Causality

Jonas Peters
University of Copenhagen

Mini course on Causality, Cambridge MIT

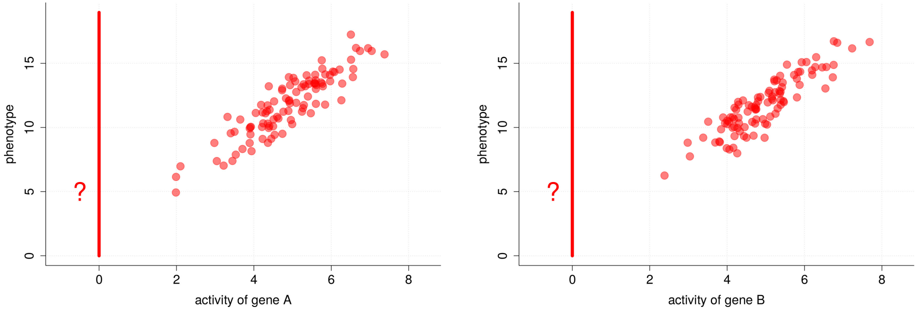
10th and 11th May 2017

UNIVERSITY OF COPENHAGEN

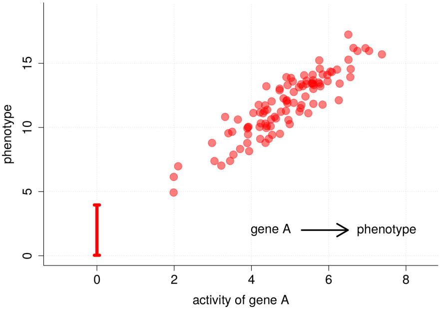
These slides contain ideas and concepts from...

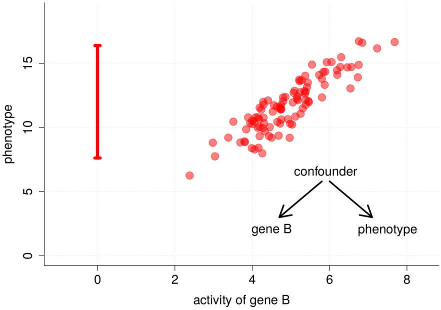
- UCLA: Judea Pearl
- CMU: Peter Spirtes, Clark Glymour, Richard Scheines, Kun Zhang
- Harvard University: Donald Rubin, Jamie Robins
- ETH Zürich: Peter Bühlmann, Nicolai Meinshausen, Stefan Bauer
- Max-Planck-Institute Tübingen: Dominik Janzing, Bernhard Schölkopf, Mateo Rojas-Carulla
- University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

Step 1: Consider the following problem.

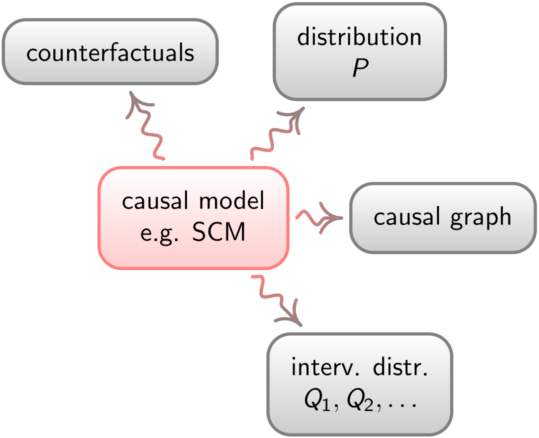


Step 2: Causality matters!



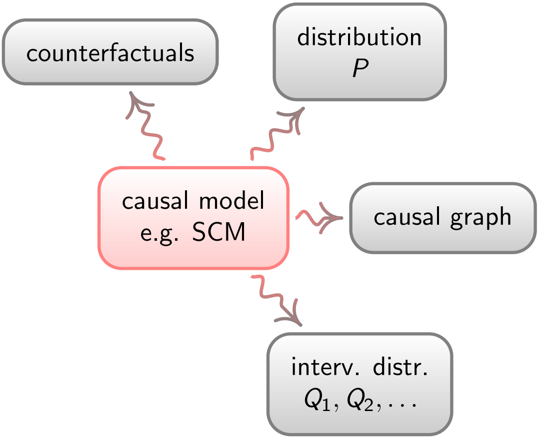


Step 3: What is a causal model?



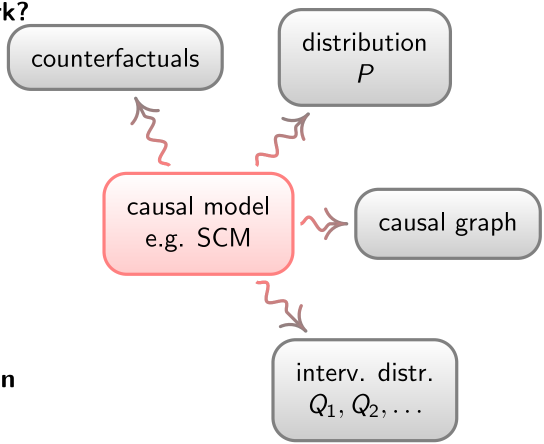
Step 4: What questions are being asked?

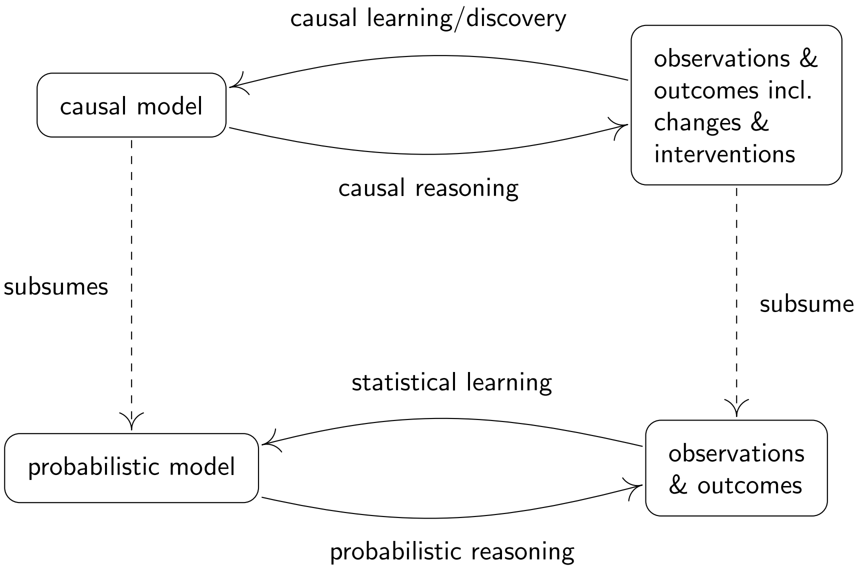
- How does the model work?
- What if there are hidden variables or feedback?
- What are nice graphical representations?
- Can we test counterfactual statements?
- Can we infer the graph structure from data?
- Is causality useful, even in classical ML/statistics settings?



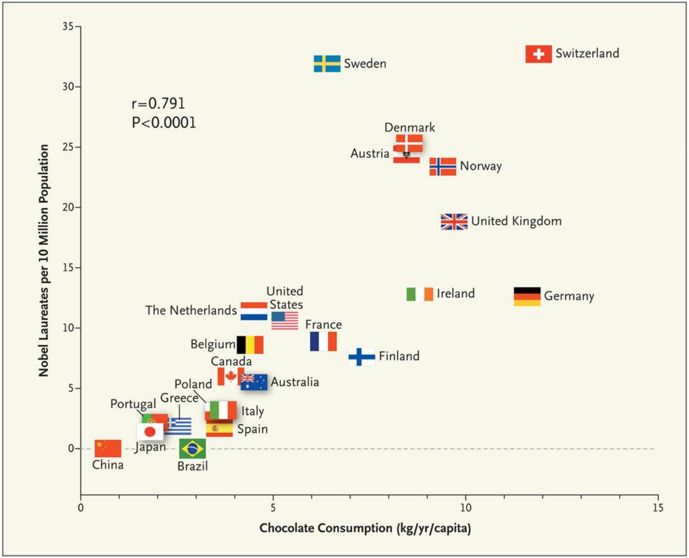
Step 4: What questions are being asked?

- **How does the model work?**
- What if there are hidden variables or feedback?
- What are nice graphical representations?
- Can we test counterfactual statements?
- **Can we infer the graph structure from data?**
- **Is causality useful, even in classical ML/statistics settings?**




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

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 |

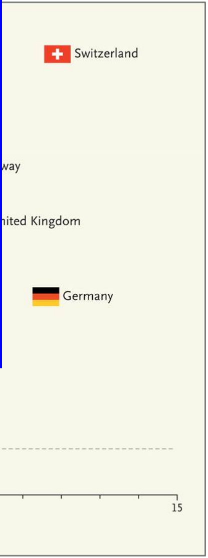
HEADLINES > REGULATION & SAFETY

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

Confectionery

HEADLINES | T

HEADLINES >

Subscribe to the Newsletter

Eating
winne

By Oliver Niebu

Related tags:
Sweden, candy

Forbes

New Posts
+10 posts this hour

Most Popular
Google's Driverless Car

Lis
Busi

+ Follow (73)

PHARMA & HEALTHCARE | 10/10/2012 @ 5:02PM | 14,700 views

Chocolate And Nobel Prize
In Study

4 comments, 2 called-out

+ Comment Now + Follow Comments

You don't have to be a genius to like chocolate, but geniuses are more likely to eat lots of chocolate, at least according to a new paper published in the August *New England Journal of Medicine*. Franz Messerli reports a highly

F. H. ... 12

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; $P<0{\cdot}001$
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; $0{\cdot}001<P<0{\cdot}01$ (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

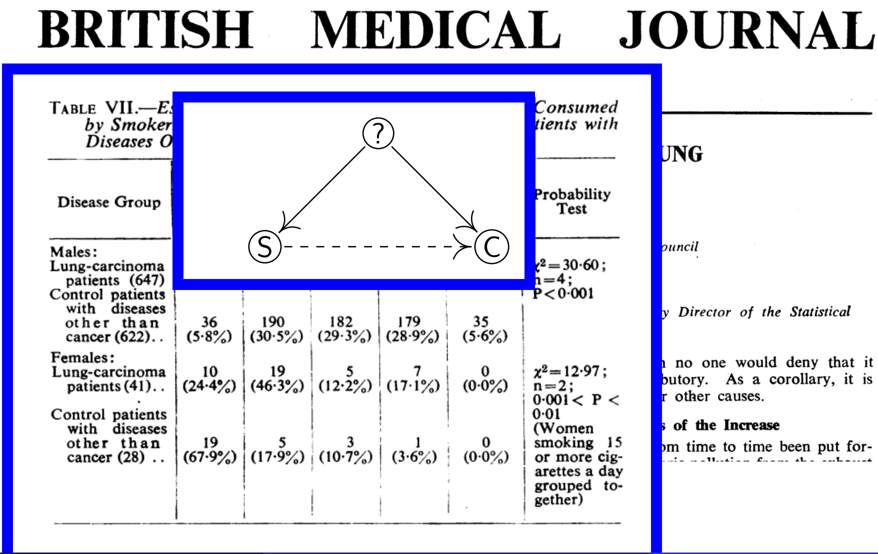
y Director of the Statistical

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.—Es… 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

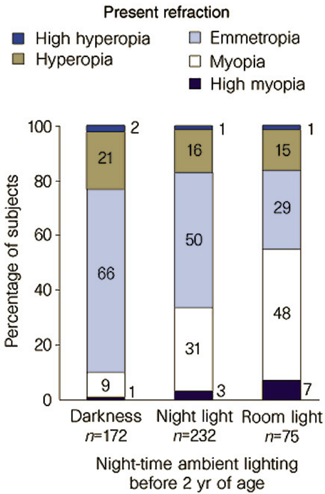
…y Director of the Statistical

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

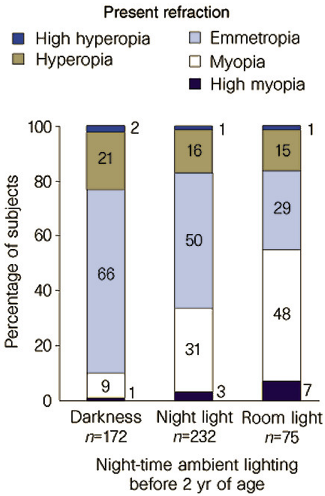
…s of the Increase

…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

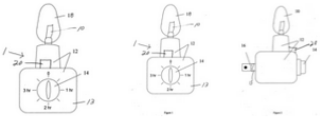
A timer a light and an optional music source is located on or in a housing of a nightlight assembly. When this assembly is plugged into a source of electric power, the timer is set to a selected time for the light and optional music to remain on. After this selected time has elapsed, the light and music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

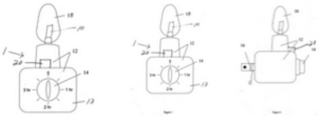
A timer a light and an optional music source is located on or in a housing of a nightlight assembly. When this assembly is plugged into a source of electric power, the timer is set to a selected time for the light and optional music to remain on. After this selected time has elapsed, the light and music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

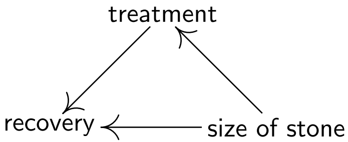
Example: kidney stones

	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

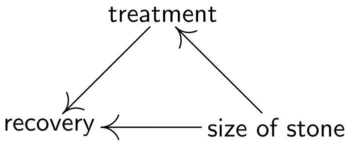
Example: kidney stones

underlying ground truth:



Example: kidney stones

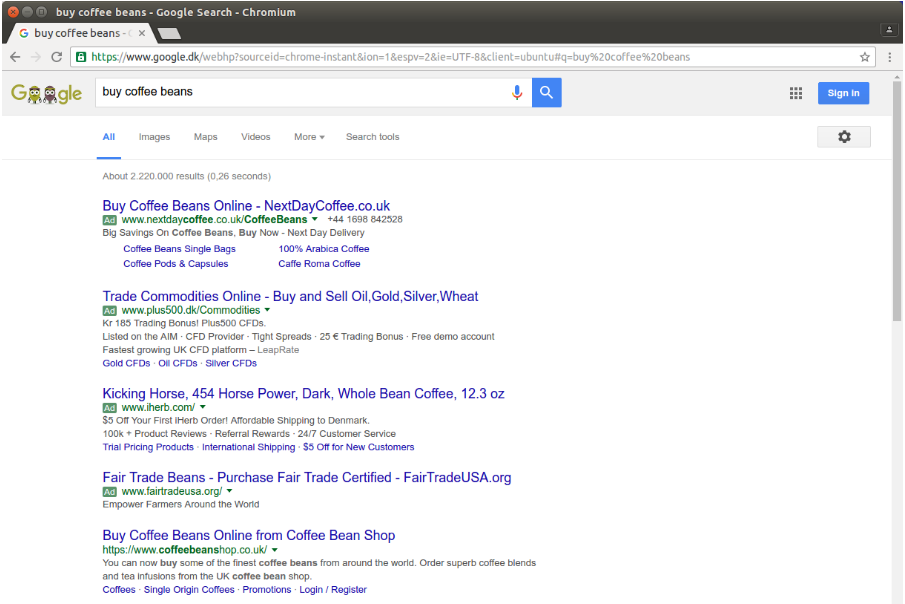
underlying ground truth:



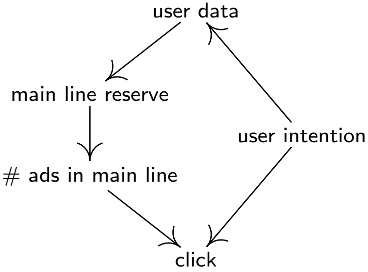
Question: What is the expected recovery if all get treatment B?

(Make treatment independent of size.)

Example: advertisement

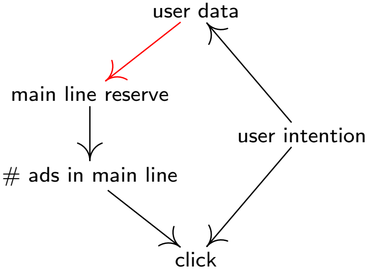


Example: advertisement



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

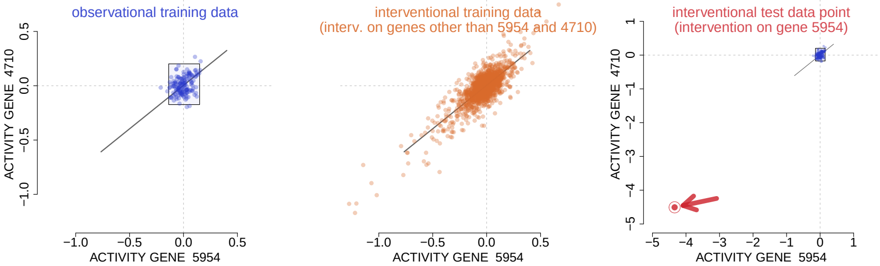
genetic perturbation experiments for yeast

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

Example: gene interactions

genetic perturbation experiments for yeast

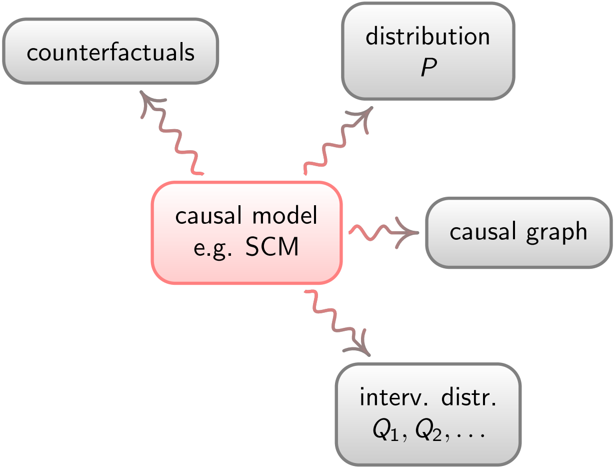
- $p = 6170$ genes
- $n_{obs} = 160$ wild-types
- $n_{int} = 1479$ gene deletions (targets known)



- 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

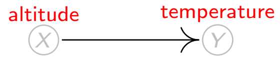

counterfactuals
distribution
P
causal model
e.g. SCM
causal graph
interv. distr.
$Q_1, Q_2, \ldots$

Example: Two variables

SCMs $(\mathbf{S}, P^{\mathbf{N}})$ model observational distributions.

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

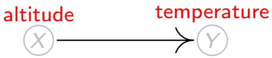
$$N_X, N_Y \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$$



Example: Two variables

SCMs $(\mathbf{S}, P^{\mathbf{N}})$ model observational distributions.

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



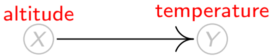
$$(X, Y) \sim \mathcal{N}\left(\left(\begin{array}{c} 0 \\ 0 \end{array}\right), \left(\begin{array}{cc} 1 & -6 \\ -6 & 37 \end{array}\right)\right)$$

Example: Two variables

SCMs $(\mathbf{S}, P^{\mathbf{N}})$ model interventions, too.

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

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



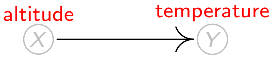
Example: Two variables

SCMs $(\mathbf{S}, P^{\mathbf{N}})$ model interventions, too.

$$X := N_X \quad X := 3$$

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

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



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

Example: Two variables

SCMs $(\mathbf{S}, P^{\mathbf{N}})$ 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 $(\mathbf{S}, P^{\mathbf{N}})$ model interventions, too.

$$X := N_X$$

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

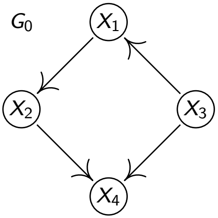
$$N_X, N_Y \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$$

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

SCMs $(\mathbf{S}, P^{\mathbf{N}})$: structural equations with noise distribution.

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

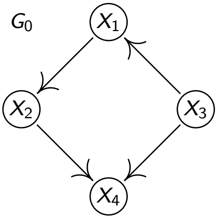
- N_i jointly independent
- G_0 has no cycles



SCMs $(\mathbf{S}, P^{\mathbf{N}})$ model observational distributions over $X_1, \ldots, X_d$. Call it P.

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

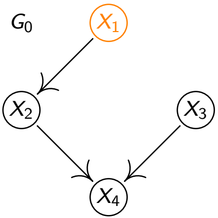
- N_i jointly independent
- G_0 has no cycles



SCMs $(\mathbf{S}, P^{\mathbf{N}})$ model interventions, too. Call it: $P_{do(X_1:=0)}$.

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

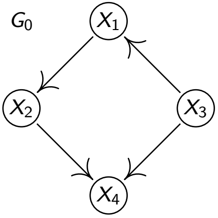
- N_i jointly independent
- G_0 has no cycles



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

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

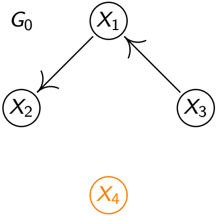
- N_i jointly independent
- G_0 has no cycles



SCMs model interventions, too. Call it $P_{do(X_4:=13)} \neq P(.\,|\,X_4 = 13)$

$$X_1 := f_1(X_3, N_1)$$
$$X_2 := f_2(X_1, N_2)$$
$$X_3 := f_3(N_3)$$
$$X_4 := 13$$

- N_i jointly independent
- G_0 has no cycles

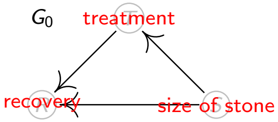


Example: kidney stones

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

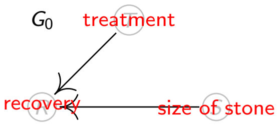


Example: kidney stones

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, ...

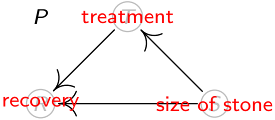
If you intervene only on X_j, you intervene only on X_j (MUTE).

Example: kidney stones

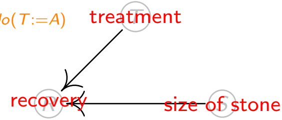
	Treatment A	Treatment B
Small Stones ($\frac{357}{700} = 0.51$)	$\frac{81}{87} = 0.93$	$\frac{234}{270} = 0.87$
Large Stones ($\frac{343}{700} = 0.49$)	$\frac{192}{263} = 0.73$	$\frac{55}{80} = 0.69$
	$\frac{273}{350} = 0.78$	$\frac{289}{350} = 0.83$
	$\frac{562}{700} = 0.80$	

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

P



$P_{do(T:=A)}$



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

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

Example: kidney stones

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

Example: kidney stones

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

This idea holds more generally.

Definition

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

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

This idea holds more generally.

Definition

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

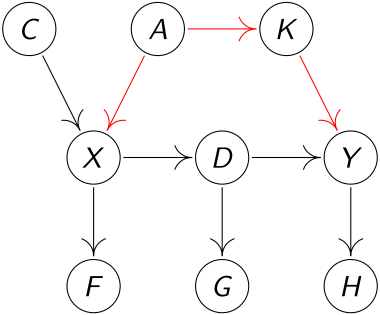
$$p_{do(X:=x)}(y) = \sum_{\mathbf{z}} p(y|x, \mathbf{z})p(\mathbf{z}) \neq p(y|x)$$

Proposition (Parent Adjustment)

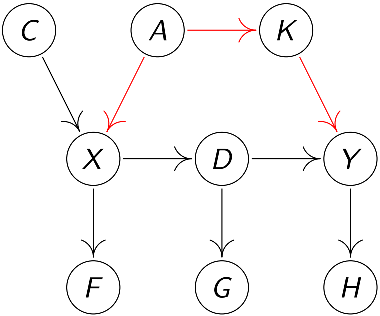
Assume $Y \notin PA(X)$. Then

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

Adjusting in Linear Gaussian Models



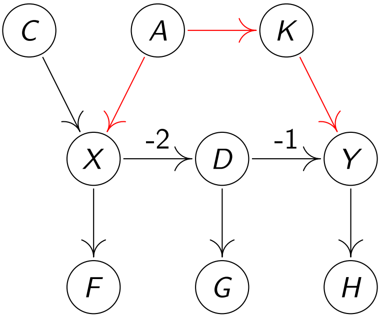
$X \leftarrow A \rightarrow K \rightarrow Y$ is a "backdoor path" from X to Y.



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

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

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



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

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

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

```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
```

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

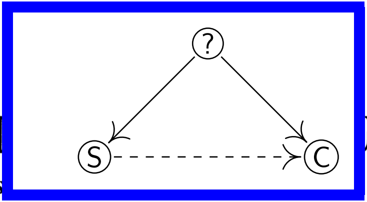
In England and Wales the phenomenal increase in the number of deaths attributed to cancer of the lung provides one of the most striking changes in the pattern of mortality recorded by the Registrar-General. For example, in the quarter of a century between 1922 and 1947 the annual number of deaths recorded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

LONDON S

SMOKING AND CARCINOMA OF THE LUNG

PRELIMINARY REPORT

BY

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

Member of the Statistical Research Unit of the Medical Research Council

AND

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

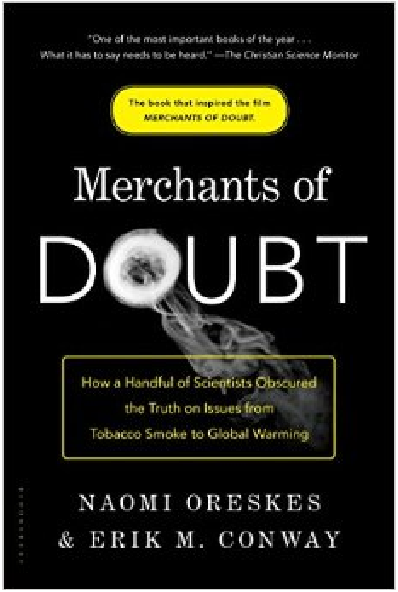
Professor of Medical Statistics, London School of Hygiene and Tropical Medicine; Honorary Director of the Statistical Research Unit of the Medical Research Council

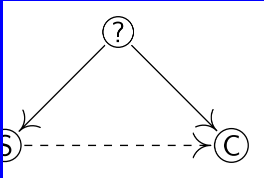
In England and Wales the phenomenal increase in the number of deaths attributed to cancer of the lung provides one of the most striking changes in the pattern of mortality recorded by the Registrar-General. For example, in the quarter of a century between 1922 and 1947 the annual number of deaths recorded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

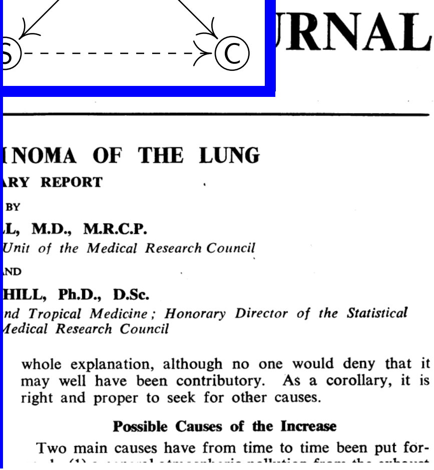
whole explanation, although no one would deny that it may well have been contributory. As a corollary, it is right and proper to seek for other causes.

