
y Director of the Statistical

n no one would deny that it
butory. As a corollary, it is
r other causes.

s of the Increase

om time to time been put for-

Example: smoking



BRITISH MEDICAL JOURNAL

TABLE VII.—E… Consumed by Smoker… …tients with Diseases O…

Disease Group						Probability Test
Males: Lung-carcinoma patients (647)						$\chi^2=30{\cdot}60$; n=4; P<0·001
Control patients with diseases other than cancer (622)..	36 (5·8%)	190 (30·5%)	182 (29·3%)	179 (28·9%)	35 (5·6%)	
Females: Lung-carcinoma patients (41)..	10 (24·4%)	19 (46·3%)	5 (12·2%)	7 (17·1%)	0 (0·0%)	$\chi^2=12{\cdot}97$; n=2; 0·001 < P < 0·01 (Women smoking 15 or more cigarettes a day grouped together)
Control patients with diseases other than cancer (28) ..	19 (67·9%)	5 (17·9%)	3 (10·7%)	1 (3·6%)	0 (0·0%)	

…UNG

…ouncil

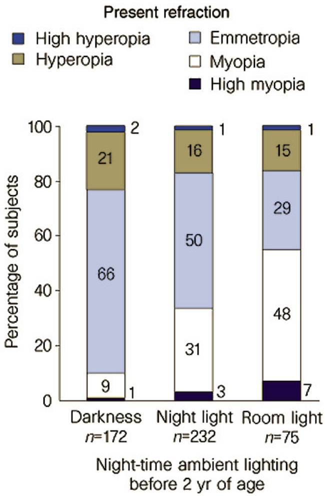
…y Director of the Statistical

…n no one would deny that it …butory. As a corollary, it is …r other causes.

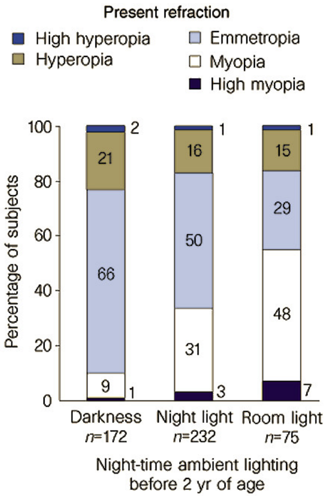
…s of the Increase

…om time to time been put for-

Example: myopia



Example: myopia



"the strength of the association . . . does suggest that the absence of a daily period of darkness during childhood is a potential precipitating factor in the development of myopia"

Quinn, Shin, Maguire, Stone: *Myopia and ambient lighting at night*, Nature 1999

Example: myopia

Patente

Night light with sleep timer

US 20050007889 A1

ZUSAMMENFASSUNG

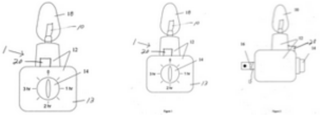
A timer a light and an optional music source is located on or in a housing of a nightlight assembly. When this assembly is plugged into a source of electric power, the timer is set to a selected time for the light and optional music to remain on. After this selected time has elapsed, the light and music automatically turns off, allowing for sleep in appropriate darkness and silence.

Veröffentlichungsnummer	US20050007889 A
Publikationstyp	Anmeldung
Anmeldenummer	US 10/614,245
Veröffentlichungsdatum	13. Jan. 2005
Eingetragen	8. Juli 2003
Prioritätsdatum	8. Juli 2003
Erfinder	Karin Peterson
Ursprünglich Bevollmächtigter	Peterson Karin Lyn
Zitat exportieren	BiBTeX, EndNote, R

Klassifizierungen (4)

Externe Links: USPTO, USPTO-Zuordnung, Esp

BILDER (3)



BESCHREIBUNG

ANSPRÜCHE (18)

Example: myopia

Patente

Night light with sleep timer

US 20050007889 A1

ZUSAMMENFASSUNG

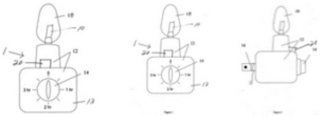
A timer a light and an optional music source is located on or in a housing of a nightlight assembly. When this assembly is plugged into a source of electric power, the timer is set to a selected time for the light and optional music to remain on. After this selected time has elapsed, the light and music automatically turns off, allowing for sleep in appropriate darkness and silence.

Veröffentlichungsnummer	US20050007889 A
Publikationstyp	Anmeldung
Anmeldenummer	US 10/614,245
Veröffentlichungsdatum	13. Jan. 2005
Eingetragen	8. Juli 2003
Prioritätsdatum	8. Juli 2003
Erfinder	Karin Peterson
Ursprünglich Bevollmächtigter	Peterson Karin Lyn
Zitat exportieren	BiBTeX, EndNote, R

Klassifizierungen (4)

Externe Links: USPTO, USPTO-Zuordnung, Esp

BILDER (3)



Question: Does the night light with sleep timer help?

BESCHREIBUNG

ANSPRÜCHE (18)

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

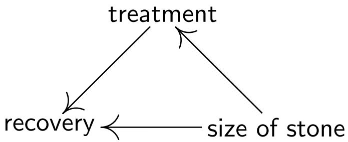
Example: kidney stones

	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

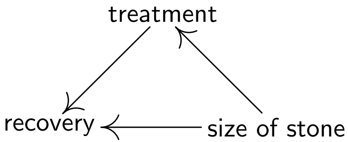
Example: kidney stones

underlying ground truth:



Example: kidney stones

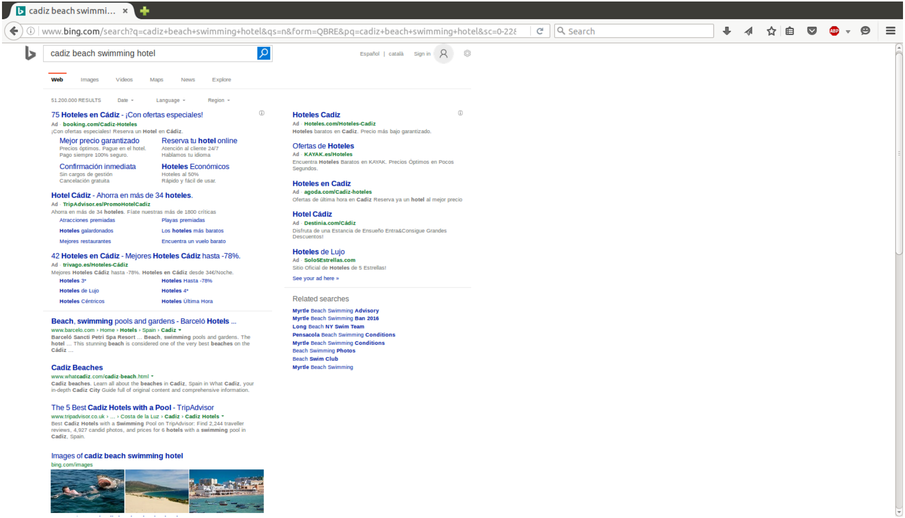
underlying ground truth:



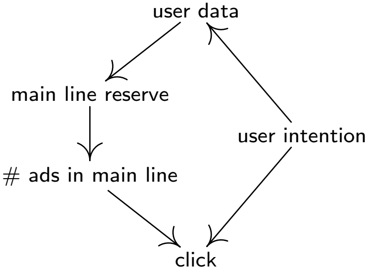
Question: What is the expected recovery if all get treatment B?

(Make treatment independent of size.)

Example: advertisement

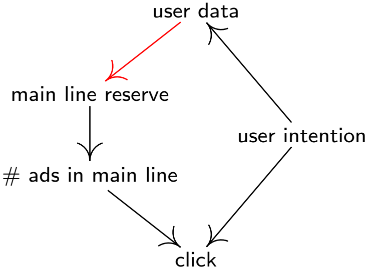


Example: advertisement



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

Example: advertisement



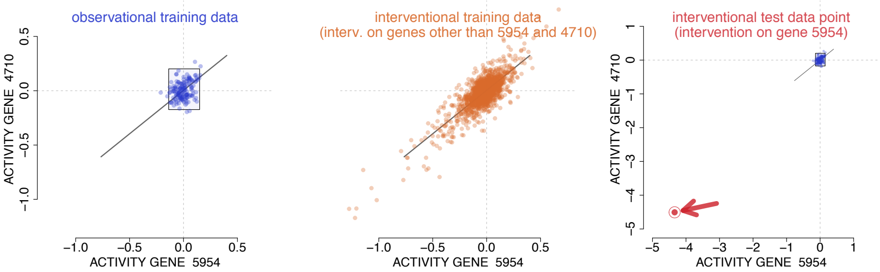
Question: How do we choose an optimal main line reserve?

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

Example: gene interactions

genetic perturbation experiments for yeast

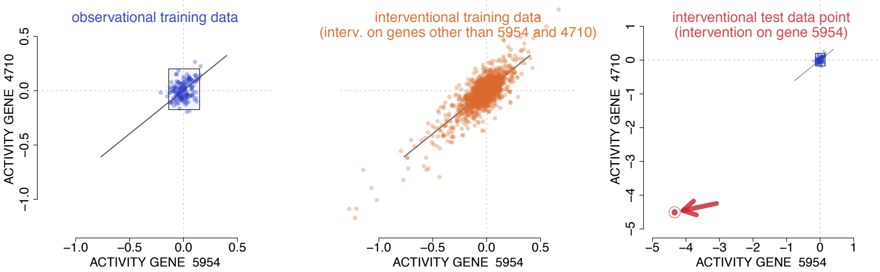
- $p = 6170$ genes
- $n_{obs} = 160$ wild-types
- $n_{int} = 1479$ gene deletions (targets known)



Example: gene interactions

genetic perturbation experiments for yeast

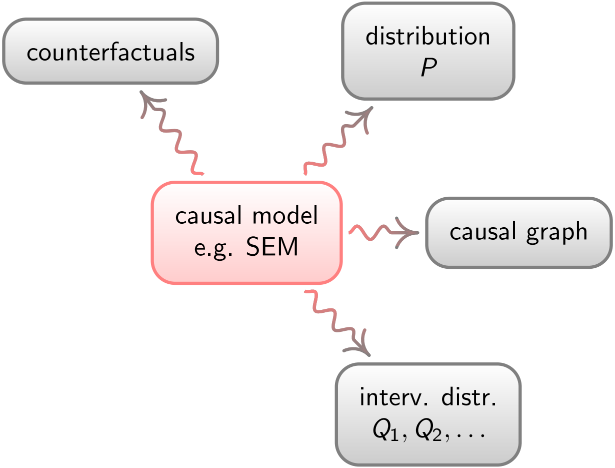
- $p = 6170$ genes
- $n_{obs} = 160$ wild-types
- $n_{int} = 1479$ gene deletions (targets known)



- Causal relationships are often stable!

Kemmeren et al.: Large-scale genetic perturbations reveal reg. networks and an abundance of gene-specific repressors. Cell, 2014

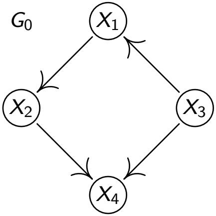
Part I: Causal Language and causal reasoning


counterfactuals
distribution
P
causal model
e.g. SEM
causal graph
interv. distr.
$Q_1, Q_2, \ldots$

SEMs: structural equations with noise distribution.

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

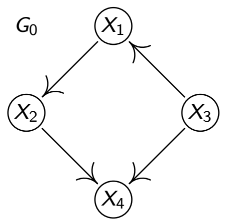
- N_i jointly independent
- G_0 has no cycles



SEMs model observational distributions over $X_1, \dots, X_d$.

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

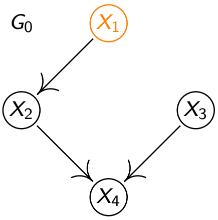
- N_i jointly independent
- G_0 has no cycles



SEMs can model interventions, too.

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

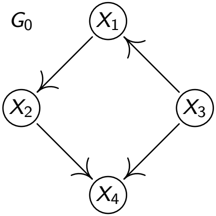
- N_i jointly independent
- G_0 has no cycles



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

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

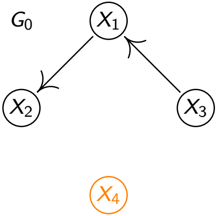
- N_i jointly independent
- G_0 has no cycles



SEMs can model interventions, too.

$$X_1 := f_1(X_3, N_1)$$
$$X_2 := f_2(X_1, N_2)$$
$$X_3 := f_3(N_3)$$
$$X_4 := 0$$

- N_i jointly independent
- G_0 has no cycles

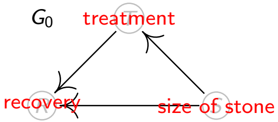


Example: kidney stones

Given: graph and P.

$$
\begin{aligned}
T &:= f_1(S, N_1)\\
R &:= f_2(T, S, N_2)\\
S &:= f_3(N_3)
\end{aligned}
$$

- N_i jointly independent
- G_0 has no cycles

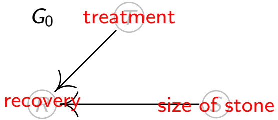


Example: kidney stones

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

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

- N_i jointly independent
- G_0 has no cycles

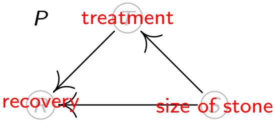


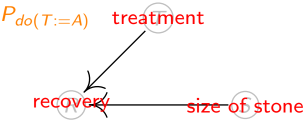
IMPORTANT: modularity, autonomy: Aldrich 1989, Pearl 2009, Schölkopf et al. 2012, ...

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

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

Definition

Given an SEM, there is a total causal effect from X to Y if one of the following equivalent statements is satisfied:

Definition

Given an SEM, there is a total causal effect from X to Y if one of the following equivalent statements is satisfied:

(i) $X \not\perp\!\!\!\perp Y \quad$ in $P_{\mathrm{do}\, X := \tilde{N}_X}$ for some random variable $\tilde{N}_X$.

(ii) There are $x^{\triangle}$ and $x^{\square}$, such that $P^Y_{\mathrm{do}\, X := x^{\triangle}} \neq P^Y_{\mathrm{do}\, X := x^{\square}}$.

(iii) There is $x^{\triangle}$, such that $P^Y_{\mathrm{do}\, X := x^{\triangle}} \neq P^Y$.

(iv) $X \not\perp\!\!\!\perp Y$ in $P^{X,Y}_{\mathrm{do}\, X := \tilde{N}_X}$ for any $\tilde{N}_X$ whose distribution has full support.

Definition

Given an SEM, there is a total causal effect from X to Y if one of the following equivalent statements is satisfied:

(i) $X \not\perp\!\!\!\perp Y \quad$ in $P_{\mathrm{do}\, X := \tilde{N}_X}$ for some random variable $\tilde{N}_X$.

(ii) There are $x^{\triangle}$ and $x^{\square}$, such that $P^{Y}_{\mathrm{do}\, X := x^{\triangle}} \neq P^{Y}_{\mathrm{do}\, X := x^{\square}}$.

(iii) There is $x^{\triangle}$, such that $P^{Y}_{\mathrm{do}\, X := x^{\triangle}} \neq P^{Y}$.

(iv) $X \not\perp\!\!\!\perp Y$ in $P^{X,Y}_{\mathrm{do}\, X := \tilde{N}_X}$ for any $\tilde{N}_X$ whose distribution has full support.

Causal strength?

Definition

Given an SEM, there is a total causal effect from X to Y if one of the following equivalent statements is satisfied:

(i) $X \not\perp\!\!\!\perp Y$ in $P_{\mathrm{do}\, X := \tilde{N}_X}$ for some random variable $\tilde{N}_X$.

(ii) There are $x^{\triangle}$ and $x^{\square}$, such that $P^{Y}_{\mathrm{do}\, X := x^{\triangle}} \neq P^{Y}_{\mathrm{do}\, X := x^{\square}}$.

(iii) There is $x^{\triangle}$, such that $P^{Y}_{\mathrm{do}\, X := x^{\triangle}} \neq P^{Y}$.

(iv) $X \not\perp\!\!\!\perp Y$ in $P^{X,Y}_{\mathrm{do}\, X := \tilde{N}_X}$ for any $\tilde{N}_X$ whose distribution has full support.

Causal strength? $\rightsquigarrow$ your next paper :)

Summary Part I:

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

Summary Part I:

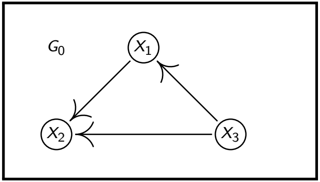
- What if interested in iid prediction, i.e. observational data? Don't worry (too much) about causality!
- But often, we are interested in a system's behaviour under intervention.

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.
- SEMs 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

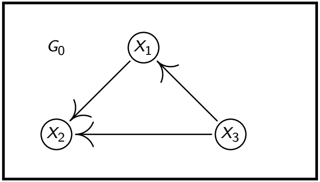


Summary Part I:

- 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.
- SEMs entail graphs, obs. distr., interventions and counterfactuals.

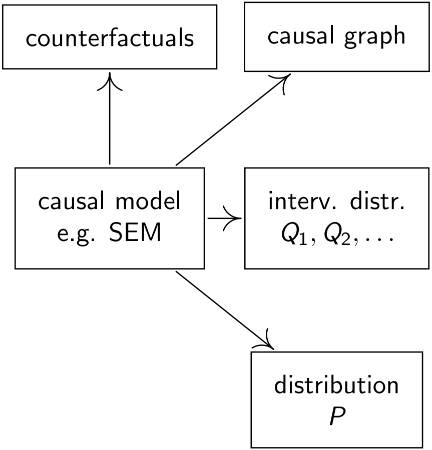
$$X_1 := f_1(X_3, N_1)$$
$$X_2 := f_2(X_1, X_3, N_2)$$
$$X_3 := f_3(N_3)$$

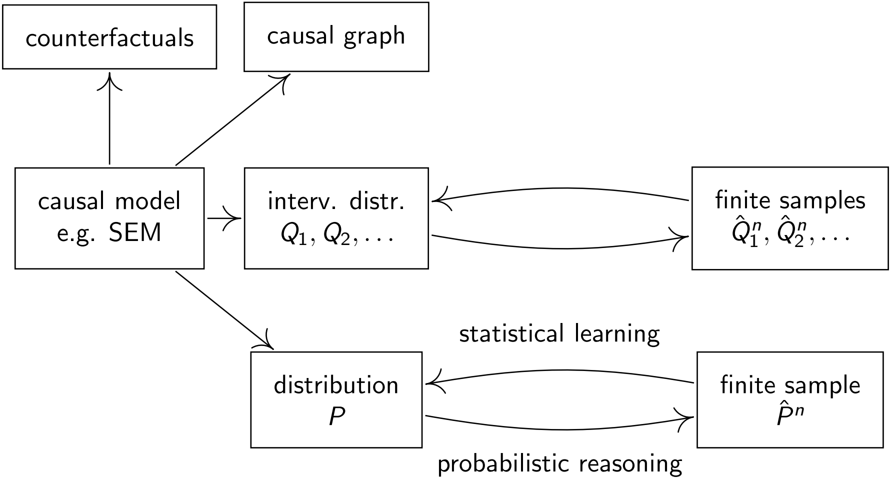
- N_i jointly independent
- G_0 has no cycles



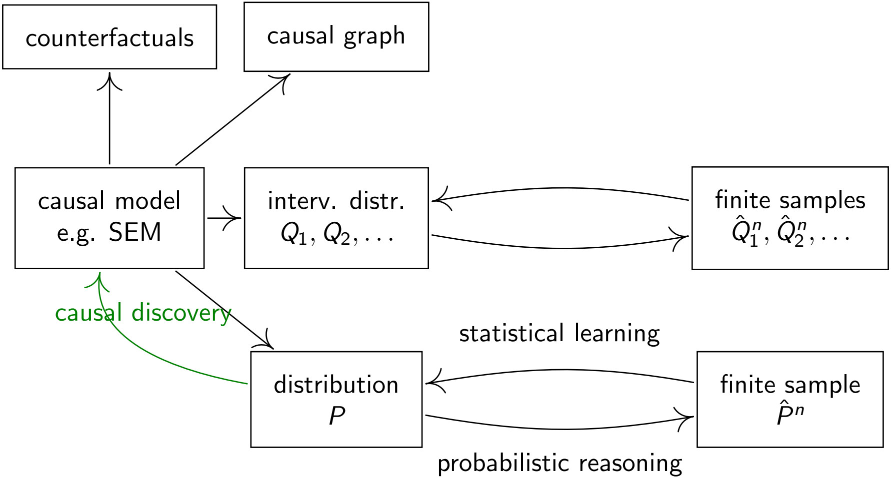
- graph + observational distribution $\rightsquigarrow$ interventions (by adjusting)
- ... even possible if there are (some) hidden variables

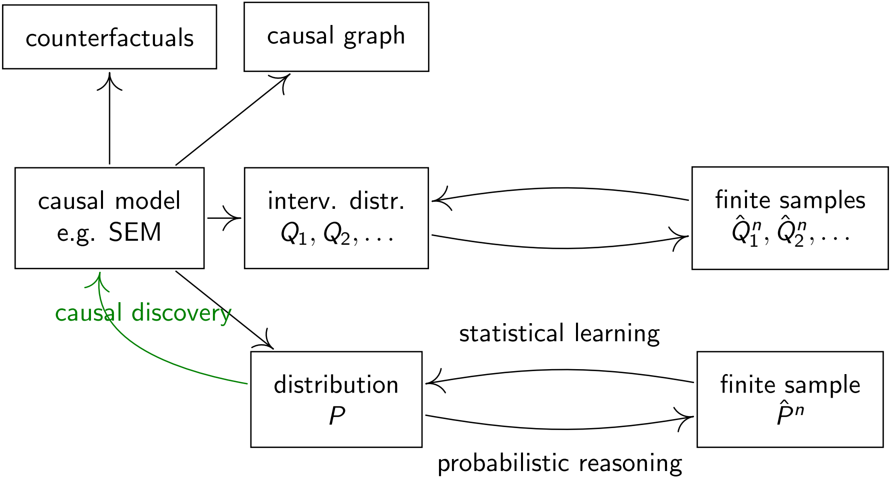
Part II: Causal Discovery


counterfactuals
causal graph
causal model
e.g. SEM
interv. distr.
$Q_1, Q_2, \ldots$
distribution
P

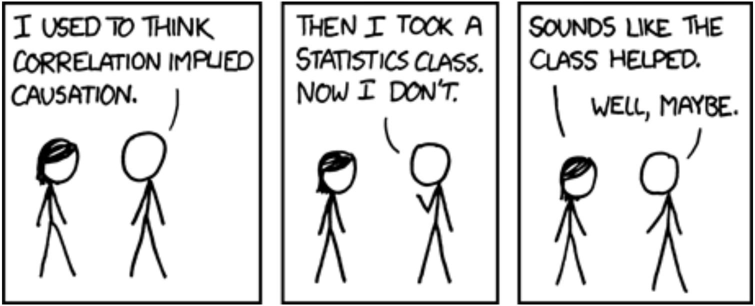


counterfactuals
causal graph
causal model
e.g. SEM
interv. distr.
$Q_1, Q_2, \ldots$
finite samples
$\hat{Q}_1^n, \hat{Q}_2^n, \ldots$
statistical learning
distribution
P
finite sample
$\hat{P}^n$
probabilistic reasoning


counterfactuals
causal graph
causal model
e.g. SEM
interv. distr.
$Q_1, Q_2, \ldots$
finite samples
$\hat{Q}_1^n, \hat{Q}_2^n, \ldots$
causal discovery
statistical learning
distribution
P
finite sample
$\hat{P}^n$
probabilistic reasoning



Required:
Relation between distribution P and SEM.


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

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:

$$\begin{tikzcd} \text{X: Spain } \heartsuit \arrow[dr] & & \text{Y: Machine Learning } \heartsuit \arrow[dl] \\ & \text{Z: attending MLSS} & \end{tikzcd}$$

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:

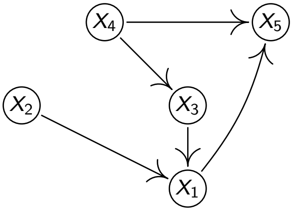
$$
\begin{tikzcd}
\text{X: Spain } \heartsuit \arrow[dr] & & \text{Y: Machine Learning } \heartsuit \arrow[dl] \\
& \text{Z: attending MLSS} &
\end{tikzcd}
$$

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

Definition: graphs

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

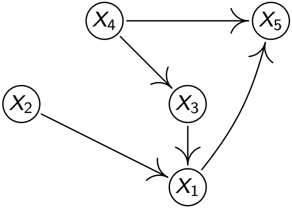
- parents, children, descendants, ancestors, ...
- paths, directed paths
- immoralities (or v-structures)
- d-separation (see next)
- ...



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 $\mathcal{S}$.

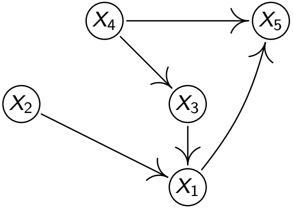
Check, whether all paths blocked!!



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!!

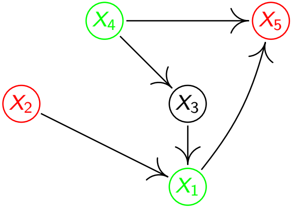


$\circ \cdots \rightarrow \circ \rightarrow \cdots \circ$ blocks a path.
$\circ \cdots \leftarrow \circ \rightarrow \cdots \circ$ blocks a path.
$\circ \cdots \rightarrow \circ \leftarrow \cdots \circ$ blocks a path.

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.

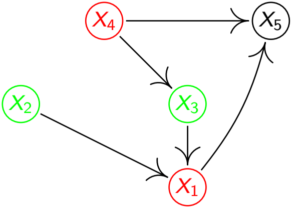
$\circ \cdots \to \circ \leftarrow \cdots \ \circ$ blocks a path.

X_2 and X_5 are d-sep. by $\{X_1, X_4\}$

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.

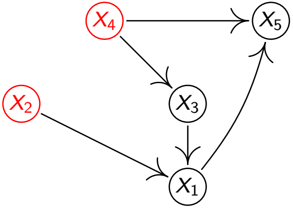
$\circ \cdots \to \circ \leftarrow \cdots \circ$ blocks a path.

X_2 and X_5 are d-sep. by $\{X_1, X_4\}$
X_4 and X_1 are d-sep. by $\{X_2, X_3\}$

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.

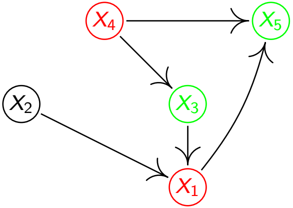
$\circ \cdots \rightarrow \circ \leftarrow \cdots \ \circ$ blocks a path.