Possible Causes of the Increase

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







URNAL

NOMA OF THE LUNG

RY REPORT

BY

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

Unit of the Medical Research Council

ND

HILL, Ph.D., D.Sc.

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

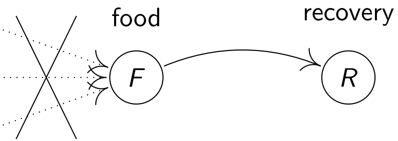
whole explanation, although no one would deny that it may well have been contributory. As a corollary, it is right and proper to seek for other causes.

Possible Causes of the Increase

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

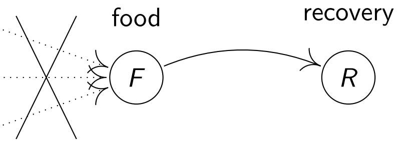
James Lind (1716–94):

James Lind (1716–94): Randomized Experiment



Then: $P_{do(F:=f)}(R) = P(R|F = f)$

James Lind (1716–94): Randomized Experiment



Then: $P_{do(F:=f)}(R) = P(R|F = f)$

"On the 20th of May 1747, I selected twelve patients in the scurvy, on board the Salisbury at sea. [...] 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 hospital surgeon [...] The consequence was, that the most sudden and visible good effects were perceived from the use of oranges and lemons; one of those who had taken them, being at the end of six days fit for duty."

Definition (Equivalence of causal models)

Two models are called

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

if they entail the same

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

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

Definition

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

Definition

Given an SCM, 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 SCM, 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 SCM, 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? :-)

Instrumental Variables?

Counterfactuals

Summary Part I:

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

Summary Part I:

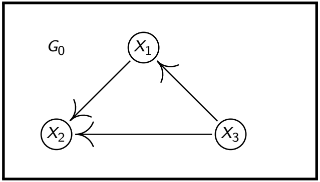
- What if interested in iid prediction, i.e. observational data? Don't worry (too much) about causality!
- But often, we are interested in a system's behaviour under intervention.

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

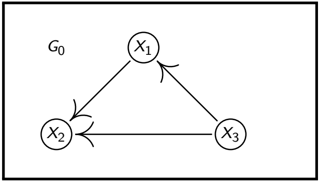


Summary Part I:

- 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



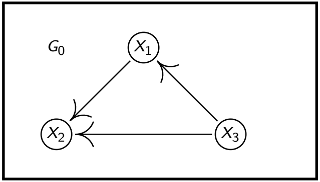
- graph + observational distribution $\rightsquigarrow$ interventions
- SCM + observational distribution $\rightsquigarrow$ counterfactuals

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



- graph + observational distribution $\rightsquigarrow$ interventions
- SCM + observational distribution $\rightsquigarrow$ counterfactuals
- Adjusting allows to compute interventions when there are (some) hidden variables

BMJ

BMJ 2011;343:d7679 doi: 10.1136/bmj.d7679 (Published 15 December 2011) Page 1 of 4

RESEARCH

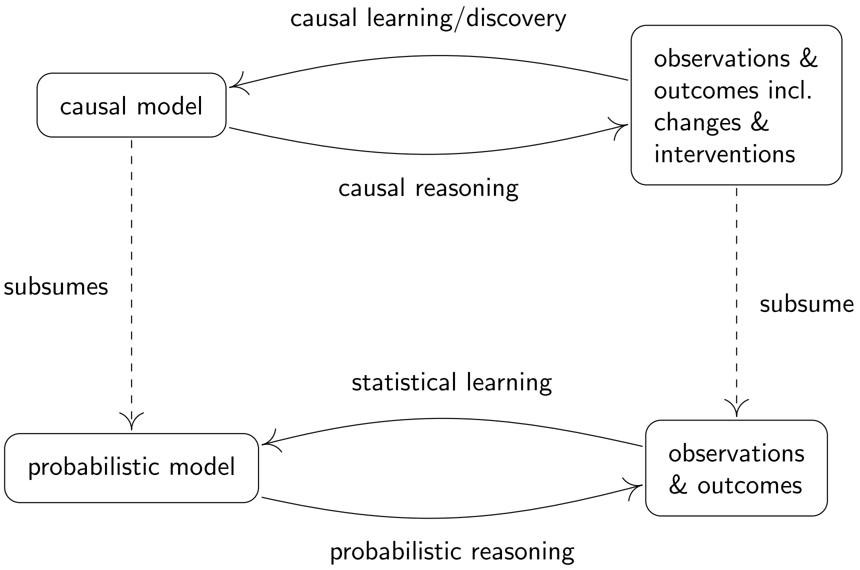
CHRISTMAS 2011: DEATH'S DOMINION

How fast does the Grim Reaper walk? Receiver operating characteristics curve analysis in healthy men aged 70 and over

Fiona F Stanaway *research fellow*[1], Danijela Gnjidic *research fellow*[2 3 4], Fiona M Blyth *deputy*

answer: $0.82 m/s$ (picture by Belle Mellor)

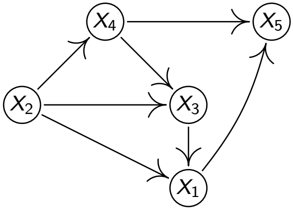
Part II: Causal Discovery

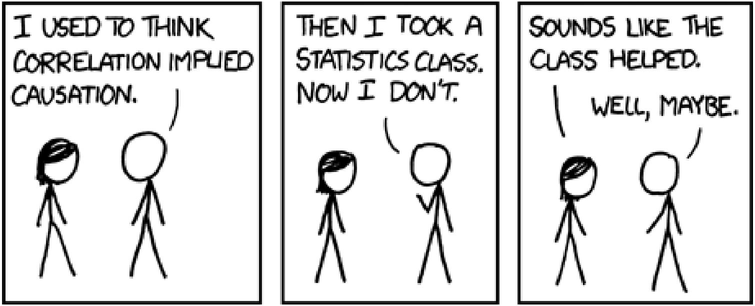

causal learning/discovery
causal model
observations & outcomes incl. changes & interventions
causal reasoning
subsumes
subsume
statistical learning
probabilistic model
observations & outcomes
probabilistic reasoning

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
⋮	⋮	⋮	⋮	⋮




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:

$$\begin{tikzcd}
\text{X: Cambridge (MA) } \heartsuit \arrow[dr] & & \text{Y: MIT } \heartsuit \arrow[dl] \\
& \text{Z: studying at MIT} &
\end{tikzcd}$$

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:

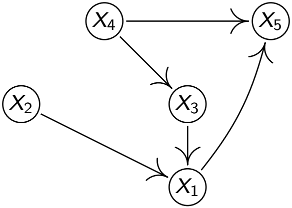
$$\begin{tikzcd}
\text{X: Cambridge (MA) } \heartsuit \arrow[dr] & & \text{Y: MIT } \heartsuit \arrow[dl] \\
& \text{Z: studying at MIT} &
\end{tikzcd}$$

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!

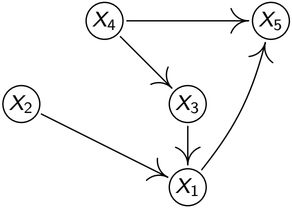
- parents, children, descendants, ancestors, ...
- paths, directed paths
- immoralities (or v-structures)
- *d*-separation (see next)
- ...



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 $\mathcal{S}$.

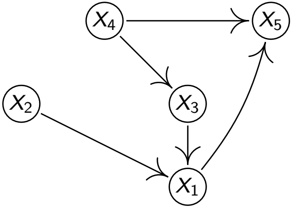
Check, whether all paths blocked!!



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.

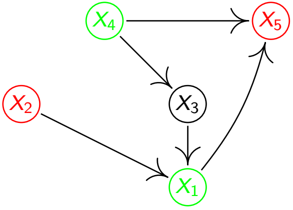
$\circ \cdots \leftarrow \circ \rightarrow \cdots \circ$ blocks a path.

$\circ \cdots \rightarrow \circ \leftarrow \cdots \circ$ 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!!



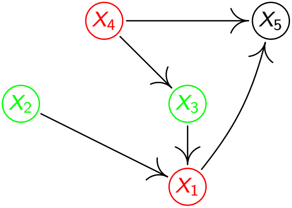
$\circ \cdots \rightarrow \circ \rightarrow \cdots \circ$ blocks a path.
$\circ \cdots \leftarrow \circ \rightarrow \cdots \circ$ blocks a path.
$\circ \cdots \rightarrow \circ \leftarrow \cdots \circ$ blocks a path.

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

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.

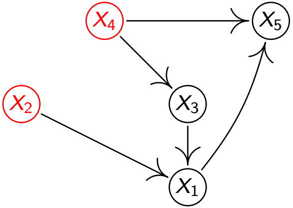
$\circ \cdots \rightarrow \circ \leftarrow \cdots \circ$ blocks a path.

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

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!!



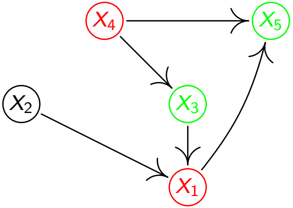
$\circ \cdots \to \circ \to \cdots \circ$	blocks a path.
$\circ \cdots \leftarrow \circ \to \cdots \circ$	blocks a path.
$\circ \cdots \to \circ \leftarrow \cdots \circ$	blocks a path.

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 $\{\}$

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.

There are three equivalent formulations of the Markov condition.

(i) **global Markov property**:

$$\mathbf{A} \text{ d-sep } \mathbf{B} \,|\, \mathbf{C} \text{ in } G \;\Rightarrow\; \mathbf{A} \perp\!\!\!\perp \mathbf{B} \,|\, \mathbf{C}$$

(ii) **local Markov property**: each variable is independent of its non-descendants given its parents, and

(iii) **Markov factorization property**:

$$p(\mathbf{x}) = p(x_1, \ldots, x_d) = \prod_{j=1}^{d} p(x_j \,|\, \mathbf{pa}_j^G).$$

(assume existence of density)

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

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

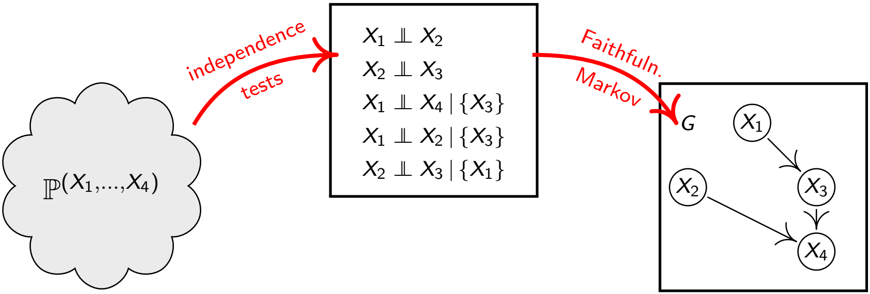
Definition

P is faithful w.r.t. G if

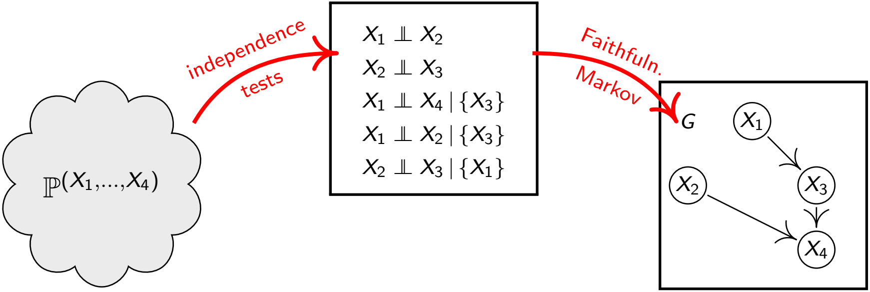
$$X_i \text{ and } X_j \text{ are } d\text{-separated by } \mathcal{S} \text{ in } G \quad \Leftarrow \quad X_i \perp\!\!\!\perp X_j \,|\, \mathcal{S}$$

Examples...

Idea 1: independence-based methods



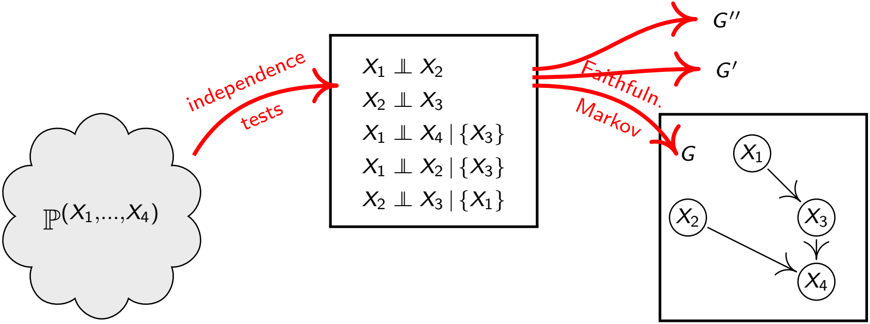
Idea 1: independence-based methods



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.

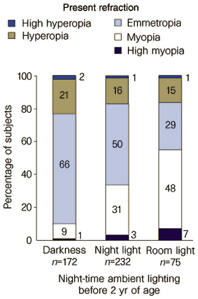
Idea 1: independence-based methods



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



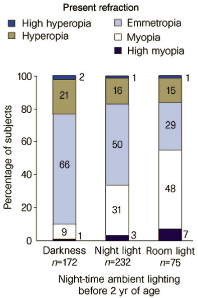
We have

- 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

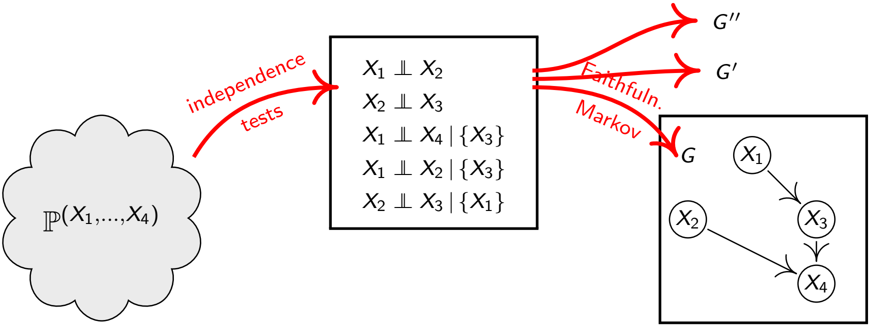
- night light $\not\perp\!\!\!\perp$ child myopia
- night light $\perp\!\!\!\perp$ child myopia | parent myopia
- no other independences

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 ...

$$
\begin{tikzcd}
 & PM \arrow[dl] \arrow[dr] & \\
NL & & CM
\end{tikzcd}
$$

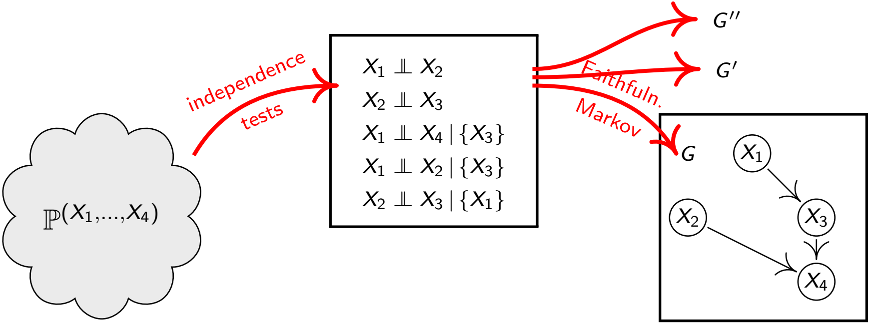
Idea 1: independence-based methods



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? (Nothing is possible in general.)

Idea 2: restricted structural causal models

Consider a distribution generated by

$$\boxed{\begin{array}{l} Y = \alpha X + N_Y \\ \text{with } N_Y, X \text{ ind.} \end{array}}$$

$$X \longrightarrow Y$$

Idea 2: restricted structural causal models

Consider a distribution generated by

$$\boxed{\begin{array}{l} Y = \alpha X + N_Y \\ \\ \text{with } N_Y, X \text{ ind.} \end{array}} \qquad X \longrightarrow Y$$

Then, if (X, N_Y) is non-Gaussian, there is no

$$\cancel{\boxed{\begin{array}{l} X = \beta Y + M_X \\ \\ \text{with } M_X, Y \text{ ind.} \end{array}} \qquad X \longleftarrow Y}$$

Shimizu et al. 2006

Idea 2: restricted structural causal models

Consider a distribution corresponding to

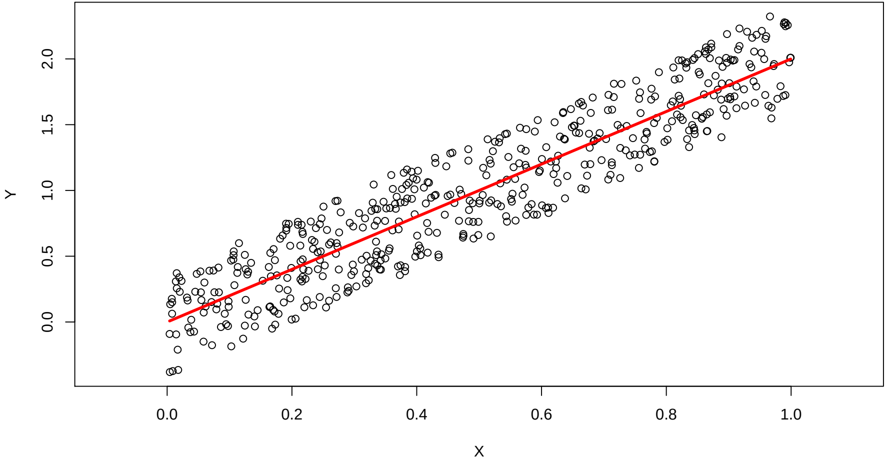
$$\boxed{\begin{array}{c} Y = 2X + N_Y \\[1em] \text{with } N_Y, X \overset{ind}{\sim} \mathcal{U} \end{array}}$$

$$X \longrightarrow Y$$

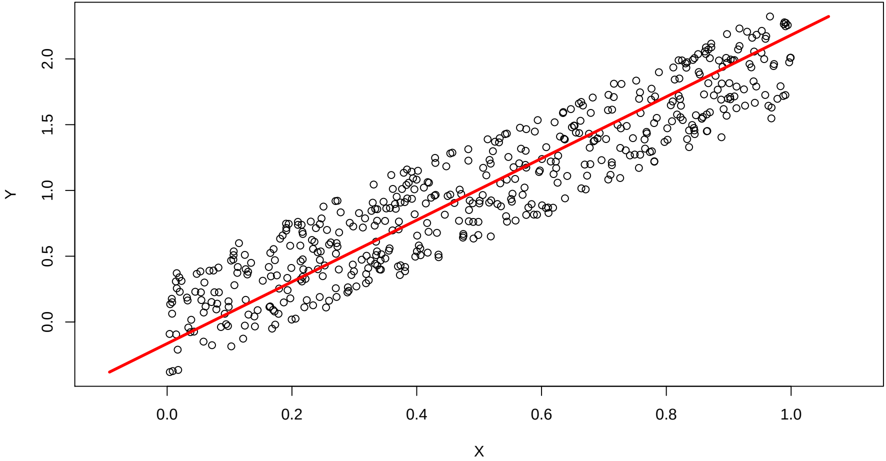
with

$$X \sim \mathcal{U}[-1, 1]$$
$$N_Y \sim \mathcal{U}[-0.4, 0.4]$$

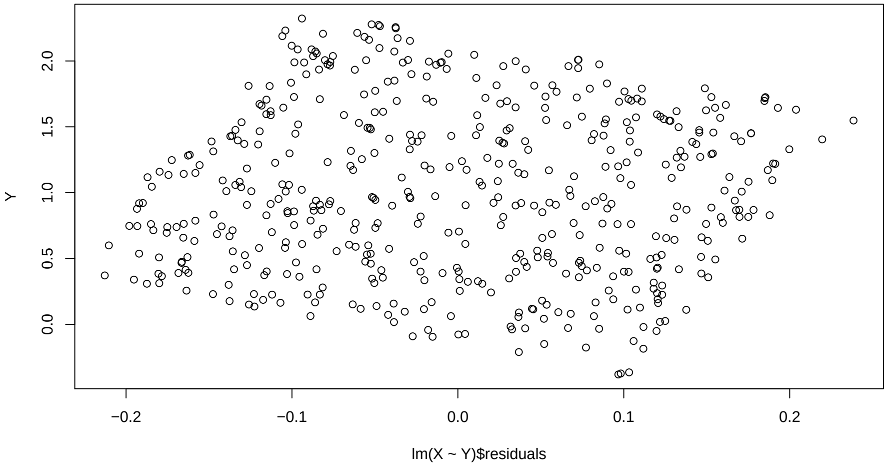
Idea 2: restricted structural causal models



Idea 2: restricted structural causal models



Idea 2: restricted structural causal models



Idea 2: restricted structural causal models

Method...

Idea 2: restricted structural causal models

Theory...

Idea 2: restricted SCMs – arrow of time

Peters et al ICML 2009 (univariate), Bauer et al ICML 2016 (multivariate)

Theorem

Let $(X_t)_t$ be a causal[a] *solution of an* $\mathrm{ARMA}(p, q)$ *process:*

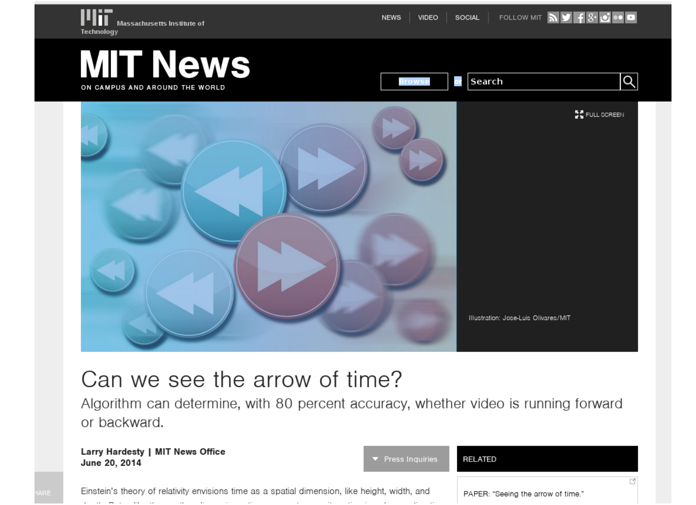
$$X_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \ldots + \theta_q Z_{t-q}.$$

Then, X_t *is time reversible, i.e., a causal solution of an* $\mathrm{ARMA}(\tilde{p}, \tilde{q})$ *process with reversed time, if and only if* $(Z_t)_t$ *is Gaussian.*

[a] $(X_t)_t$ causal iff $Z_t \perp\!\!\!\perp X_{t-k}$, $k > 0$.

Idea 2: restricted SCMs – arrow of time

Pickup et al. 2014:

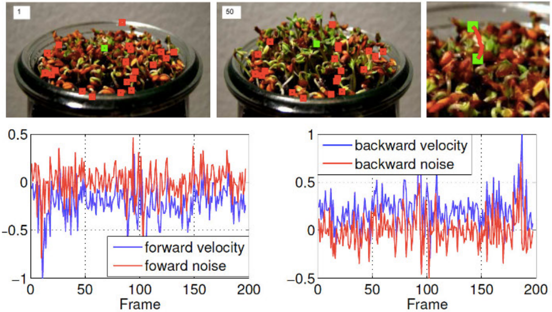


Idea 2: restricted SCMs – arrow of time

Pickup et al. 2014:

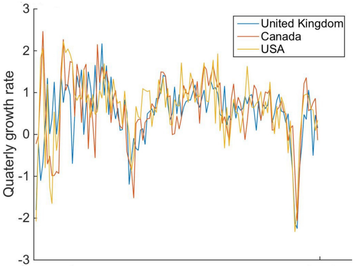
Method #3: Auto-regressive model

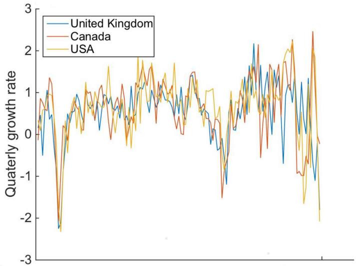
If object motion is linear, then the current velocity of the object should be affected only by the past. Noise on this motion will be asymmetric in the forward and backward directions, and fitting an auto-regressive model to the linear motion ought to yeild independence between the noise and signal only in the forwards-time direction. This method attempts to find the forward direction by looking at the independence of AR fitting error on motion trajectories.



Top: tracked points from a sequence, and an example track. Bottom: Forward-time (left) and backward-time (right) vertical trajectory components, and the corresponding model residuals. Trajectories should be independent from model residuals (noise) in the forward-time direction only. For the example track shown, p-values for the forward and backward directions are 0.52 and 0.016 respectively, indicating that forwards time is more likely.

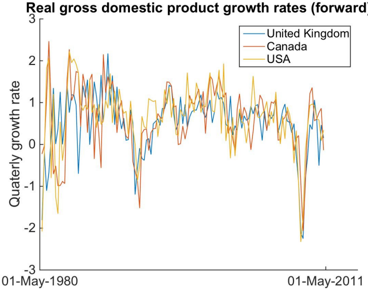
Idea 2: restricted SCMs – arrow of time

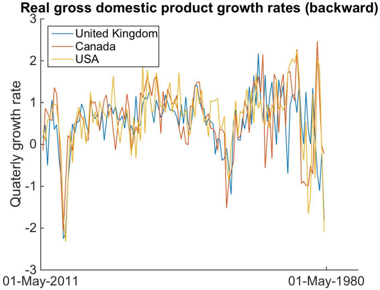




Quarterly growth rates in percentage of GDP for the UK, Canada and USA (Tsay et al 2014).

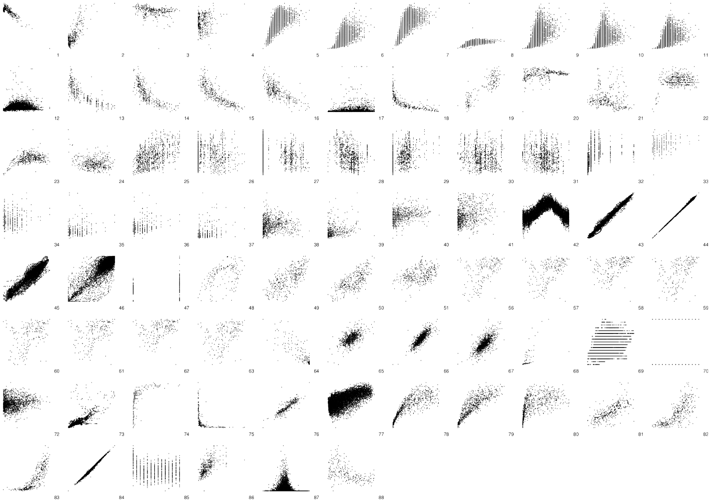
Idea 2: restricted SCMs – arrow of time





Quarterly growth rates in percentage of GDP for the UK, Canada and USA (Tsay et al 2014).