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 $\{\}$

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 is Markov w.r.t. G if

$$X_i \text{ and } X_j \text{ are } d\text{-separated by } \mathcal{S} \text{ in } G \quad \Rightarrow \quad X_i \perp\!\!\!\perp X_j \,|\, \mathcal{S}$$

Definition

P is Markov w.r.t. G if

$$X_i \text{ and } X_j \text{ are } d\text{-separated by } \mathcal{S} \text{ in } G \quad \Rightarrow \quad X_i \perp\!\!\!\perp X_j \,|\, \mathcal{S}$$

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 is Markov w.r.t. G if

$$X_i \text{ and } X_j \text{ are } d\text{-separated by } \mathcal{S} \text{ in } G \quad \Rightarrow \quad X_i \perp\!\!\!\perp X_j \,|\, \mathcal{S}$$

Definition

P is Markov w.r.t. G if

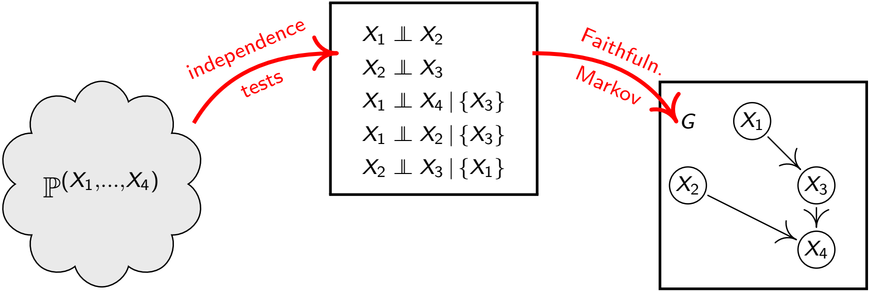
$$X_i \text{ and } X_j \text{ are } d\text{-separated by } \mathcal{S} \text{ in } G \quad \Rightarrow \quad X_i \perp\!\!\!\perp X_j \,|\, \mathcal{S}$$

Definition

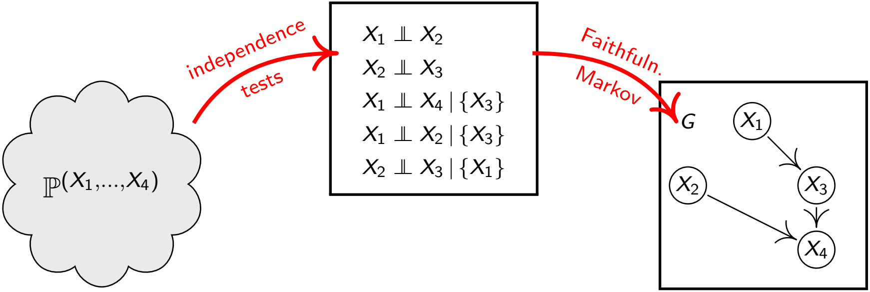
P is faithful w.r.t. G if

$$X_i \text{ and } X_j \text{ are } d\text{-separated by } \mathcal{S} \text{ in } G \quad \Leftarrow \quad X_i \perp\!\!\!\perp X_j \,|\, \mathcal{S}$$

Idea 1: independence-based methods



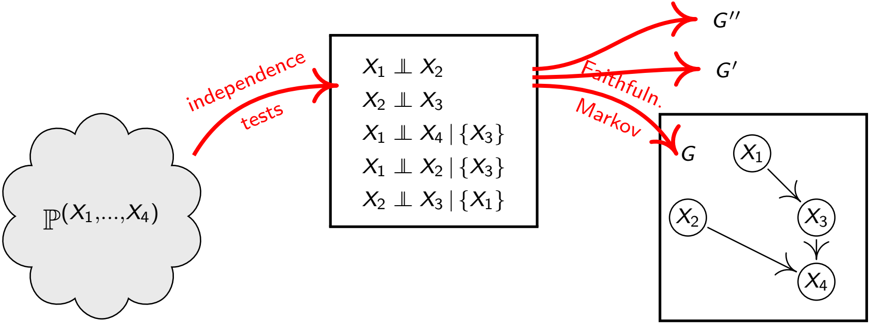
Idea 1: independence-based methods



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.

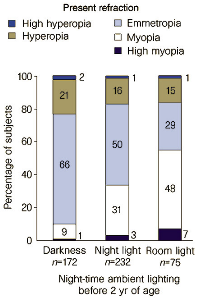
Idea 1: independence-based methods



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.

Example: myopia



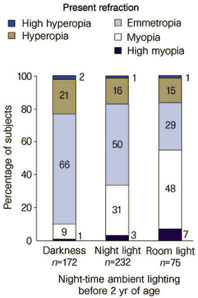
We have

- night light $\not\perp\!\!\!\perp$ child myopia
- night light $\perp\!\!\!\perp$ child myopia $|$ parent myopia
- no other independences

Quinn et al.: *Myopia and ambient lighting at night*, Nature 1999
Zadnik et al.: *Vision: Myopia and ambient night-time light.*, Nature 2000
Gwiazda et al.: *Vision: Myopia and ambient night-time light.*, Nature 2000

and therefore ...

Example: myopia

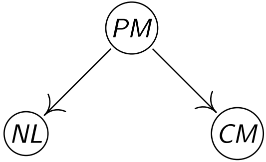


We have

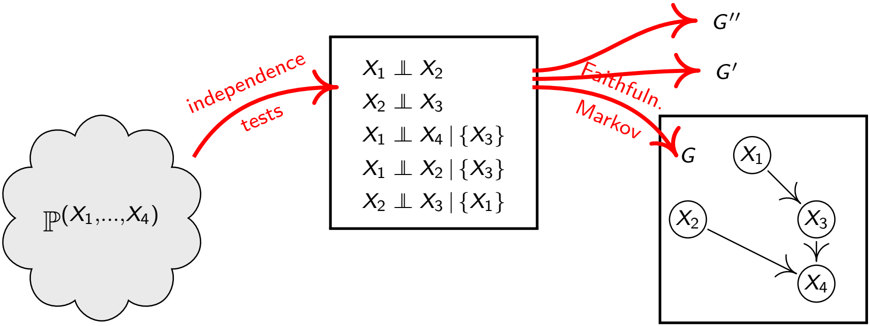
- night light $\not\perp\!\!\!\perp$ child myopia
- night light $\perp\!\!\!\perp$ child myopia | parent myopia
- no other independences

Quinn et al.: *Myopia and ambient lighting at night*, Nature 1999
Zadnik et al.: *Vision: Myopia and ambient night-time light.*, Nature 2000
Gwiazda et al.: *Vision: Myopia and ambient night-time light.*, Nature 2000

and therefore ...



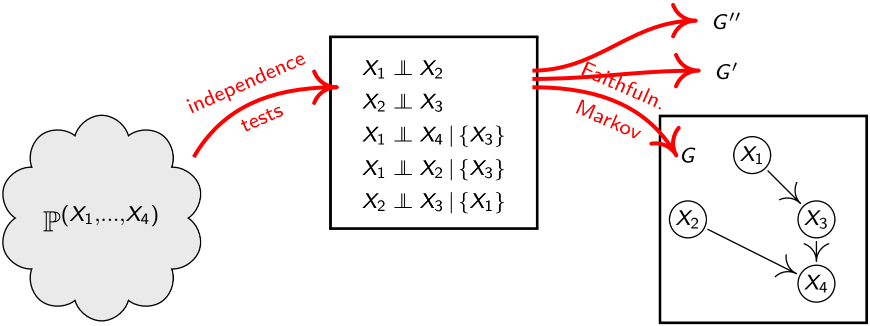
Idea 1: independence-based methods



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

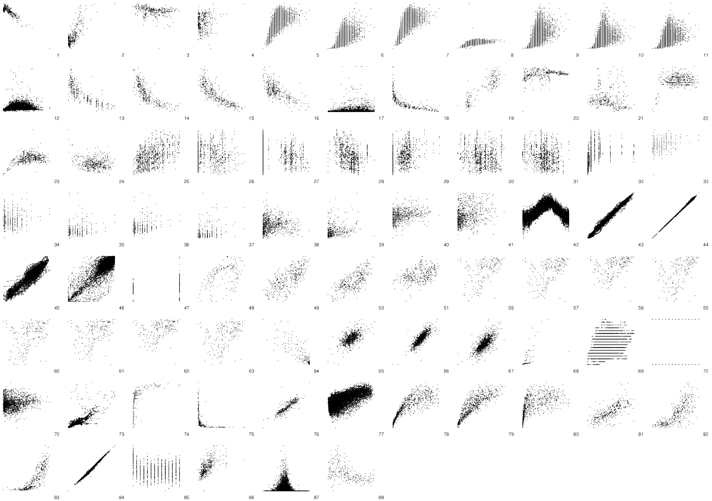


Method: ~~IC (Pearl 2009);~~ PC, FCI (Spirtes et al., 2000)

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

What do we do with two variables?

Idea 2: restricted structural equation models



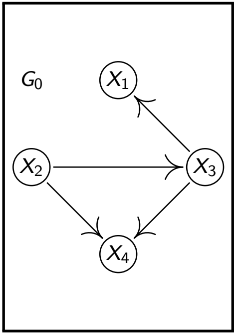
Mooij, JP, Janzing, Zscheischler, Schölkopf: *Disting. cause from effect using obs. data: methods and benchm.*, submitted

Idea 2: restricted structural equation 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



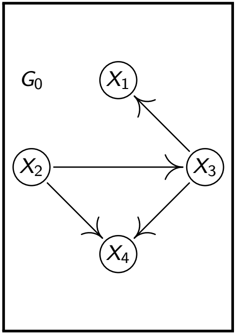
Structural equation model.
Can the DAG be recovered from $P(X_1, \ldots, X_4)$?

Idea 2: restricted structural equation 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



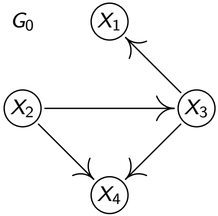
Structural equation model.
Can the DAG be recovered from $P(X_1, \ldots, X_4)$? **No.**

Idea 2: restricted structural equation 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 2: restricted structural equation 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 2: restricted structural equation models

Consider a distribution entailed by

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

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

$$X \longrightarrow Y$$

Then, if f is nonlinear, there is no

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

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

$$X \longleftarrow Y$$

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

Idea 2: restricted structural equation models

Consider a distribution corresponding to

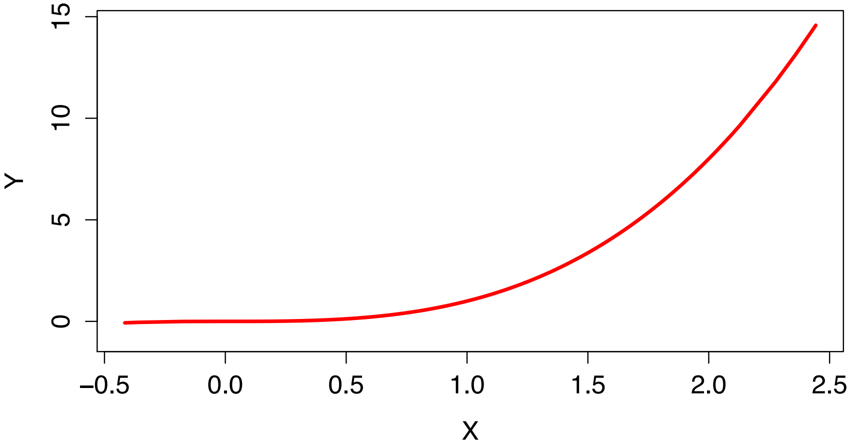
$$\boxed{\begin{array}{c} Y = X^3 + N_Y \\[1em] \text{with } N_Y, X \overset{ind}{\sim} \mathcal{N} \end{array}}$$

$$X \longrightarrow Y$$

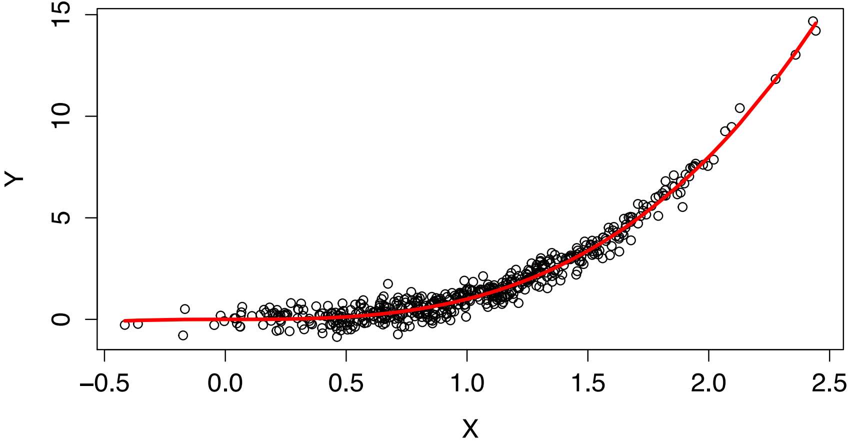
with

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

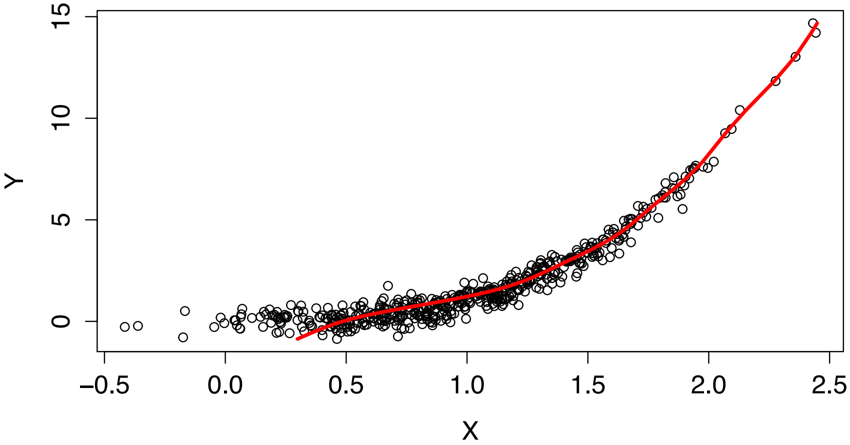
Idea 2: restricted structural equation models



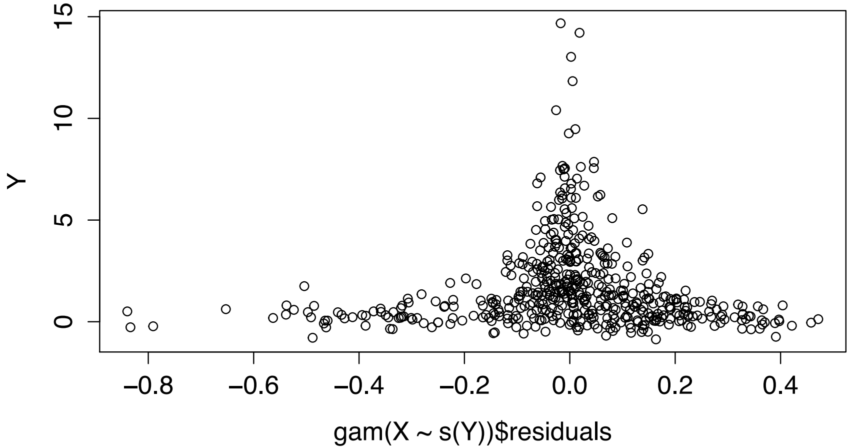
Idea 2: restricted structural equation models



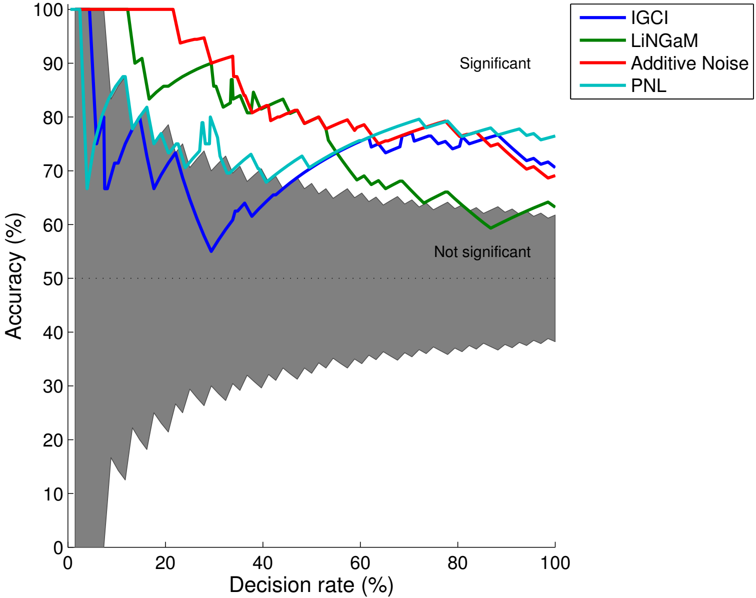
Idea 2: restricted structural equation models



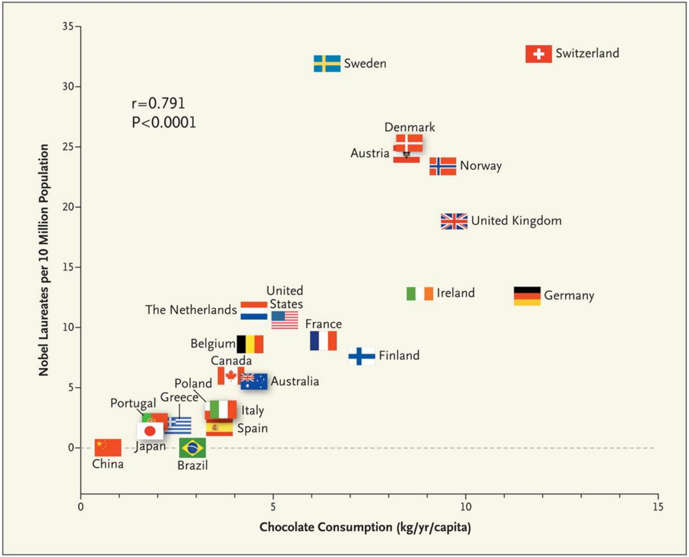
Idea 2: restricted structural equation models



Real Data: cause-effect pairs



Example: chocolate



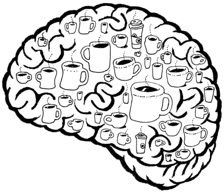
F. H. Messerli: *Chocolate Consumption, Cognitive Function, and Nobel Laureates*, N Engl J Med 2012

Example: chocolate

No (not enough) data for chocolate

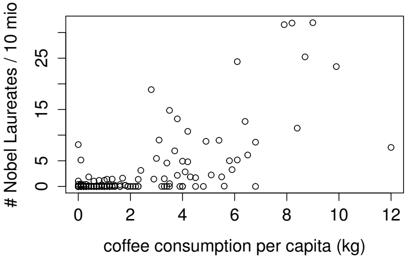
Example: chocolate

No (not enough) data for chocolate



... but we have data for coffee!

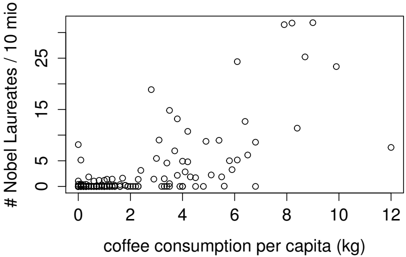
Example: chocolate



Correlation: 0.698

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

Example: chocolate

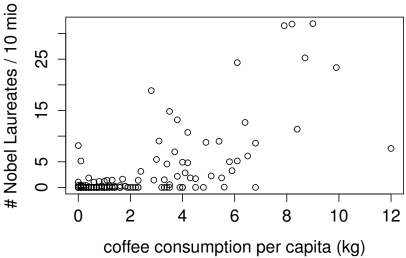


Correlation: 0.698
p-value: $< 2.2 \cdot 10^{-16}$

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: chocolate



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?

Idea 2: restricted structural equation 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 2: restricted structural equation models

Slightly surprising:

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

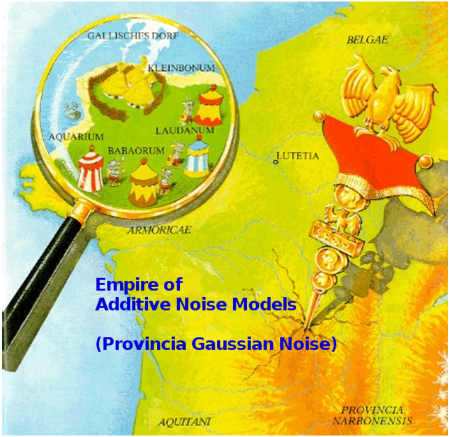
Peters et al.: *Identifiability of Causal Graphs using Functional Models*, UAI 2011

Let $P(X_1, \ldots, X_p)$ be entailed by an ...

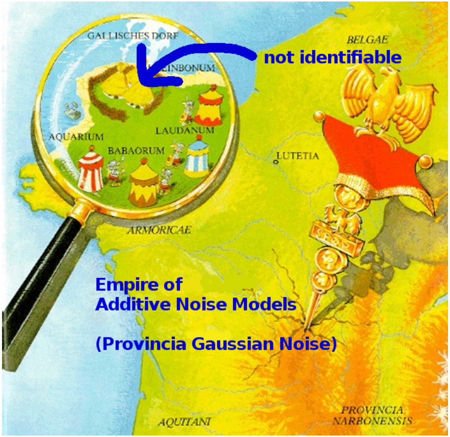
		conditions	identif.
structural equation model:	$X_i = f_i(X_{\mathbf{PA}_i}, N_i)$	-	✗
additive noise model:	$X_i = f_i(X_{\mathbf{PA}_i}) + N_i$	nonlin. fct.	✓
causal additive model:	$X_i = \sum_{k \in \mathbf{PA}_i} f_{ik}(X_k) + N_i$	nonlin. fct.	✓
linear Gaussian model:	$X_i = \sum_{k \in \mathbf{PA}_i} \beta_{ik} X_k + N_i$	linear fct.	✗

(results hold for Gaussian noise)

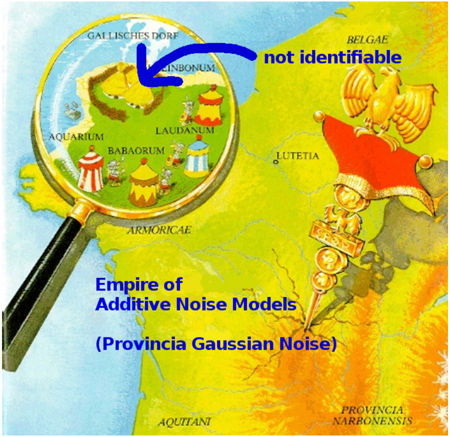
Idea 2: restricted structural equation models



Idea 2: restricted structural equation models

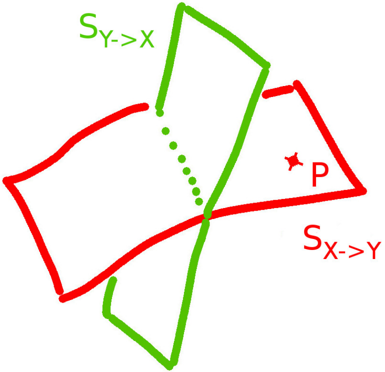


Idea 2: restricted structural equation models

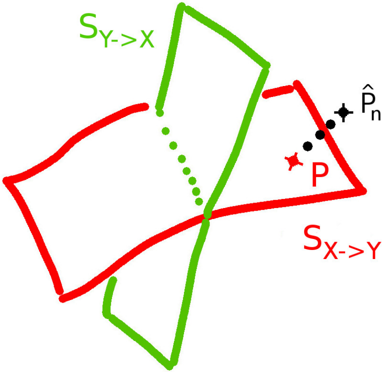


~~GAUL~~ GAUSS
"the LINEAR"

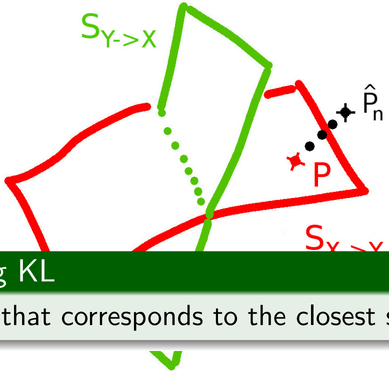
Idea 2: restricted structural equation models



Idea 2: restricted structural equation models



Idea 2: restricted structural equation models



Method: Minimizing KL

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

Idea 2: restricted structural equation 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 2: restricted structural equation models

Consider model classes

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

Define

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

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

Idea 2: restricted structural equation models

Consider model classes

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

Define

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

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

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

Idea 2: restricted structural equation models

Consider model classes

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

Define

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

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

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

Idea 2: restricted structural equation models

p	number of DAGs with p 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
23	69743329837281492647141549700245804876504274990515985894109106401549811985510951501377122074625

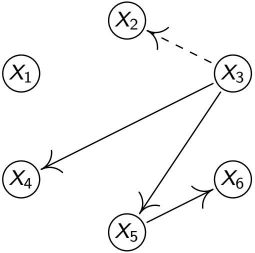
`https://oeis.org/A003024/b003024.txt`

Idea 2: restricted structural equation models

E.g. greedy search!

–	0.2	0.1	0.1	0.1	0.3
0.4	–	0.1	0.1	0.1	0.1
0.1	0.6	–	–	–	0.4
0.1	0.1	–	–	0.1	0.1
0.1	0.1	–	0.1	–	–
0.3	0.1	–	0.1	–	–

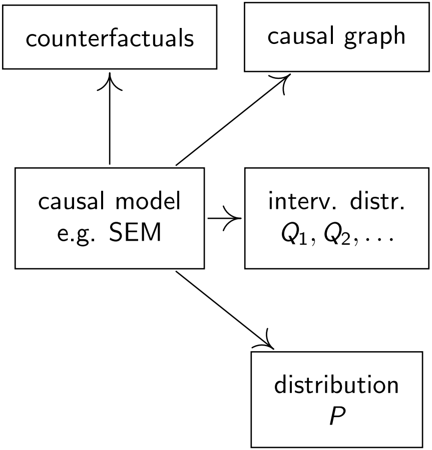
$\xrightarrow[\text{recompute column}]{\text{include best edge}}$



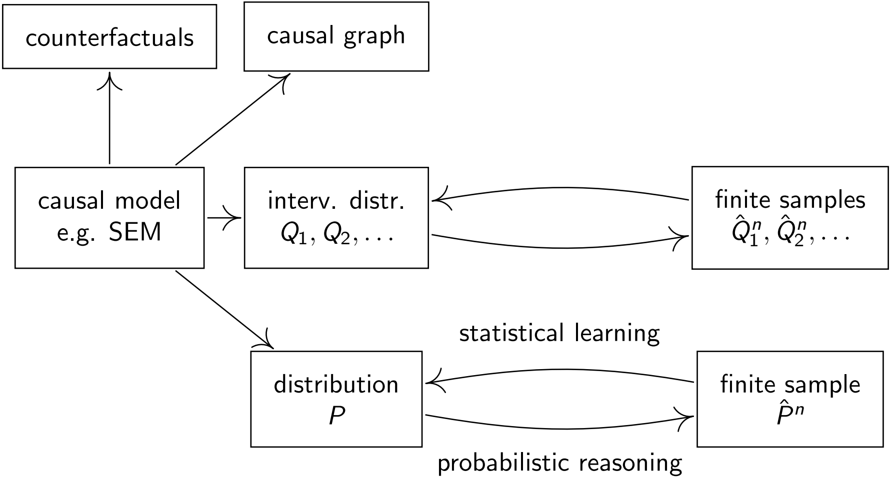
Greedy Addition (e.g. Chickering 2002). Include the edge that leads to the largest increase of the log-likelihood.