Idea 2: restricted structural causal models



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

Idea 2: restricted structural causal models

Consider a distribution entailed by

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

$$X \longrightarrow Y$$

Idea 2: restricted structural causal models

Consider a distribution entailed by

$$\boxed{\begin{array}{c} Y = f(X) + N_Y \\[1em] \text{with } N_Y, X \overset{ind}{\sim} \mathcal{N} \end{array}} \qquad \textcircled{X} \longrightarrow \textcircled{Y}$$

Then, if f is nonlinear, there is no

$$\boxed{\begin{array}{c} X = g(Y) + M_X \\[1em] \text{with } M_X, Y \overset{ind}{\sim} \mathcal{N} \end{array}} \qquad \textcircled{X} \longleftarrow \textcircled{Y}$$

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

Idea 2: restricted structural causal models

Consider a distribution corresponding to

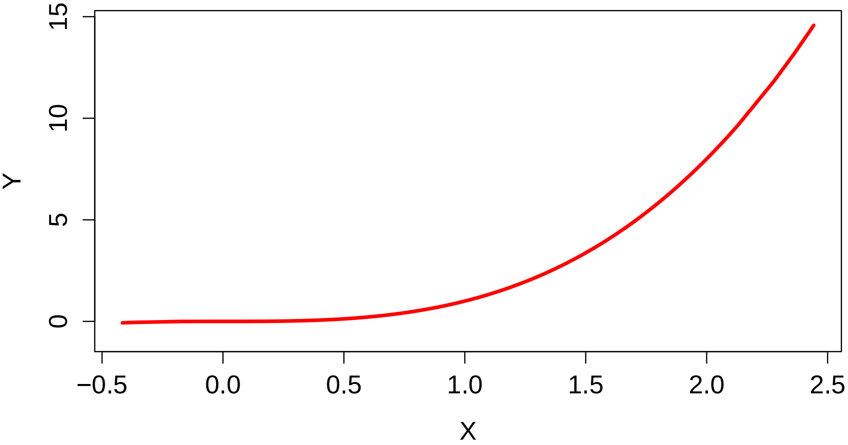
$$\boxed{\begin{array}{c} Y = X^3 + N_Y \\ \\ \text{with } N_Y, X \overset{ind}{\sim} \mathcal{N} \end{array}}$$

$$X \longrightarrow Y$$

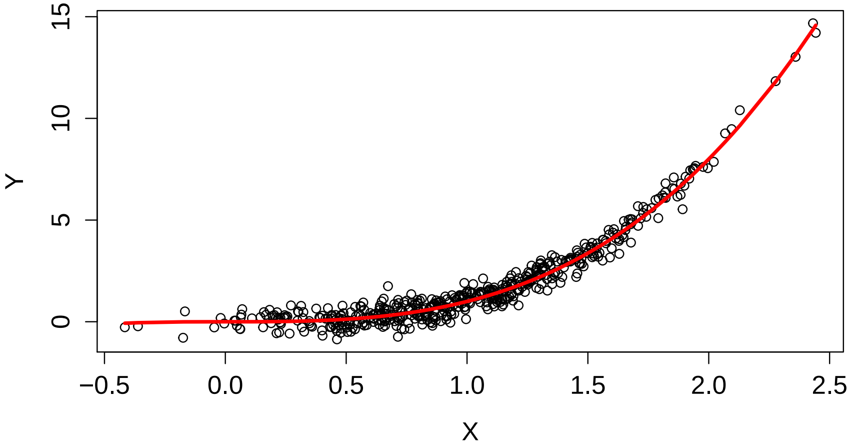
with

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

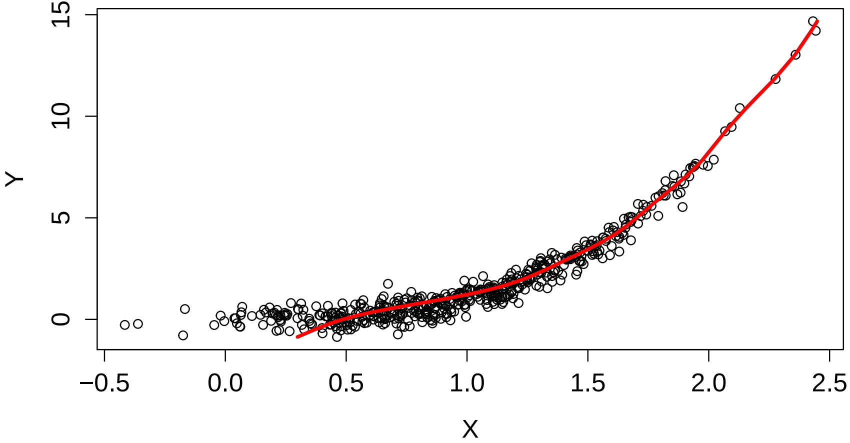
Idea 2: restricted structural causal models



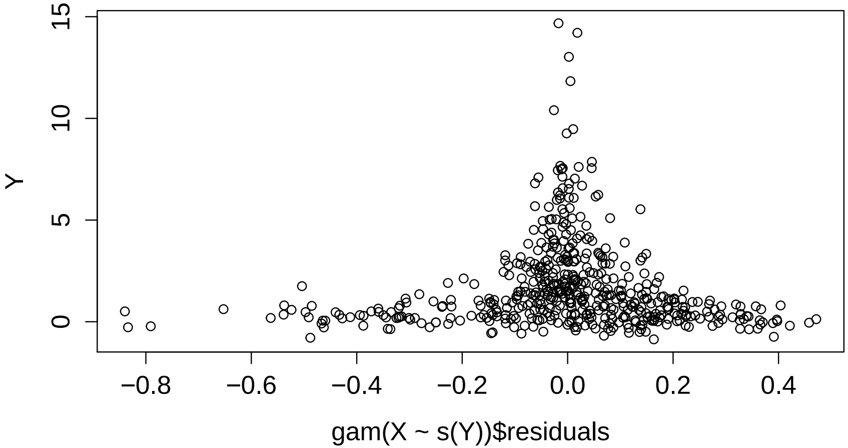
Idea 2: restricted structural causal models



Idea 2: restricted structural causal models

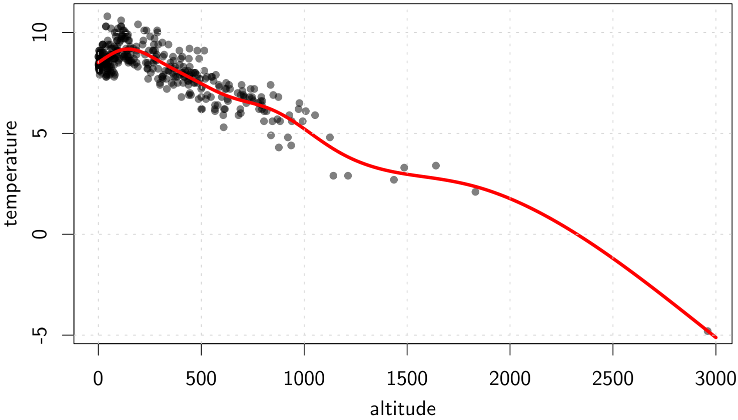


Idea 2: restricted structural causal models

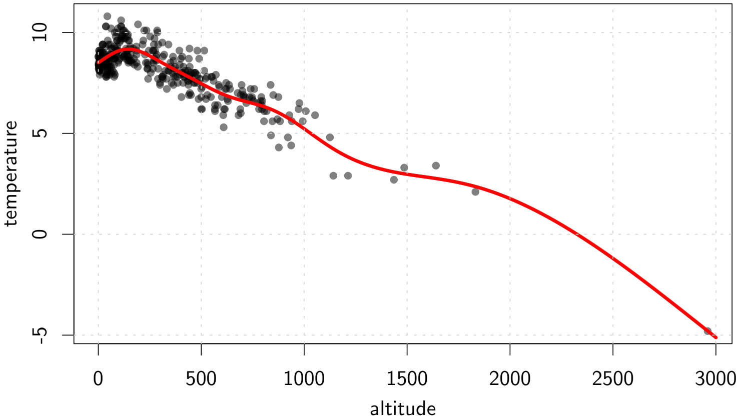


Method... (code)

Real Data: altitude and temperature



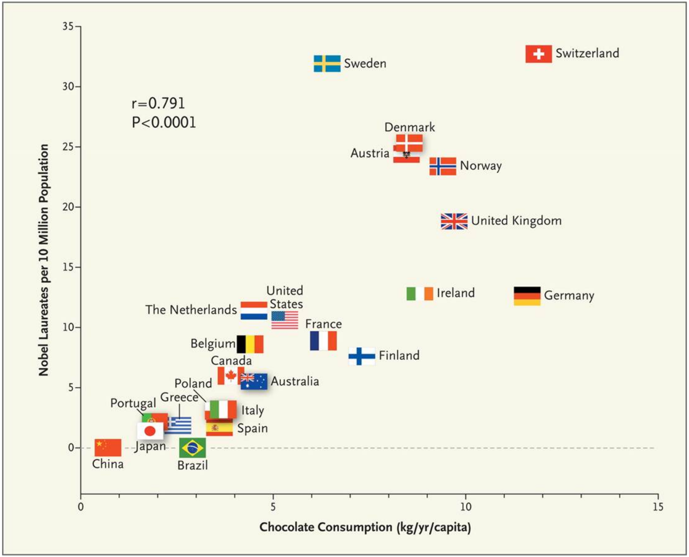
Real Data: altitude and temperature



p-value forward: 0.024

p-value backward: 0.0000000000019

Example: chocolate



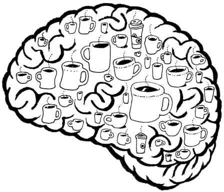
F. H. Messerli: *Chocolate Consumption, Cognitive Function, and Nobel Laureates*, N Engl J Med 2012

Example: chocolate

No (not enough) data for chocolate

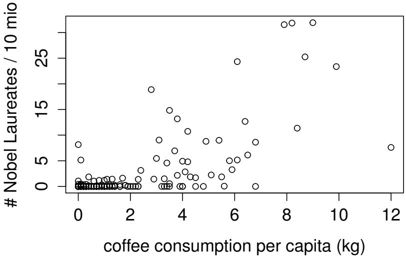
Example: chocolate

No (not enough) data for chocolate



... but we have data for coffee!

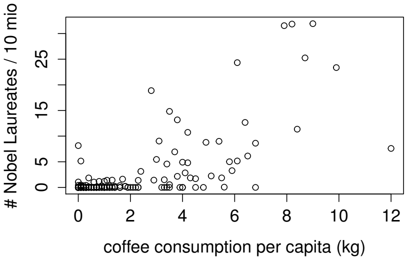
Example: chocolate



Correlation: 0.698

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

Example: chocolate

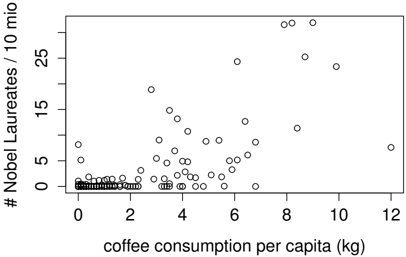


Correlation: 0.698
p-value: $< 2.2 \cdot 10^{-16}$

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

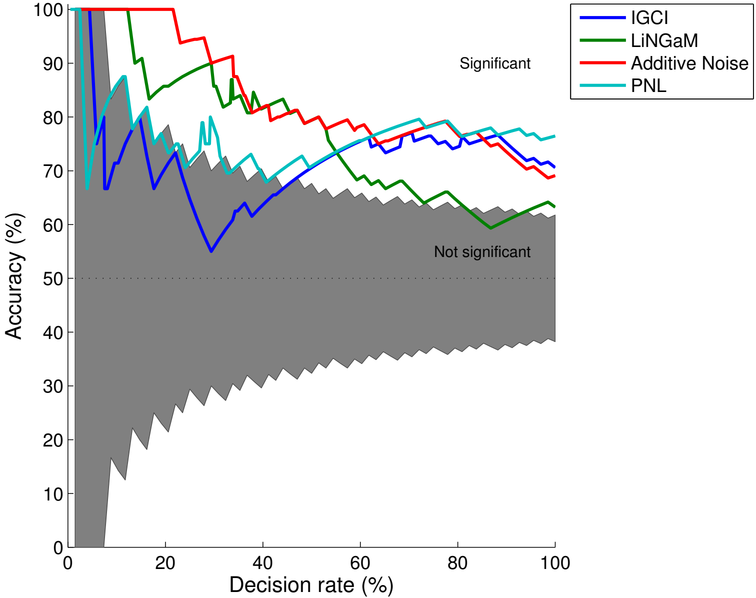


Correlation: 0.698
p-value: $< 2.2 \cdot 10^{-16}$

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?

Real Data: cause-effect pairs



Idea 2: restricted structural causal models

Slightly surprising:

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

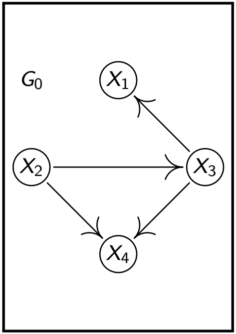
Peters et al.: *Identifiability of Causal Graphs using Functional Models*, UAI 2011

Idea 2: 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



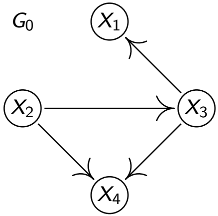
Structural causal model.
Can the DAG be recovered from $P(X_1, \ldots, X_4)$?