Bühlmann, JP, Ernest: *CAM: Causal add. models, high-dim. order search and penalized regr.*, Annals of Statistics 2014

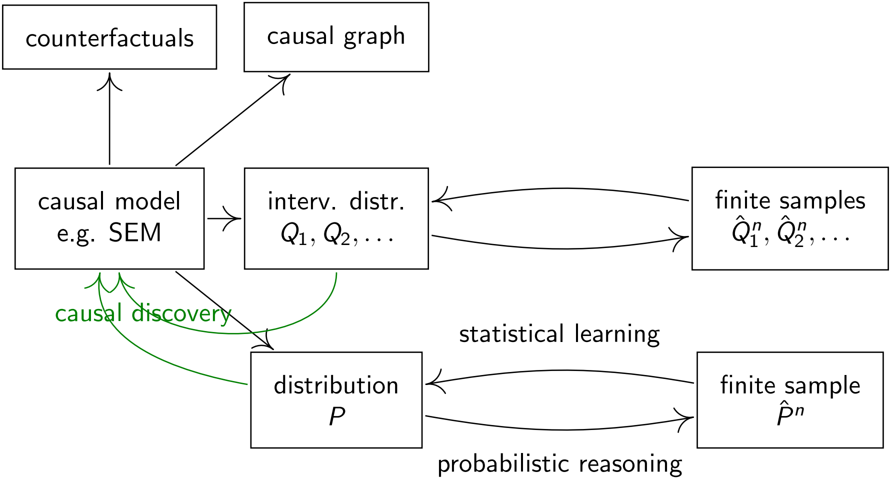
Idea 3: invariant causal prediction



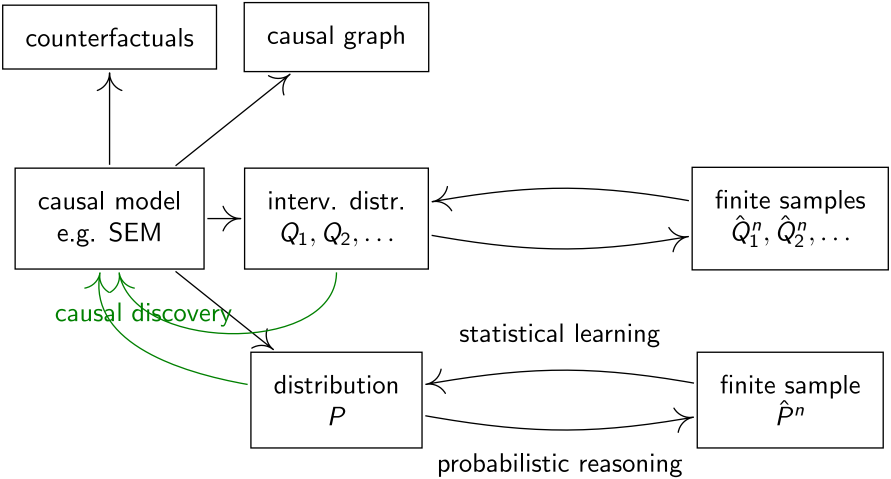
Idea 3: invariant causal prediction



Idea 3: invariant causal prediction

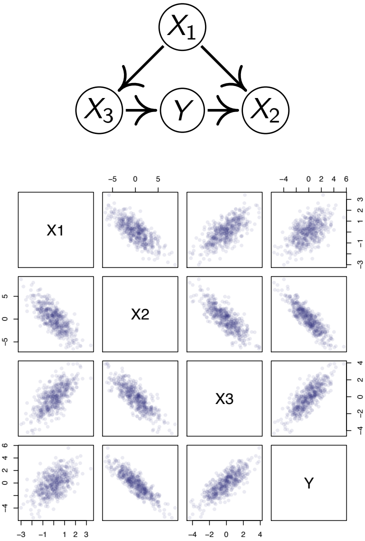


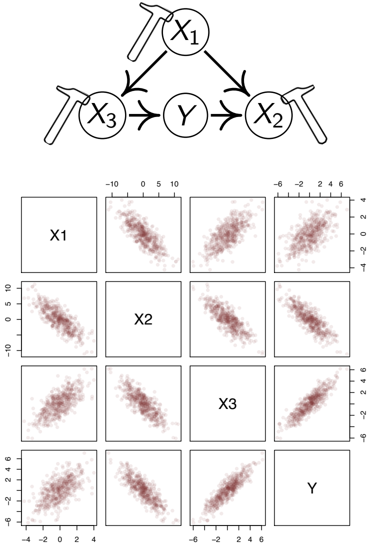
Idea 3: invariant causal prediction



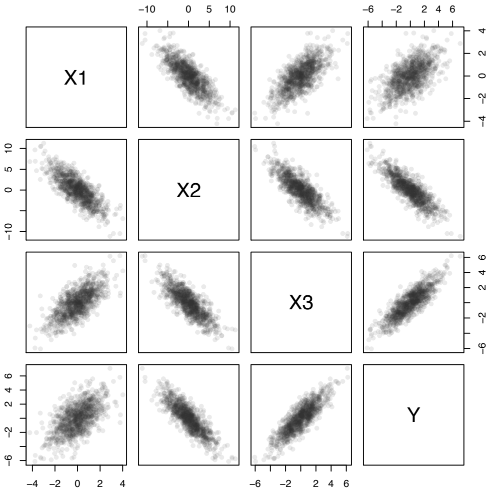
Problem:

- Find the causal parents of a target variable Y from $\hat{P}^n, \hat{Q}_1^n, \hat{Q}_2^n, \ldots$
- Confidence statements?


X_1
X_3
Y
X_2
X1
X2
X3
Y


X_1
X_3
Y
X_2
X1
X2
X3
Y

pooled data ($n = 1000$)



infer parents of Y from pooled data?

linear model

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

Call:
lm(formula = YY ~ XX)


Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.000322   0.025858   0.012     0.99
X1          -0.444534   0.034306 -12.958   <2e-16 ***
X2          -0.402398   0.016471 -24.430   <2e-16 ***
X3           0.603502   0.025642  23.536   <2e-16 ***
```

ICP (R-package `InvariantCausalPrediction`)

```
> ExpInd
```

[1]1111111111111111111111111111111...2222222222222...

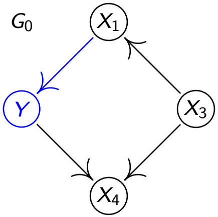
```
> icp <- ICP(X,Y,ExpInd)
```

```
             LOWER BOUND  UPPER BOUND  MAXIMIN EFFECT  P-VALUE
Variable_1       -0.11         0.10              0.00   1.0000
Variable_2       -0.33         0.00              0.00   1.0000
Variable_3        0.47         1.05              0.47   0.0012 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Key idea:
$P(Y \mid \mathbf{PA}_Y)$ remains invariant if the struct. equ. for Y does not change.

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

- N_i jointly independent
- G_0 has no cycles

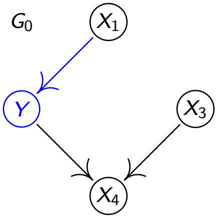


IMPORTANT: modularity, autonomy
Haavelmo 1944, Aldrich 1989, Pearl 2009, Schölkopf et al. 2012, Barenboim et al. 2013, Hauser et al. 2013, ...

Key idea:
$P(Y \mid \mathbf{PA}_Y)$ remains invariant if the struct. equ. for Y does not change.

$$
\begin{aligned}
X_1 &:= \tilde{f}_1(\tilde{N}_1)\\
Y &:= f_2(X_1, N_2)\\
X_3 &:= f_3(N_3)\\
X_4 &:= f_4(Y, X_3, N_4)
\end{aligned}
$$

- N_i jointly independent
- G_0 has no cycles



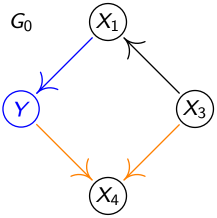
IMPORTANT: modularity, autonomy
Haavelmo 1944, Aldrich 1989, Pearl 2009, Schölkopf et al. 2012, Barenboim et al. 2013, Hauser et al. 2013, . . .

Key idea:

$P(Y \,|\, \mathbf{PA}_Y)$ remains invariant if the struct. equ. for Y does not change.

$$
\begin{aligned}
X_1 &:= f_1(X_3, N_1) \\
Y &:= f_2(X_1, N_2) \\
X_3 &:= f_3(N_3) \\
X_4 &:= \tilde{f}_4(Y, X_3, \tilde{N}_4)
\end{aligned}
$$

- N_i jointly independent
- G_0 has no cycles



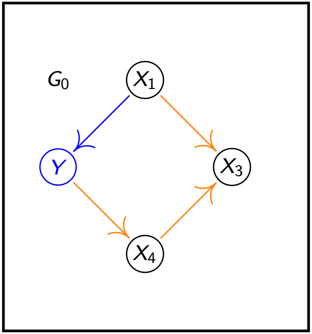
IMPORTANT: modularity, autonomy

Haavelmo 1944, Aldrich 1989, Pearl 2009, Schölkopf et al. 2012, Barenboim et al. 2013, Hauser et al. 2013, ...

Key idea:
$P(Y \mid \mathbf{PA}_Y)$ remains invariant if the struct. equ. for Y does not change.

$$
\begin{aligned}
X_1 &:= \tilde{f}_1(\tilde{N}_1)\\
Y &:= f_2(X_1, N_2)\\
X_3 &:= \tilde{f}_3(X_1, X_4, \tilde{N}_3)\\
X_4 &:= \tilde{f}_4(Y, \tilde{N}_4)
\end{aligned}
$$

- N_i jointly independent
- G_0 has no cycles



IMPORTANT: modularity, autonomy
Haavelmo 1944, Aldrich 1989, Pearl 2009, Schölkopf et al. 2012, Barenboim et al. 2013, Hauser et al. 2013, ...

Assumption

Let S^* be the indices of parents(Y).

$$\text{for all } e \in \mathcal{E}: \quad X^e \text{ has an arbitrary distribution} \;\; \text{and}$$
$$Y^e \,|\, X^e_{S^*} = x \quad \text{invariant}\,.$$

Assumption

Let S^* be the indices of parents(Y). There exists γ^* with support S^* that satisfies

$$\text{for all } e \in \mathcal{E}: \quad X^e \text{ has an arbitrary distribution and}$$

$$\cancel{Y^e \,|\, X^e_{S^*} = x \quad \text{invariant}.}$$

$$Y^e = X^e \gamma^* + \varepsilon^e, \quad \varepsilon^e \sim F_\varepsilon \text{ and } \varepsilon^e \perp\!\!\!\perp X^e_{S^*}.$$

Assumption

Let S^* be the indices of parents(Y). There exists γ^* with support S^* that satisfies

$$\text{for all } e \in \mathcal{E}: \quad X^e \text{ has an arbitrary distribution and}$$
$$\cancel{Y^e \,|\, X^e_{S^*} = x \quad \text{invariant}.}$$
$$Y^e = X^e \gamma^* + \varepsilon^e, \quad \varepsilon^e \sim F_\varepsilon \text{ and } \varepsilon^e \perp\!\!\!\perp X^e_{S^*}\,.$$

We say:

"S^* satisfies invariant prediction." or "$H_{0,S^*}(\mathcal{E})$ is true."

Assumption

Let S^* be the indices of parents(Y). There exists γ^* with support S^* that satisfies

$$\text{for all } e \in \mathcal{E}: \quad X^e \text{ has an arbitrary distribution and}$$
$$\cancel{Y^e \mid X^e_{S^*} = x \quad \text{invariant}.}$$
$$Y^e = X^e \gamma^* + \varepsilon^e, \quad \varepsilon^e \sim F_\varepsilon \text{ and } \varepsilon^e \perp\!\!\!\perp X^e_{S^*}.$$

We say:

"S^* satisfies invariant prediction." or "$H_{0,S^*}(\mathcal{E})$ is true."

Goal: Find S^*.

Given: Data from different environments $e \in \mathcal{E}$.

Assumption

Let S^* be the indices of parents(Y). There exists γ^* with support S^* that satisfies

$$\text{for all } e \in \mathcal{E}: \quad X^e \text{ has an arbitrary distribution} \quad \text{and}$$
$$\cancel{Y^e \mid X^e_{S^*} = x \quad \text{invariant}.}$$
$$Y^e = X^e\gamma^* + \varepsilon^e, \quad \varepsilon^e \sim F_\varepsilon \text{ and } \varepsilon^e \perp\!\!\!\perp X^e_{S^*}.$$

We say:

"S^* satisfies invariant prediction." or "$H_{0,S^*}(\mathcal{E})$ is true."

Goal: Find S^*.
Given: Data from different environments $e \in \mathcal{E}$.
Idea: Check $H_{0,S}(\mathcal{E})$ for several candidates S.

$$H_{0,S}(\mathcal{E}) = \begin{cases} \text{not rejected} \\ \text{rejected} \end{cases}$$

$$H_{0,S}(\mathcal{E}) = \begin{cases} \text{not rejected} \\ \text{rejected} \end{cases}$$

$$\hat{S}(\mathcal{E}) \; := \bigcap_{S \,:\, H_{0,S}(\mathcal{E}) \text{ not rej.}} S$$

$$H_{0,S}(\mathcal{E}) = \begin{cases} \text{not rejected} \\ \text{rejected} \end{cases}$$

$$\hat{S}(\mathcal{E}) := \bigcap_{S \,:\, H_{0,S}(\mathcal{E}) \text{ not rej.}} S$$

set	$\{3,5\}$	$\{3,7\}$	$S^* = \{1,3,6\}$	$\{2\}$	$\{3,8\}$	$\cdots$
inv. pred.	✓	✗	✓	✗	✓	$\cdots$

$\hat{S}(\mathcal{E}) = \{3\}$

$$H_{0,S}(\mathcal{E}) = \begin{cases} \text{not rejected} \\ \text{rejected} \end{cases}$$

$$\hat{S}(\mathcal{E}) := \bigcap_{S\,:\,H_{0,S}(\mathcal{E}) \text{ not rej.}} S$$

set	$\{3,5\}$	$\{3,7\}$	$S^* = \{1,3,6\}$	$\{2\}$	$\{3,8\}$	$\cdots$
inv. pred.	✓	✗	✓	✗	✓	$\cdots$

$$\hat{S}(\mathcal{E}) = \{3\}$$

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

infinite data P

finite data $\hat{P}_n$

$$H_{0,S}(\mathcal{E}) = \begin{cases} \text{correct} \\ \text{false} \end{cases}$$

$$H_{0,S}(\mathcal{E}) = \begin{cases} \text{not rejected} \\ \text{rejected} \end{cases}$$

$$S(\mathcal{E}) \;:= \bigcap_{S \,:\, H_{0,S}(\mathcal{E}) \text{ is true}} S$$

$$\hat{S}(\mathcal{E}) \;:= \bigcap_{S \,:\, H_{0,S}(\mathcal{E}) \text{ not rej.}} S$$

set	$\{3,5\}$	$\{3,7\}$	$S^* = \{1,3,6\}$	$\{2\}$	$\{3,8\}$	$\cdots$
inv. pred.	✓	✗	✓	✗	✓	$\cdots$

$S(\mathcal{E}) = \{3\}$

$\hat{S}(\mathcal{E}) = \{3\}$

$S(\mathcal{E}) \subseteq S^*$

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

Theorem (PBM 2016)

- *No mistakes:*

$$S(\mathcal{E}) \subseteq S^* \quad \text{and} \quad P(\hat{S}(\mathcal{E}) \subseteq S^*) \geq 1 - \alpha\,.$$

Theorem (PBM 2016)

- *No mistakes:*

$$S(\mathcal{E}) \subseteq S^* \quad \textit{and} \quad P(\hat{S}(\mathcal{E}) \subseteq S^*) \geq 1 - \alpha \,.$$

- *Seeing more environments helps:*

$$S(\mathcal{E}_1) \subseteq S(\mathcal{E}_2) \subseteq S^* \quad \textit{if} \quad \mathcal{E}_1 \subseteq \mathcal{E}_2$$

Theorem (PBM 2016)

- *No mistakes:*

$$S(\mathcal{E}) \subseteq S^* \quad \textit{and} \quad P(\hat{S}(\mathcal{E}) \subseteq S^*) \geq 1 - \alpha\,.$$

- *Seeing more environments helps:*

$$S(\mathcal{E}_1) \subseteq S(\mathcal{E}_2) \subseteq S^* \quad \textit{if} \quad \mathcal{E}_1 \subseteq \mathcal{E}_2$$

- *Sufficient conditions for $S(\mathcal{E}) = S^*$ exist.*

Theorem (PBM 2016)

- *No mistakes:*

$$S(\mathcal{E}) \subseteq S^* \quad \textit{and} \quad P(\hat{S}(\mathcal{E}) \subseteq S^*) \geq 1 - \alpha\,.$$

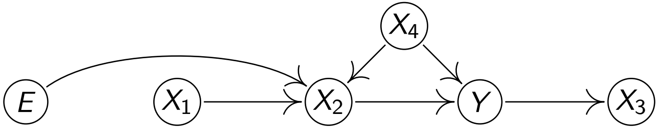
- *Seeing more environments helps:*

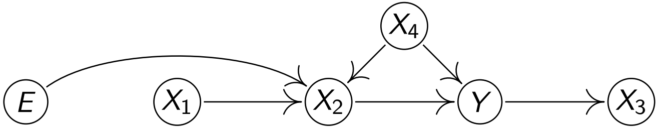
$$S(\mathcal{E}_1) \subseteq S(\mathcal{E}_2) \subseteq S^* \quad \textit{if} \quad \mathcal{E}_1 \subseteq \mathcal{E}_2$$

- *Sufficient conditions for* $S(\mathcal{E}) = S^*$ *exist.*

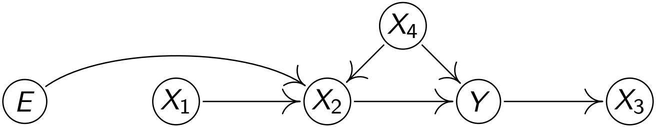
Identifiability improves if we have more and stronger interventions, at better places, more heterogeneity in the data.

JP, P. Bühlmann, N. Meinshausen: *Causal inference using invariant prediction: conf. interv.*, JRSS-B 2016.

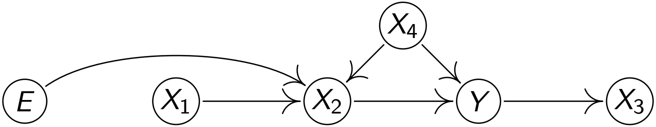

X_4
E
X_1
X_2
Y
X_3



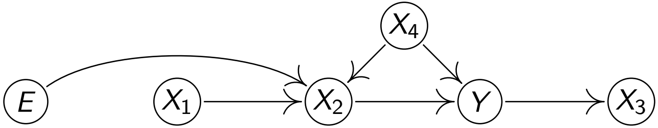

```
> Y <- X[,2] + X[,4] + noise
> ICP(X,Y,ExpInd)
```




```
> Y <- X[,2] + X[,4] + noise
> ICP(X,Y,ExpInd)

accepted set of variables: 2,4
accepted set of variables: 1,2,4
accepted set of variables: 2,3,4
accepted set of variables: 1,2,3,4

     LOWER BOUND  UPPER BOUND  MAXIMIN EFFECT  P-VALUE
X1        -0.03         0.01            0.00     0.48
X2         0.98         1.01            0.98  < 1e-09 ***
X3        -0.07         0.00            0.00     0.48
X4         0.95         1.01            0.95  2.6e-05 ***
```



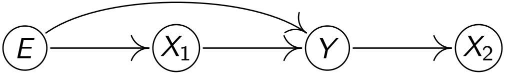

```
> Y <- X[,2]^2 + X[,4] + noise
> ICP(X,Y,ExpInd)
```



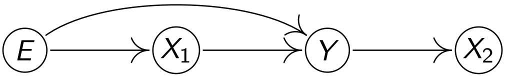

```
> Y <- X[,2]^2 + X[,4] + noise
> ICP(X,Y,ExpInd)

  empty set
  (all models rejected)
```

Model violation: nonlinear models

$\rightsquigarrow$ usually leads to loss of power, not coverage



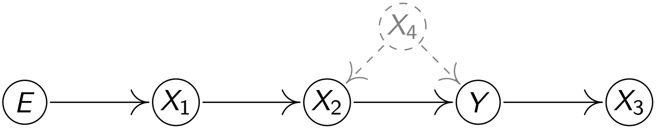

```
> Y <- X[,1] + E + noise
> ICP(X,Y,ExpInd)
```



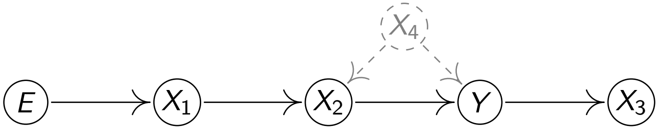

```
> Y <- X[,1] + E + noise
> ICP(X,Y,ExpInd)

  empty set
  (all models rejected)
```

Model violation: intervention on Y

$\rightsquigarrow$ usually leads to loss of power, not coverage



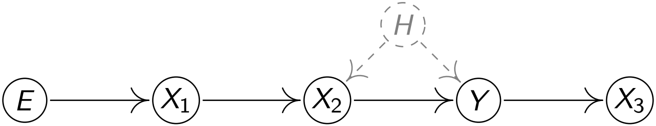

```
> Y <- X[,2] + X[,4] + noise
> ICP(X[,1:3],Y,ExpInd)
```




```
> Y <- X[,2] + X[,4] + noise
> ICP(X[,1:3],Y,ExpInd)

accepted set of variables: 1
accepted set of variables: 1,2
accepted set of variables: 1,3
accepted set of variables: 1,2,3

     LOWER BOUND  UPPER BOUND  MAXIMIN EFFECT  P-VALUE
X1        -0.87         1.05            0.00   <1e-09 ***
X2         0.00         1.86            0.00     1.00
X3        -1.61         0.00            0.00     0.73
```

Model violation: hidden variables

$\rightsquigarrow$ coverage still holds if we consider ancestors instead of parents



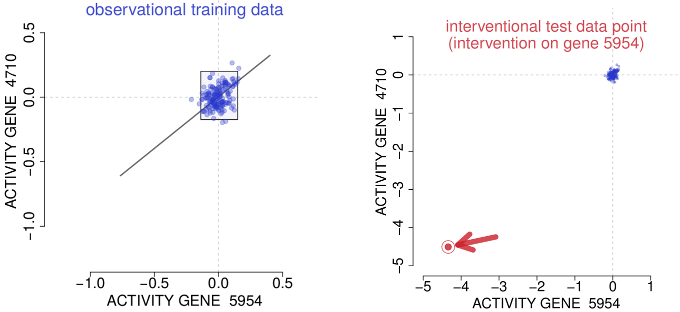
Theorem (PBM 2016)

Assume that the joint distribution over $(Y, X_1, \ldots, X_p, H_1, \ldots, H_q, E)$ *is faithful w.r.t. the augmented graph. Then*

$$S(\mathcal{E}) := \bigcap_{S \,:\, H_{0,S}(\mathcal{E}) \text{ is true}} S \;\subseteq\; \mathbf{AN}(Y) \cap \{X_1, \ldots, X_p\}\,.$$

Real data: genetic perturbation experiments for yeast (Kemmeren et al., 2014)

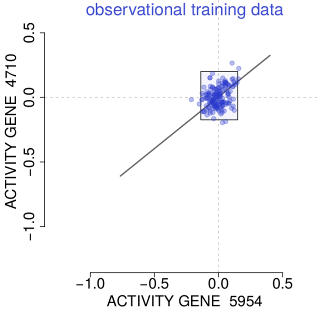
- $p = 6170$ genes
- $n_{obs} = 160$ wild-types
- $n_{int} = 1479$ gene deletions (targets known)

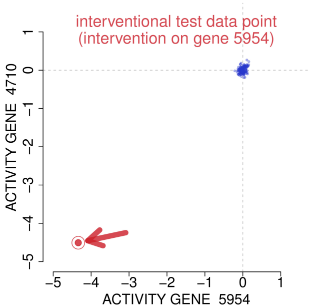


- true hits: $\approx 0.1\%$ of pairs

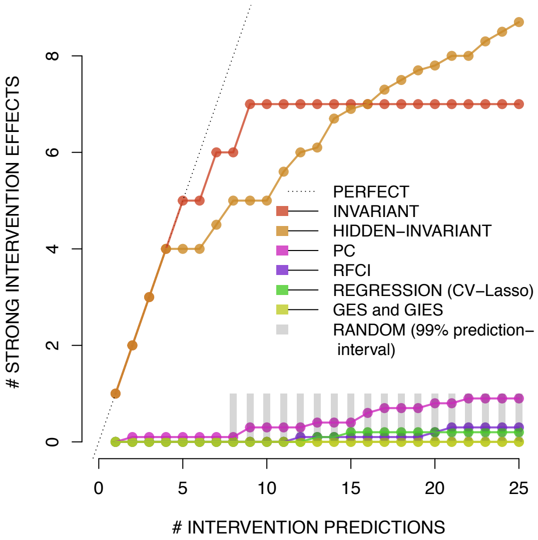
Real data: genetic perturbation experiments for yeast (Kemmeren et al., 2014)

- $p = 6170$ genes
- $n_{obs} = 160$ wild-types
- $n_{int} = 1479$ gene deletions (targets known)

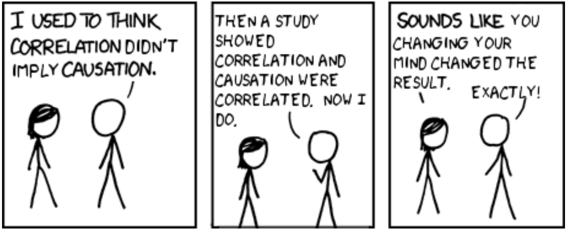




- true hits: $\approx 0.1\%$ of pairs
- our method: $\mathcal{E} = \{obs, int\}$

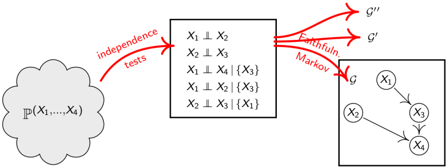


STRONG INTERVENTION EFFECTS
0
2
4
6
8
0
5
10
15
20
25
INTERVENTION PREDICTIONS
PERFECT
INVARIANT
HIDDEN−INVARIANT
PC
RFCI
REGRESSION (CV−Lasso)
GES and GIES
RANDOM (99% prediction−
interval)


I USED TO THINK CORRELATION DIDN'T IMPLY CAUSATION.
THEN A STUDY SHOWED CORRELATION AND CAUSATION WERE CORRELATED. NOW I DO.
SOUNDS LIKE YOU CHANGING YOUR MIND CHANGED THE RESULT.
EXACTLY!

Summary Part II:

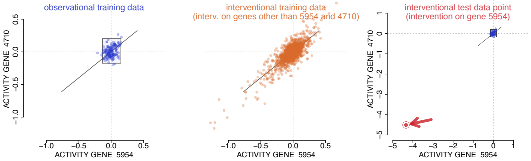
- Idea 1: independence-based methods (single environment)



- Idea 2: additive noise (single environment)

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

- Idea 3: invariant prediction (the more heterogeneity the better!)



Open Questions

- Causal Basics: What is a good definition of causal strength?
- Restricted SEMs: do we still have identifiability of causal structures if there are hidden variables?
- Real data: can we solve practically relevant problems?
- Causality and Machine Learning: do causal ideas help for "classical" tasks in machine learning?

Open Questions

- Causal Basics: What is a good definition of causal strength?
- Restricted SEMs: do we still have identifiability of causal structures if there are hidden variables?
- Real data: can we solve practically relevant problems?
- Causality and Machine Learning: do causal ideas help for "classical" tasks in machine learning?

General References

- Pearl: Causality.
- Spirtes, Glymour, Scheines: Causation, Prediction and Search.
- Peters: Causality (Script - see homepage)

Dankeschön!!