Idea 2: 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



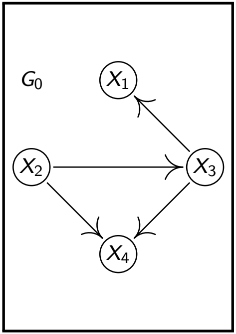
Structural causal model.
Can the DAG be recovered from $P(X_1, \ldots, X_4)$? **No.** (Prop. 7.1. in book)

Idea 2: restricted structural causal models

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

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

- $N_i \sim \mathcal{N}(0, \sigma_i^2)$ 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*

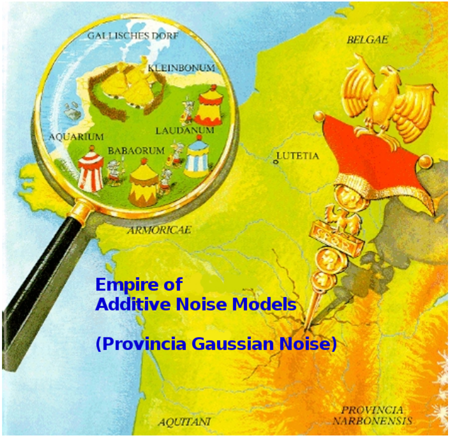
P. Bühlmann, JP, J. Ernest: *CAM: Causal add. models, high-dim. order search and penalized regr.*, Annals of Statistics 2014

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

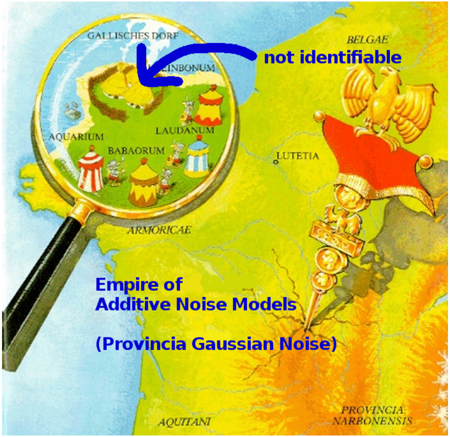
		conditions	identif.
structural causal model:	$X_i = f_i(X_{\mathbf{PA}_i}, N_i)$	-	✗
additive noise model:	$X_i = f_i(X_{\mathbf{PA}_i}) + N_i$	nonlin. fct.	✓
causal additive model:	$X_i = \sum_{k \in \mathbf{PA}_i} f_{ik}(X_k) + N_i$	nonlin. fct.	✓
linear Gaussian model:	$X_i = \sum_{k \in \mathbf{PA}_i}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

(results hold for Gaussian noise)

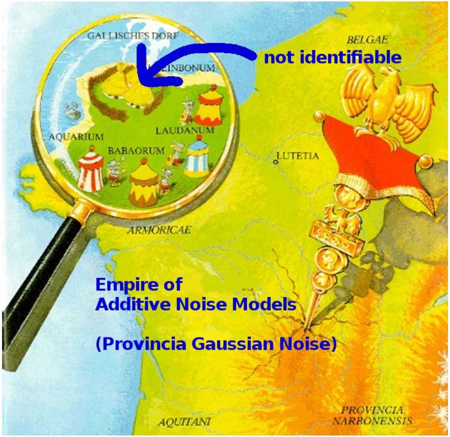
Idea 2: restricted structural causal models



Idea 2: restricted structural causal models

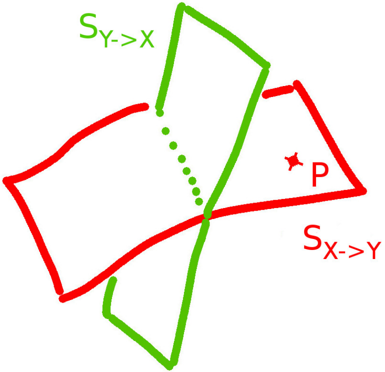


Idea 2: restricted structural causal models

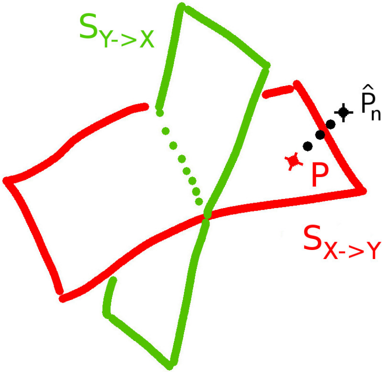


~~GAUL~~ GAUSS
"the LINEAR"

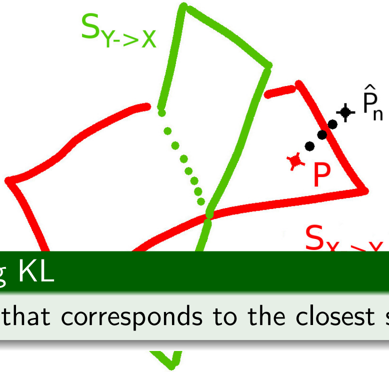
Idea 2: restricted structural causal models



Idea 2: restricted structural causal models



Idea 2: restricted structural causal models



Method: Minimizing KL

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

Idea 2: 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 2: restricted structural causal models

Consider model classes

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

Define

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

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

Idea 2: restricted structural causal models

Consider model classes

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

Define

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

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

Idea 2: restricted structural causal models

Consider model classes

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

Define

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

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

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

Idea 2: restricted structural causal models

Consider model classes

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

Define

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

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

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

Idea 2: restricted structural causal 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	10758229217257614936529561793276243265737276628091852181040900000500559527511693495107583
23	6974332983728149264714154970024580487650427499051598589410910640154981198551095150137712 2074625

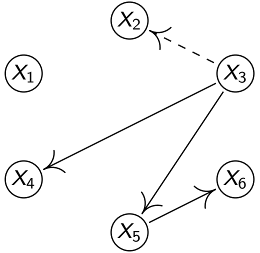
https://oeis.org/A003024/b003024.txt

Idea 2: restricted structural causal 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	–	–

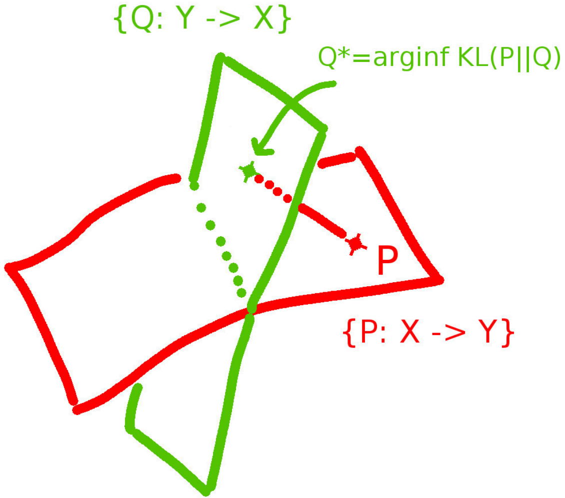
$\xrightarrow[\text{recompute column}]{\text{include best edge}}$



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

Can we characterize identifiability?



Can we characterize identifiability?

Proposition

Assume $P(X, Y)$ is generated by

$$Y = \beta X^2 + N_Y$$

with independent $X \sim \mathcal{N}(0, \sigma_X^2)$ and $N_Y \sim \mathcal{N}(0, \sigma_{N_Y}^2)$.

Can we characterize identifiability?

Proposition

Assume $P(X, Y)$ is generated by

$$Y = \beta X^2 + N_Y$$

with independent $X \sim \mathcal{N}(0, \sigma_X^2)$ and $N_Y \sim \mathcal{N}(0, \sigma_{N_Y}^2)$.

Then

$$\inf_{Q \in \{Q : Y \to X\}} \mathrm{KL}(P \,||\, Q) > 0 \qquad \text{if } \beta \neq 0 \,.$$

Can we characterize identifiability?

Proposition

Assume $P(X, Y)$ is generated by

$$Y = \beta X^2 + N_Y$$

with independent $X \sim \mathcal{N}(0, \sigma_X^2)$ and $N_Y \sim \mathcal{N}(0, \sigma_{N_Y}^2)$.

Then

$$\inf_{Q \in \{Q: Y \to X\}} \mathrm{KL}(P \,||\, Q) = \frac{1}{2} \log \left(1 + 2\beta^2 \frac{\sigma_X^4}{\sigma_{N_Y}^2} \right)$$

Leonardo da Vinci: Mould of the Horses Head

Leonardo da Vinci: Mould of the Horses Head

Given an original drawing (left) and a copy. How good is the copy?

Leonardo da Vinci: Mould of the Horses Head

Given an original drawing (left) and a copy. How good is the copy?

Given a true causal graph G and an estimate $\hat{G}$. How good is the estimate $\hat{G}$?

Leonardo da Vinci: Mould of the Horses Head

Given an original drawing (left) and a copy. How good is the copy?

Given a true causal graph G and an estimate $\hat{G}$. How good is the estimate $\hat{G}$?

What do we want do with it?

Definition: Structural Intervention Distance

For each pair (X, Y) check whether $\mathbf{PA}_X^{\hat{G}}$ is a valid adjustment set for (X, Y) in G for all distributions Markov w.r.t. G.

Definition: Structural Intervention Distance

For each pair (X, Y) check whether $\mathbf{PA}_X^{\hat{G}}$ is a valid adjustment set for (X, Y) in G for all distributions Markov w.r.t. G. - does not depend on P

$\mathrm{SID}(G, \hat{G})$ equals the number of pairs, for which this is not the case.

Definition: Structural Intervention Distance

For each pair (X, Y) check whether $\mathbf{PA}_X^{\hat{G}}$ is a valid adjustment set for (X, Y) in G for all distributions Markov w.r.t. G. - does not depend on P

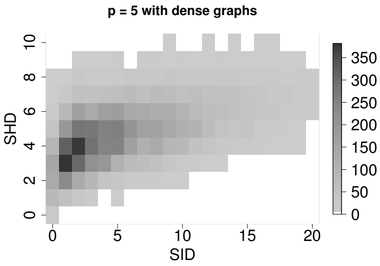
$\text{SID}(G, \hat{G})$ equals the number of pairs, for which this is not the case.

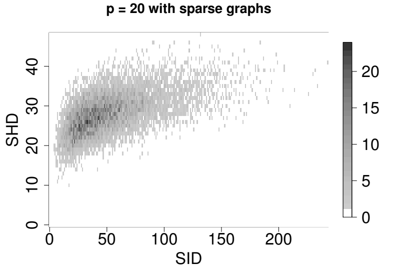
Graphical representation of the SID available!!

based on Shpitser et al: "On the validity of covariate adjustment for estimating causal effects", UAI 2010.

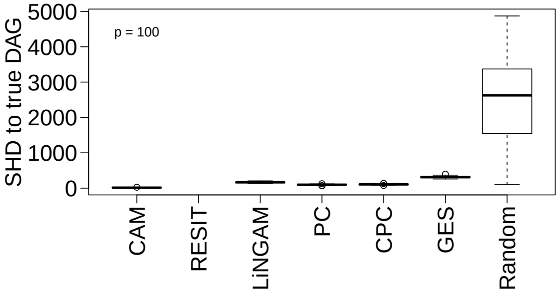
SHD and SID are quite different!

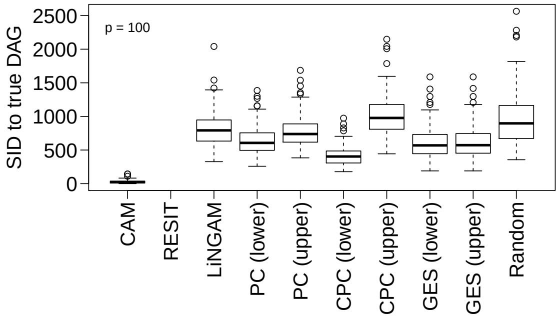
10, 000 random DAGs



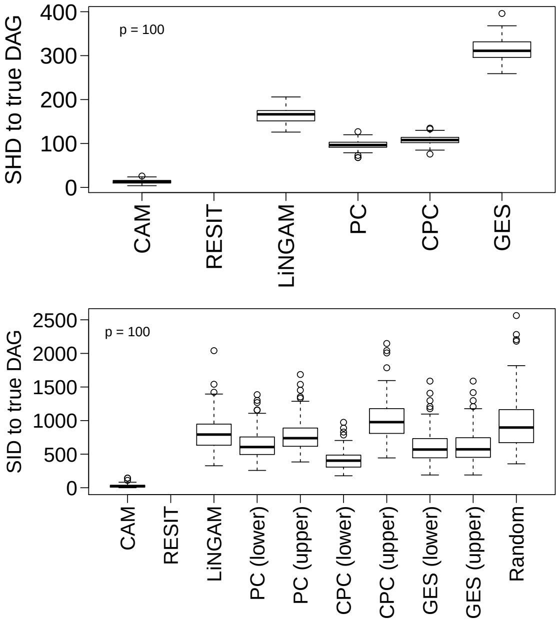


SHD and SID may lead to different conclusions (nonlinear, Gaussian).





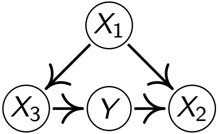
SHD and SID may lead to different conclusions (nonlinear, Gaussian).

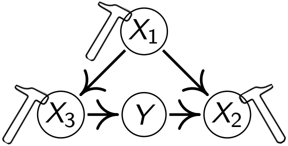


Idea 3: invariant causal prediction

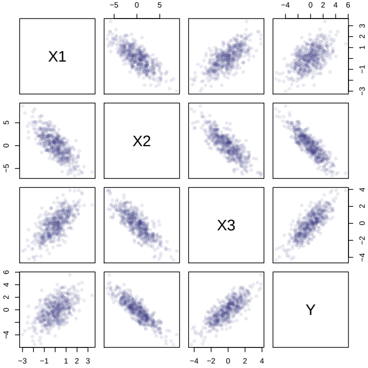
Concentrate on one target variable.