Part III: Applications to Machine Learning

Idea 1: semi-supervised learning

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

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

Idea 1: 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 1: 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:

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

Idea 1: semi-supervised learning

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

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

Modularity suggests:

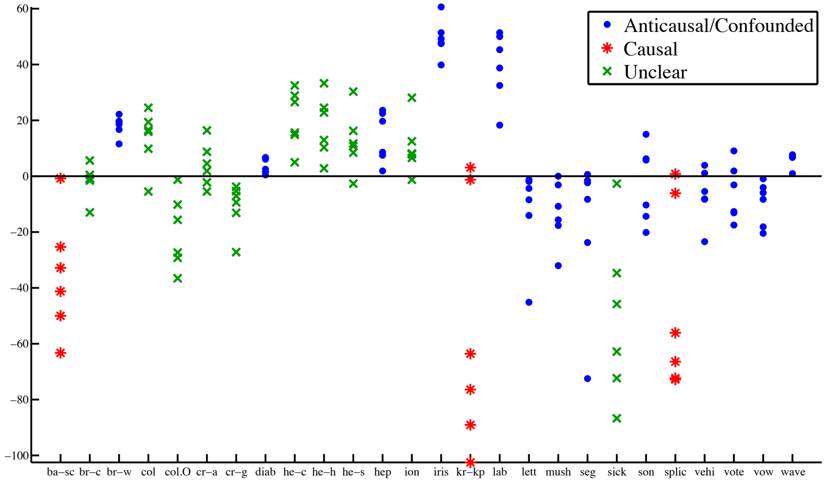
$$p(x_1 \,|\, x_{pa(1)}), \dots, p(x_d \,|\, x_{pa(d)}) \text{ are “independent”}$$

Special case:

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

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

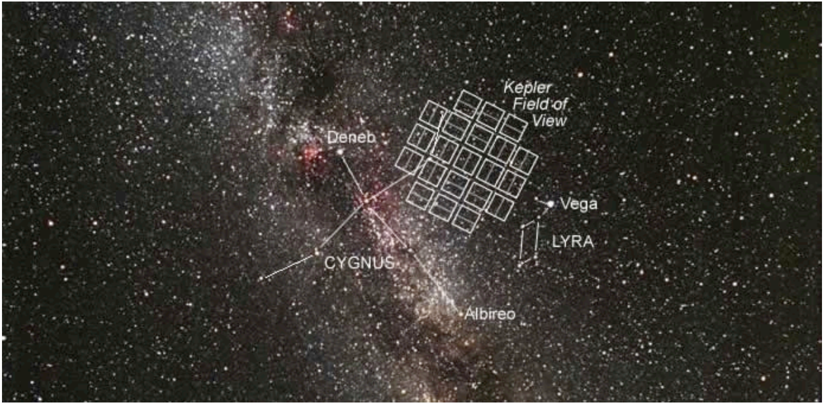
Idea 1: semi-supervised learning



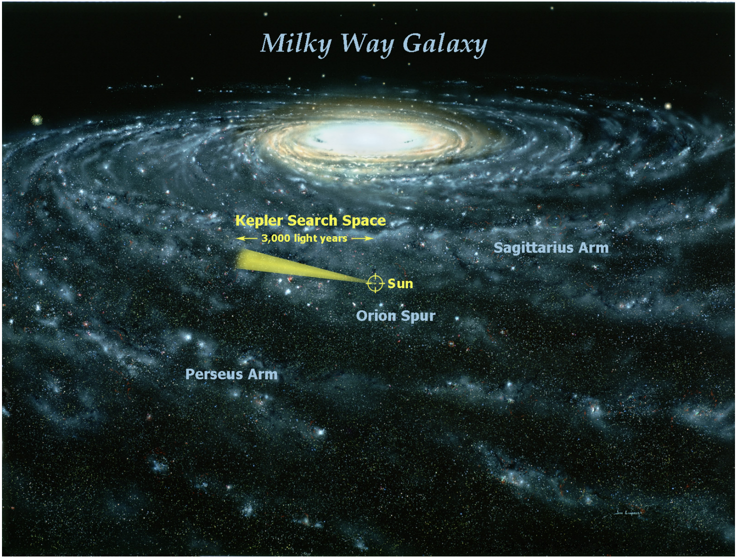
comparison to base classifier

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

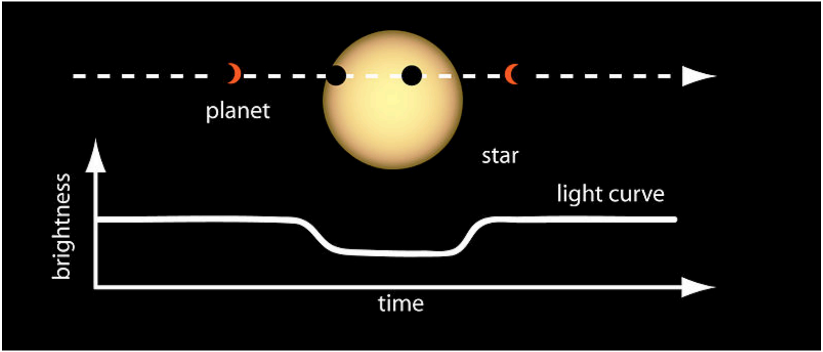
Idea 2: half-sibling regression



Idea 2: half-sibling regression



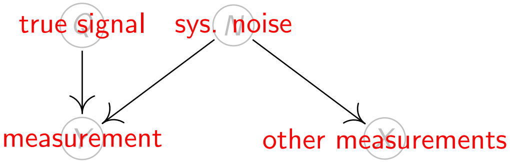
Idea 2: half-sibling regression



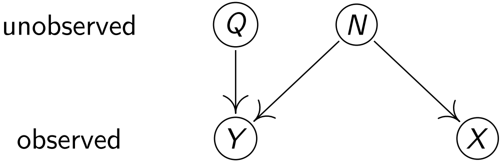
Idea 2: half-sibling regression

unobserved

observed

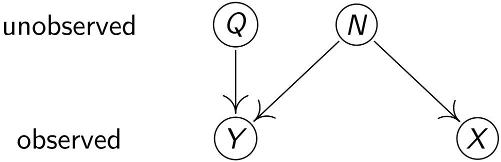


Idea 2: half-sibling regression



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

Idea 2: half-sibling regression

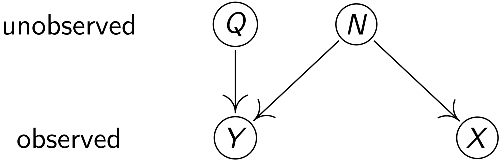


Assume $Y = f(N) + Q$.

Proposed idea:

Remove everything from Y explained by X.

Idea 2: half-sibling regression



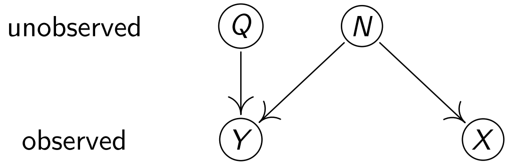
Assume $Y = f(N) + Q$.

Proposed idea:

Remove everything from Y explained by X.

Or: $\hat{Q} := Y - \mathbf{E}[Y \,|\, X]$.

Idea 2: half-sibling regression



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

Proposed idea:

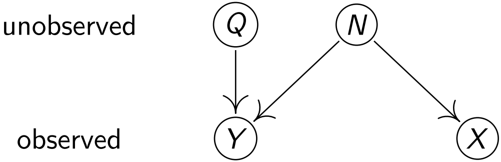
Remove everything from Y explained by X. Or: $\hat{Q} := Y - \mathbf{E}[Y \mid X]$.

Proposition

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

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

Idea 2: half-sibling regression



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

Proposed idea:

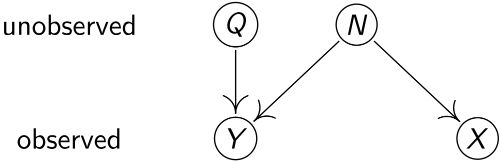
Remove everything from Y explained by X. Or: $\hat{Q} := Y - \mathbf{E}[Y \mid X]$.

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 2: half-sibling regression



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

Proposed idea:

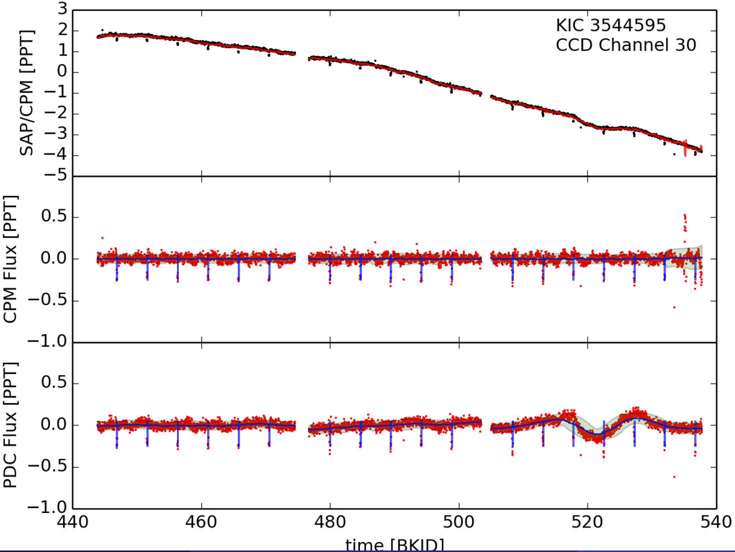
Remove everything from Y explained by X. Or: $\hat{Q} := Y - \mathbf{E}[Y \mid X]$.

Proposition

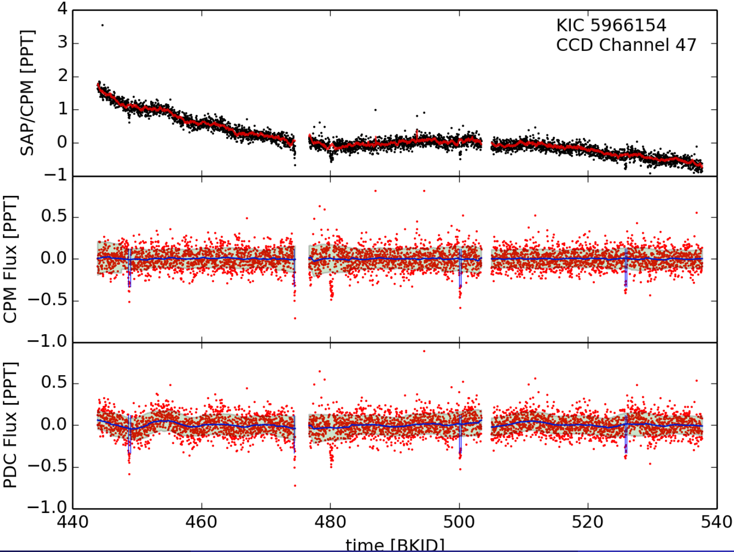
Convergence against "correct" signal Q (up to reparameterization) if

- *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:* $X_i = g_i(N) + R_i$, $i = 1, \ldots, \infty$

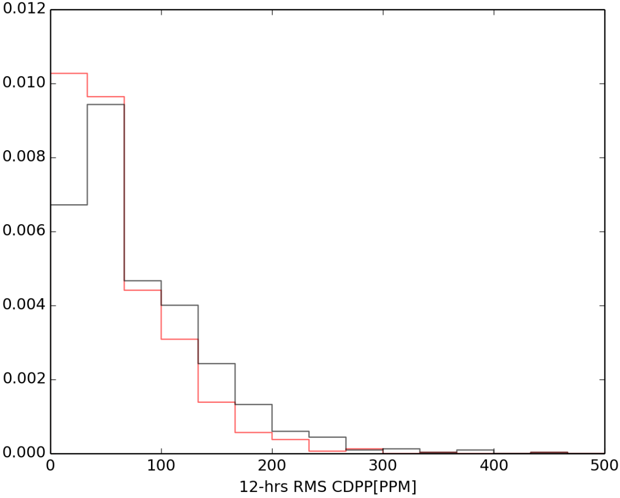
Idea 2: half-sibling regression



Idea 2: half-sibling regression



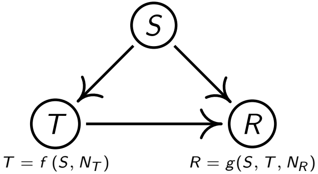
Idea 2: half-sibling regression



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

Idea 3: reinforcement learning

Recall the kidney stones:

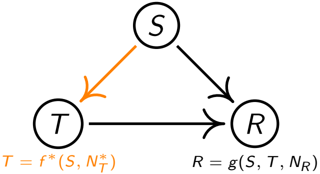


$$p(r, t, s) = p(r \mid t, s) \cdot \quad p(t \mid s) \quad \cdot p(s)$$

Question: What would happen if...?

Idea 3: Blackjack

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...?

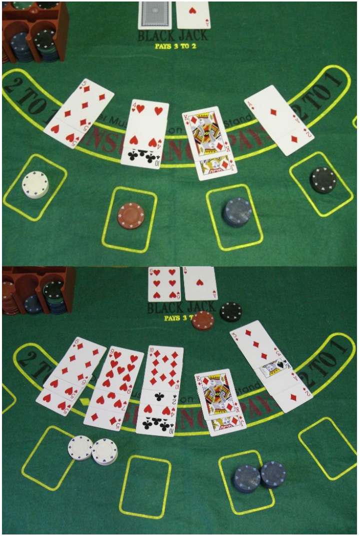
What is $\sup_{p^*} \mathbf{E}_{p^*} R$?

Idea 3: 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·bet.

Idea 3: Blackjack



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

Idea 3: 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).

Idea 3: 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 $\mid$ game state).

Questions:

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

Idea 3: 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 $\mathbf{E}_{p^*} \ell$?

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 3: Blackjack

Computation: Means

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

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

Idea 3: Blackjack

Computation: Means

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

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

Idea 3: Blackjack

Computation: Means

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

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

Estimate η by

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

Idea 3: Blackjack

Computation: Means

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

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

Estimate η by

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

Confidence intervals available!

Idea 3: Blackjack

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

Which p^* is best?

Idea 3: Blackjack

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

Which p^* is best? Parameterize and estimate

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

Idea 3: Blackjack

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

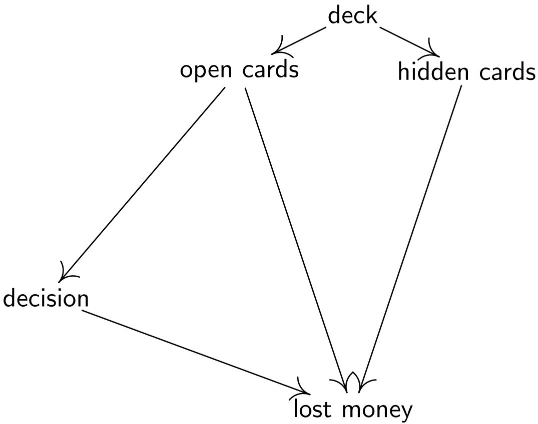
Which p^* is best? Parameterize and estimate

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

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, ...

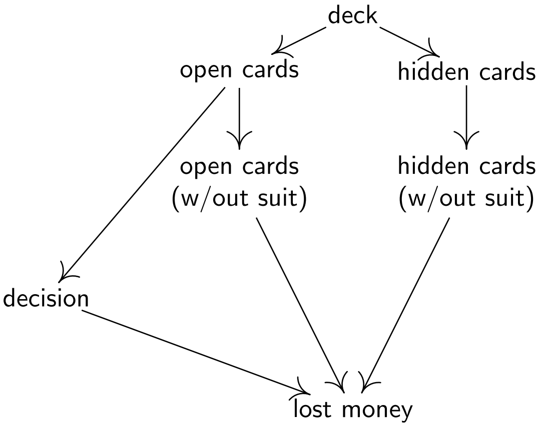
Idea 3: Blackjack

How to exploit causal structure:



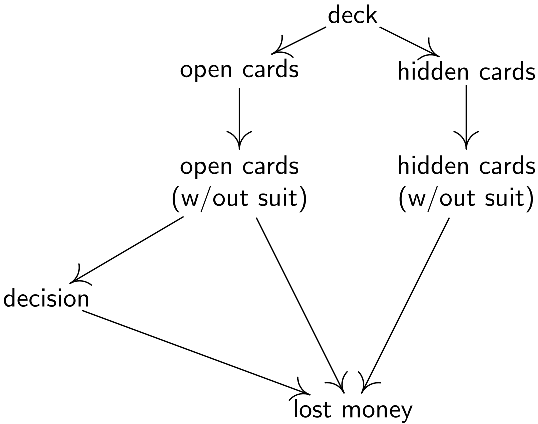
Idea 3: Blackjack

How to exploit causal structure:

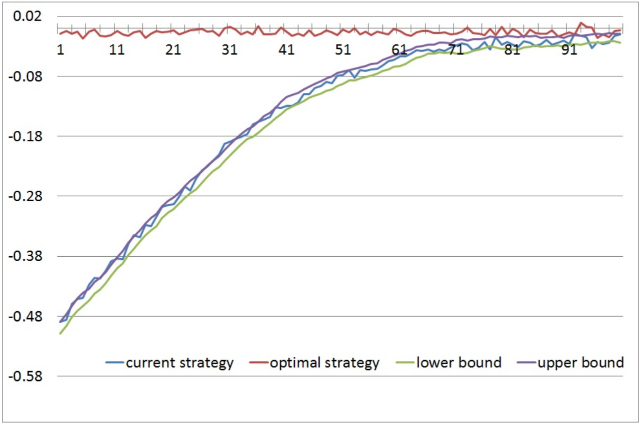


Idea 3: Blackjack

How to exploit causal structure:



Idea 3: Blackjack

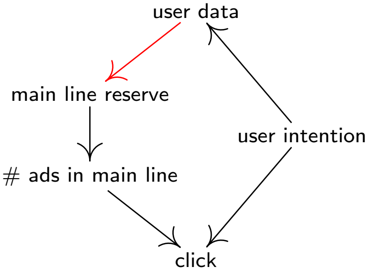


Idea 3: Blackjack

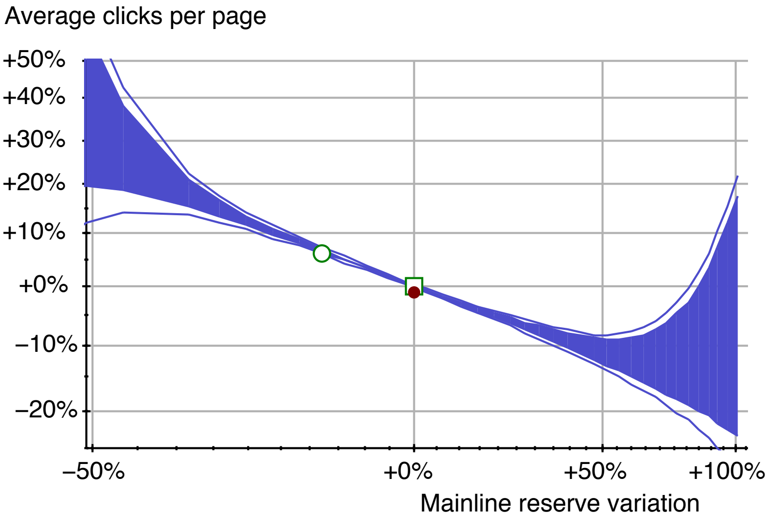
What can we do with $100,000$ samples?

	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

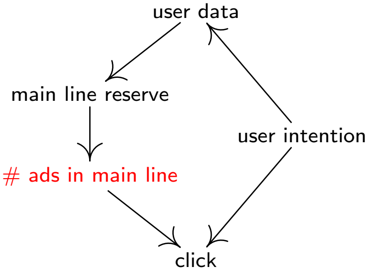
Idea 3: advertisement



Idea 3: advertisement

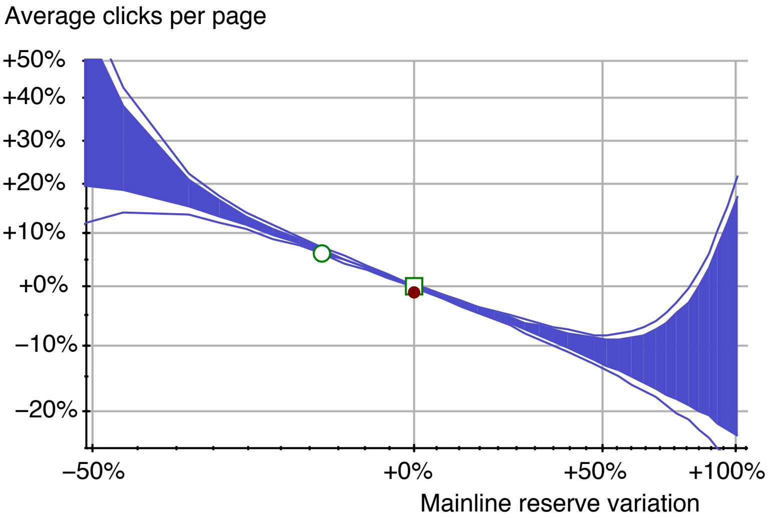


Idea 3: advertisement



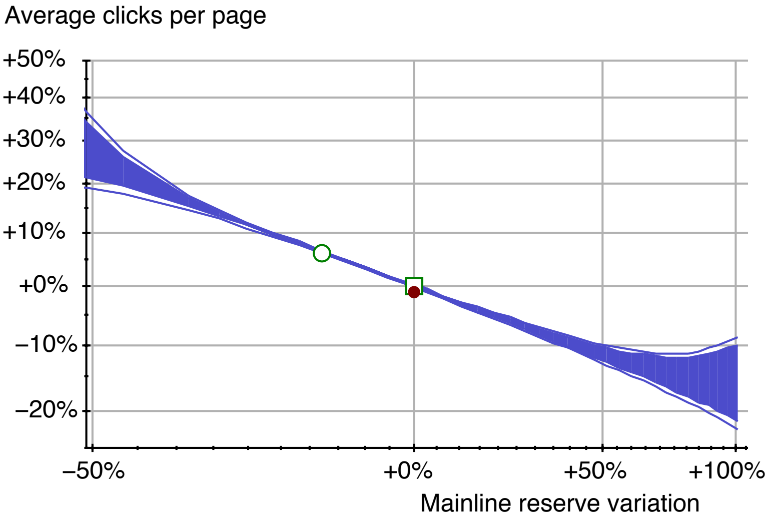
Idea 3: advertisement

Old:



Idea 3: advertisement

Using discrete variable (ads shown in mainline):



Idea 4: domain adaptation

method	training data from	test domain
transfer learning (TL)	$(\mathbf{X}^1, Y^1), \ldots, (\mathbf{X}^D, Y^D)$	$T := D + 1$
multi-task learning (MTL)	$(\mathbf{X}^1, Y^1), \ldots, (\mathbf{X}^D, Y^D)$	$T := D$

Idea 4: domain adaptation

method	training data from	test domain
transfer learning (TL)	$(\mathbf{X}^1, Y^1), \ldots, (\mathbf{X}^D, Y^D)$	$T := D + 1$
multi-task learning (MTL)	$(\mathbf{X}^1, Y^1), \ldots, (\mathbf{X}^D, Y^D)$	$T := D$

Invariant prediction for training:

$$Y^e | \mathbf{X}_S^e \stackrel{d}{=} Y^{e'} | \mathbf{X}_S^{e'} \qquad \text{for all } e \neq e' \in \{1, \ldots, D\}\,.$$

Invariant prediction in test domain T:

$$Y^e | \mathbf{X}_S^e \stackrel{d}{=} Y^T | \mathbf{X}_S^T \qquad \text{for all } e \in \{1, \ldots, D\}\,.$$

Idea 4: domain adaptation

method	training data from	test domain
transfer learning (TL)	$(\mathbf{X}^1, Y^1), \ldots, (\mathbf{X}^D, Y^D)$	$T := D + 1$
multi-task learning (MTL)	$(\mathbf{X}^1, Y^1), \ldots, (\mathbf{X}^D, Y^D)$	$T := D$

Invariant prediction for training:

$$Y^e | \mathbf{X}_S^e \stackrel{d}{=} Y^{e'} | \mathbf{X}_S^{e'} \qquad \text{for all } e \neq e' \in \{1, \ldots, D\}\,.$$

Invariant prediction in test domain T:

$$Y^e | \mathbf{X}_S^e \stackrel{d}{=} Y^T | \mathbf{X}_S^T \qquad \text{for all } e \in \{1, \ldots, D\}\,.$$

Assume for now S is known.

Idea 4: domain adaptation

Transfer learning (data in training but not in test domain):

$$f_S : \begin{array}{ccc} \mathcal{X} & \rightarrow & \mathcal{Y} \\ \mathbf{x} & \mapsto & \mathbf{E}\left[Y^1 \,|\, \mathbf{X}_S^1 = \mathbf{x}\right] \end{array} . \tag{1}$$

$\leadsto$ optimality in adversarial settings:

Idea 4: domain adaptation

Transfer learning (data in training but not in test domain):

$$f_S : \begin{array}{ccc} \mathcal{X} & \to & \mathcal{Y} \\ \mathbf{x} & \mapsto & \mathbf{E}\left[Y^1 \,|\, \mathbf{X}^1_S = \mathbf{x}\right] \end{array} . \tag{1}$$

$\rightsquigarrow$ optimality in adversarial settings:

Theorem

Consider D tasks $(\mathbf{X}^1, Y^1) \sim P^1, \ldots, (\mathbf{X}^D, Y^D) \sim P^D$ that satisfy invariant prediction in training. The estimator (1) satisfies

$$f_S \in \operatorname*{argmin}_{f \in C^0} \sup_{P^T \in \mathcal{P}} \mathbf{E}_{(\mathbf{X},Y)\sim P^T} \left(Y - f(\mathbf{X})\right)^2 ,$$

where $\mathcal{P}$ contains all distributions over $(\mathbf{X}, Y)$ that are absolutely continuous with respect to Lebesgue measure and that satisfy $Y \,|\, \mathbf{X} \overset{d}{=} Y^1 \,|\, \mathbf{X}^1$.

Idea 4: domain adaptation

Multi-task Learning - linear (data in training and test domain):

learn part of model in training domains

Idea 4: domain adaptation

Multi-task Learning - linear (data in training and test domain):

learn part of model in training domains

Theorem

Assume

$$Y^e = \alpha_S^t \mathbf{X}_S^e + \epsilon \qquad \textit{for } e \in \{1, \ldots, D\} \quad \textit{and}$$
$$\mathbf{X}_N^T = \alpha_N^T Y^T + \epsilon_N^T ,$$

where ϵ *and* ϵ_N^T *are jointly independent and* ϵ *is independent of* $\mathbf{X}_S$*. Then,*

$$\beta_N^T = \mathbb{E}(\epsilon^2) M^{-1} \alpha_N , \qquad \beta_S^T = \alpha_S \left(1 - (\alpha_N^T)^t \beta_N^T\right) - \Sigma_{X,S}^{-1} \Sigma_{X,N} \beta_N^T ,$$

where $M = \mathbb{E}(\epsilon^2) \alpha_S \alpha_S^t + \Sigma_N - \Sigma_{X,N} \Sigma_{X,S}^{-1} \Sigma_{X,N}$ *is LSE on the test domain.*

M. Rojas-Carulla, B. Schölkopf, R. Turner, JP: *A Causal Perspective on Domain Adaptation*, arXiv, 1507.05333

Idea 4: domain adaptation

What if S is unknown?

Idea 4: domain adaptation

What if S is unknown?
How to learn a good predictor from data

$$\beta^{inv} = \underset{\beta}{\text{argmin}} \underbrace{\sum_{e=1}^{D} \|R_\beta^e\|^2}_{\text{data fit}} + \lambda \cdot \underbrace{\ell(R_\beta^1, \ldots, R_\beta^D)}_{\text{invariance}},$$

with

- residuals $R_\beta^e := Y^e - \beta^t \mathbf{X}^e$ and
- $\ell(R_\beta^1, \ldots, R_\beta^D)$ penalizing different distributions of $R_\beta^1, \ldots, R_\beta^D$.

M. Rojas-Carulla, B. Schölkopf, R. Turner, JP: *A Causal Perspective on Domain Adaptation*, arXiv, 1507.05333

Summary Part III:

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

Summary Part III:

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

More details: (about all parts)

`http://people.tuebingen.mpg.de/jpeters/scriptChapter1-4.pdf`

Summary Part III:

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

More details: (about all parts)

`http://people.tuebingen.mpg.de/jpeters/scriptChapter1-4.pdf`

Dankeschön!