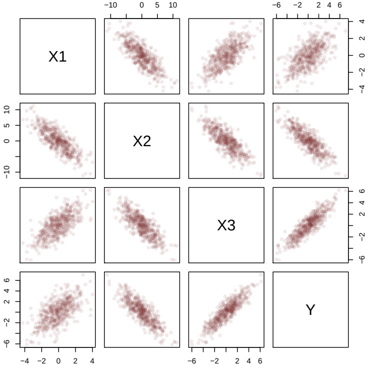
unknown:





known:





linear model

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

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...22222222222222...

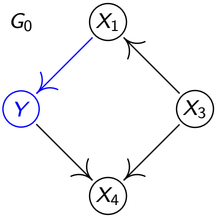
```
> icp <- ICP(X,Y,ExpInd)

            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: MUTE ... or:

$P(Y \mid \mathbf{PA}_Y)$ 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



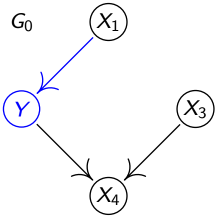
IMPORTANT: modularity, autonomy

Haavelmo 1944, Aldrich 1989, Pearl 2009, Schölkopf et al. 2012, Bareinboim et al. 2014, Hauser et al. 2014, ...

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

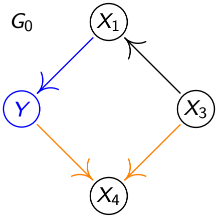


IMPORTANT: modularity, autonomy
Haavelmo 1944, Aldrich 1989, Pearl 2009, Schölkopf et al. 2012, Bareinboim et al. 2014, Hauser et al. 2014, ...

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

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

- N_i jointly independent
- G_0 has no cycles



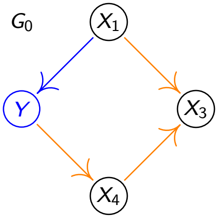
IMPORTANT: modularity, autonomy
Haavelmo 1944, Aldrich 1989, Pearl 2009, Schölkopf et al. 2012, Bareinboim et al. 2014, Hauser et al. 2014, ...

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

G_0



IMPORTANT: modularity, autonomy
Haavelmo 1944, Aldrich 1989, Pearl 2009, Schölkopf et al. 2012, Bareinboim et al. 2014, Hauser et al. 2014, . . .

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

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

Proposition

Let $S^* = \mathbf{PA}_Y$. Then, $H_{0,S^*}(\mathcal{E})$ is true, i.e.,

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

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

Proposition

Let $S^* = \mathbf{PA}_Y$. Then, $H_{0,S^*}(\mathcal{E})$ is true, i.e., there exists γ^* with support S^* that satisfies

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

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

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

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

Proposition

Let $S^* = \mathbf{PA}_Y$. Then, $H_{0,S^*}(\mathcal{E})$ is true, i.e., there exists γ^* with support S^* that satisfies

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

Goal: Find S^*.

Idea: Check $H_{0,S}(\mathcal{E})$ for several candidates S.

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

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

Theorem (PBM 2016)

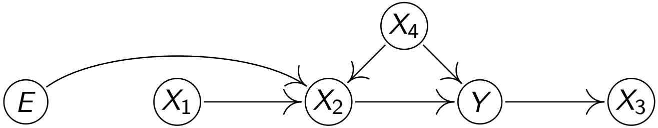
$$P(\hat{S}(\mathcal{E}) \subseteq S^*) \geq 1 - \alpha \, .$$

Theorem (PBM 2016)

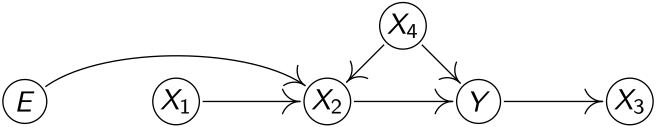
$$P(\hat{S}(\mathcal{E}) \subseteq S^*) \geq 1 - \alpha\,.$$

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

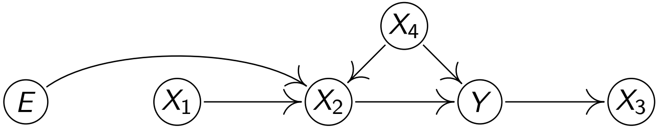
JP, P. Bühlmann, N. Meinshausen: *Causal inference using invariant prediction: conf. interv.*, JRSS-B 2016 (w/ discussion).



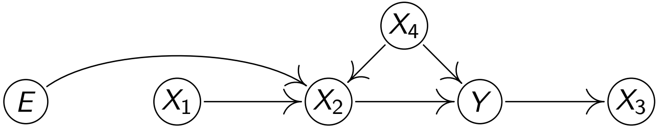
X_4
E
X_1
X_2
Y
X_3



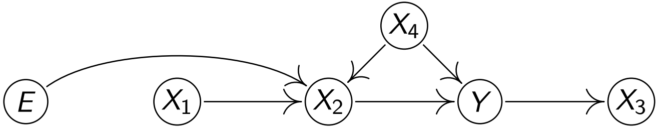

```
> Y <- X[,2] + X[,4] + noise
> ICP(X,Y,ExpInd)
```




```
> 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 ***
```



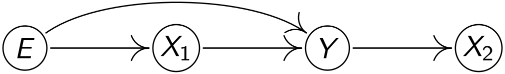

```
> Y <- X[,2]^2 + X[,4] + noise
> ICP(X,Y,ExpInd)
```



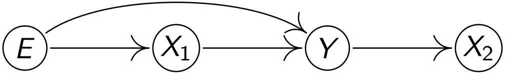

```
> Y <- X[,2]^2 + X[,4] + noise
> ICP(X,Y,ExpInd)

  empty set
  (all models rejected)
```

Model violation: nonlinear models

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



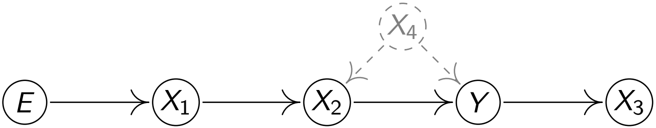

```
> Y <- X[,1] + E + noise
> ICP(X,Y,ExpInd)
```



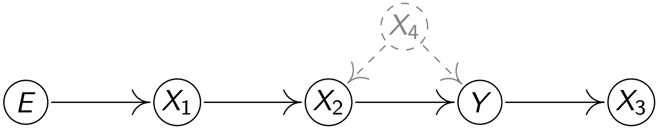

```
> Y <- X[,1] + E + noise
> ICP(X,Y,ExpInd)

  empty set
  (all models rejected)
```

Model violation: intervention on Y

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



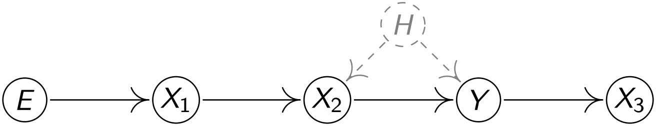

```
> Y <- X[,2] + X[,4] + noise
> ICP(X[,1:3],Y,ExpInd)
```




```
> 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



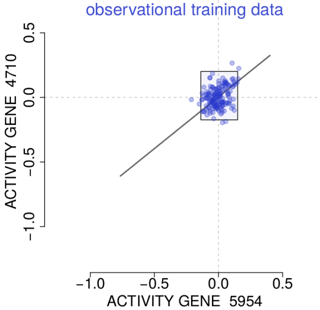
Theorem (PBM 2016)

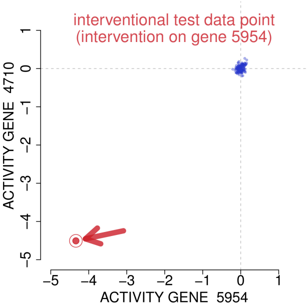
Assume that the joint distribution over $(Y, X_1, \dots, X_p, H_1, \dots, H_q, E)$ *is faithful w.r.t. the augmented graph. Then*

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

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)

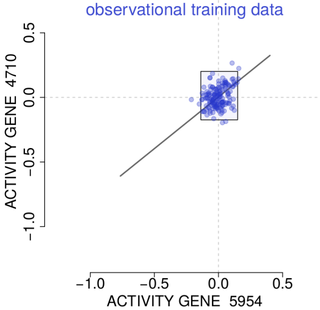


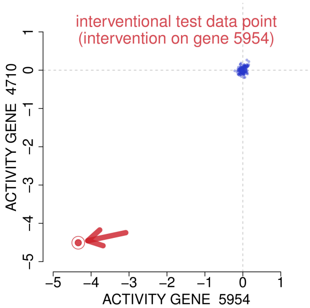


- true hits: $\approx 0.1\%$ of pairs

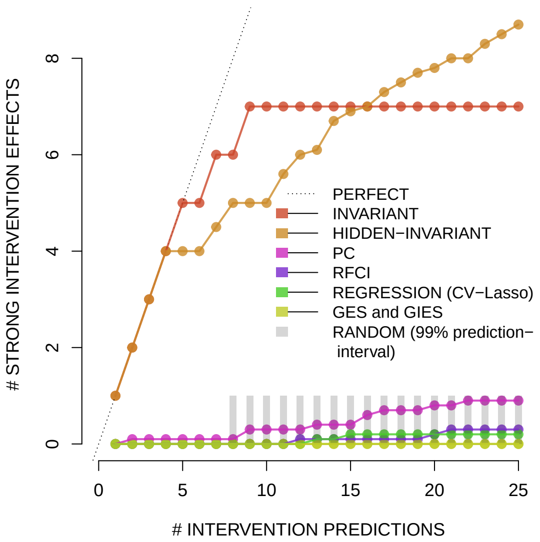
Real data: genetic perturbation experiments for yeast (Kemmeren et al., 2014)

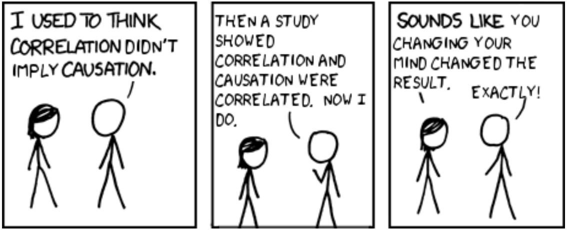
- $p = 6170$ genes
- $n_{obs} = 160$ wild-types
- $n_{int} = 1479$ gene deletions (targets known)





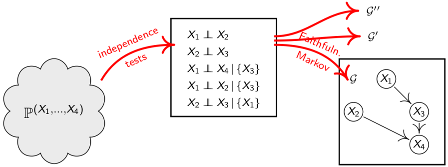
- true hits: $\approx 0.1\%$ of pairs
- our method: $\mathcal{E} = \{obs, int\}$


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


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!)

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:

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

Modularity suggests:

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

Idea 1: semi-supervised learning

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

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

Modularity suggests:

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

Special case:

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

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:

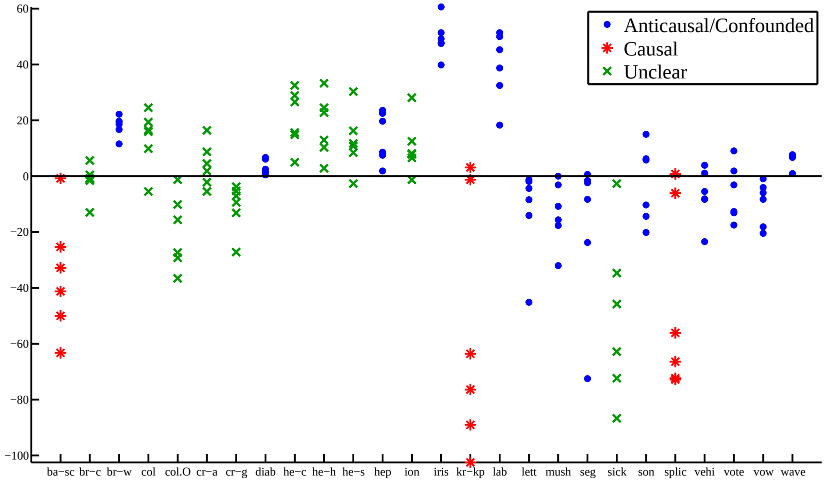
$$p(x_1 \,|\, x_{pa(1)}), \ldots, p(x_d \,|\, x_{pa(d)}) \text{ are “independent”}$$

Special case:

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

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

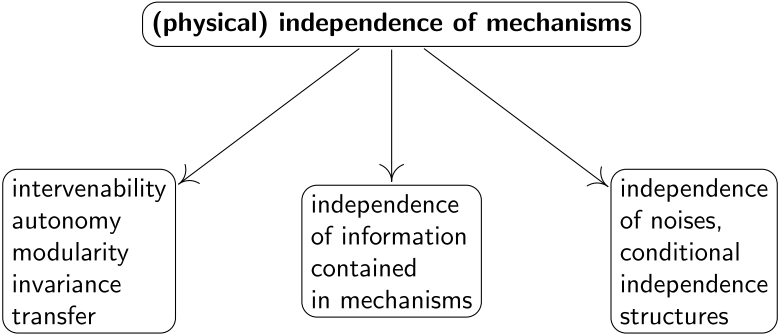
Idea 1: semi-supervised learning



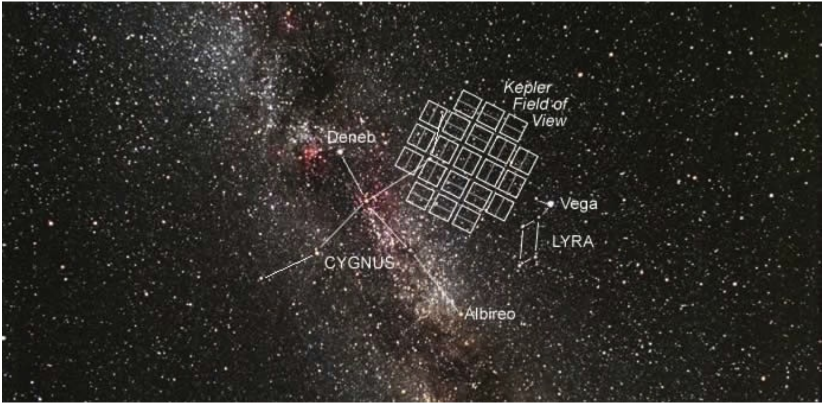
comparison to base classifier

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

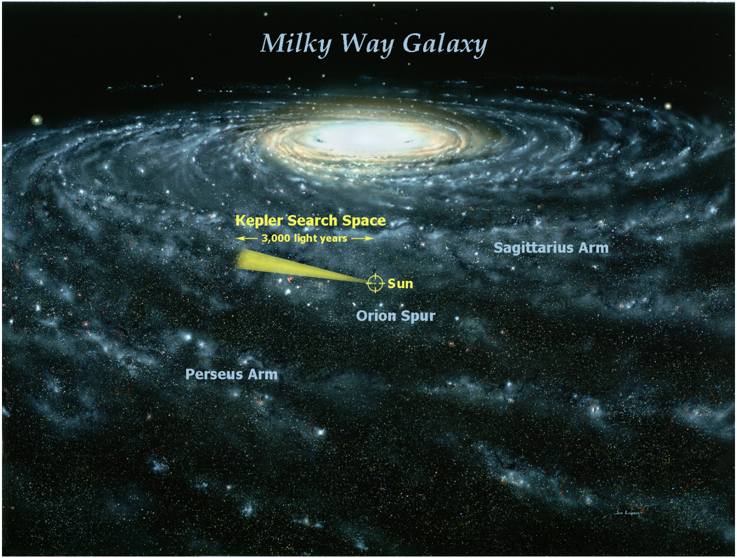
Idea 1: semi-supervised learning



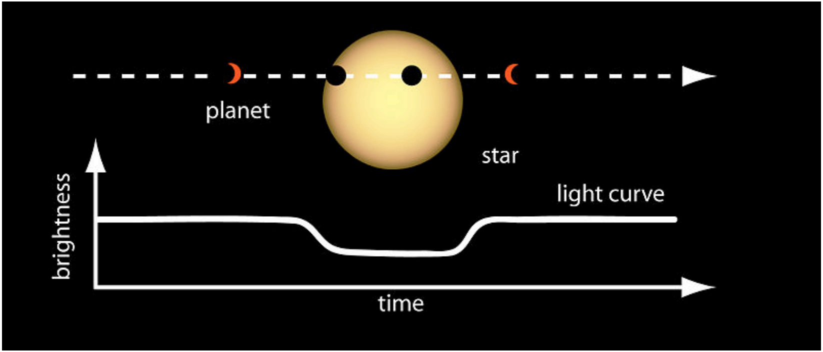
Idea 2: half-sibling regression



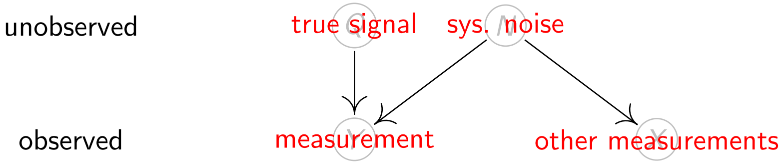
Idea 2: half-sibling regression



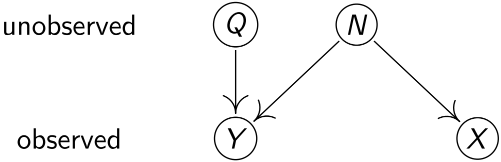
Idea 2: half-sibling regression



Idea 2: half-sibling regression

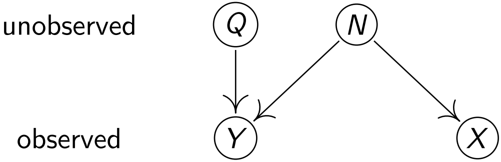


Idea 2: half-sibling regression



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

Idea 2: half-sibling regression



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

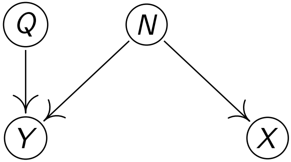
Proposed idea:

Remove everything from Y explained by X.

Idea 2: half-sibling regression

unobserved

observed



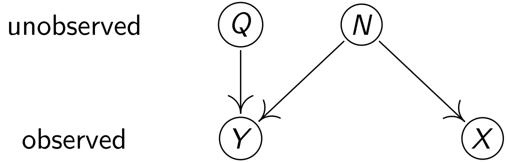
Assume $Y = f(N) + Q$.

Proposed idea:

Remove everything from Y explained by X.

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

Idea 2: half-sibling regression



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

Proposed idea:

Remove everything from Y explained by X.

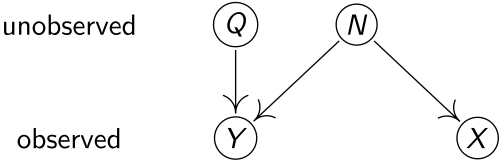
Or: $\hat{Q} := Y - \mathbf{E}[Y \mid X]$.

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

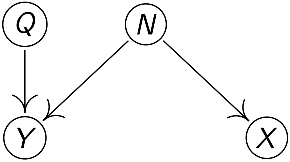
Convergence against "correct" signal Q (up to reparameterization) if

- *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

unobserved

observed



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

Proposed idea:

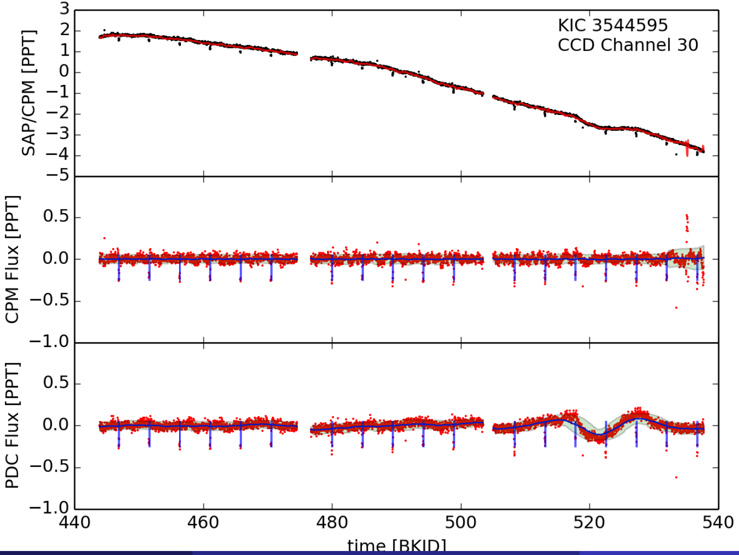
Remove everything from Y explained by X. Or: $\hat{Q} := Y - \mathbf{E}[Y \mid X]$.

Proposition

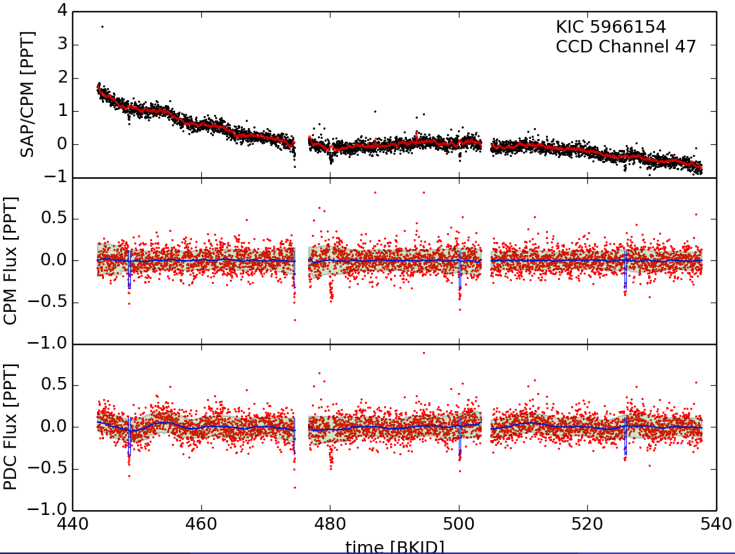
Convergence against "correct" signal Q (up to reparameterization) if

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

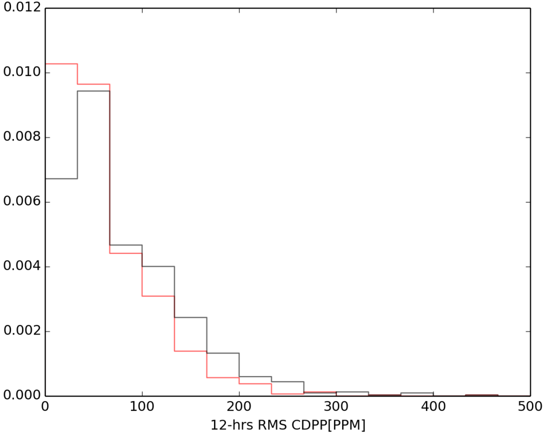
Idea 2: half-sibling regression



Idea 2: half-sibling regression



Idea 2: half-sibling regression

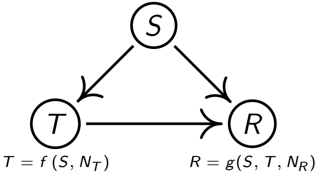


Schölkopf et al.: *Removing systematic errors for exoplanet search via latent causes*, ICML 2015

Schölkopf et al.: *Modeling Confounding by Half-Sibling Regression*, PNAS 2016

Idea 3: Blackjack (reinforcement learning)

Recall the kidney stones:

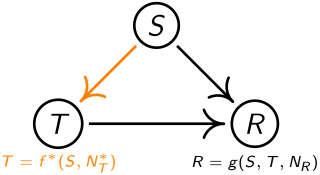


$$p(r, t, s) = p(r \mid t, s) \cdot \quad p(t \mid s) \quad \cdot p(s)$$

Question: What would happen if...?

Idea 3: Blackjack (reinforcement learning)

Recall the kidney stones:



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

$$\downarrow$$

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

Question: What would happen if...?

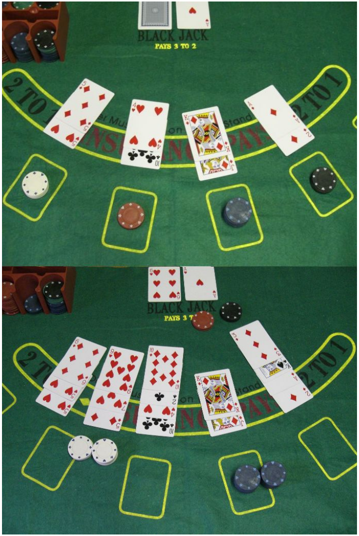
What is $\sup_{p^*} \mathbf{E}_{p^*} R$?

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

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).

Idea 3: Blackjack

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

Idea 3: Blackjack

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 | 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
⋮	⋮	⋮	⋮

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
⋮	⋮	⋮	⋮

Idea 3: Blackjack

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

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

Idea 3: Blackjack

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

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

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

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

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

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

Idea 3: Blackjack

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

Which p^* is best? Parameterize and estimate

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

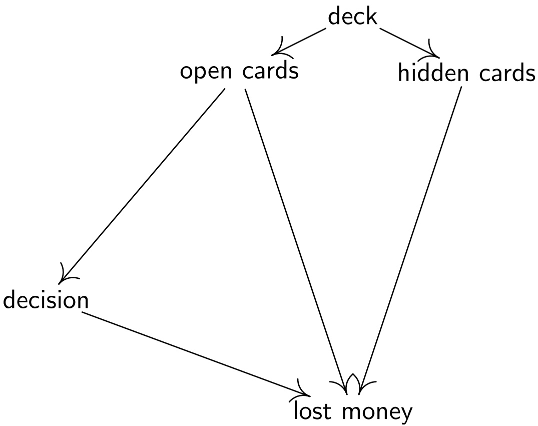
Goal: Optimize $\mathbf{E}_{p_\theta}\ell$

Idea: Use gradient $\nabla_\theta \mathbf{E}_{p_\theta}\ell$ and optimize step-by-step.

Issues: confidence intervals, step size, ...

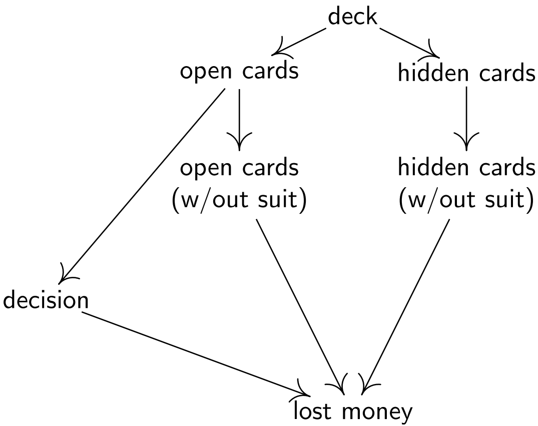
Idea 3: Blackjack

How to exploit causal structure (state simplification):



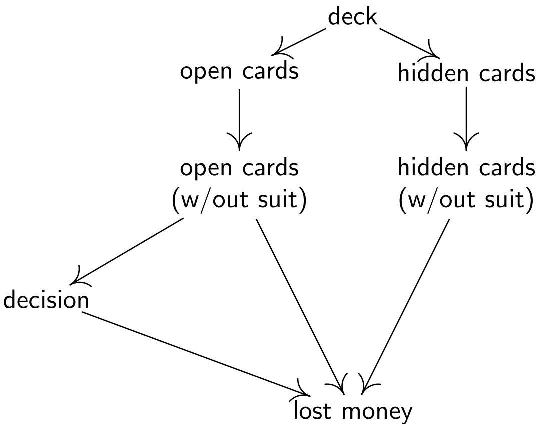
Idea 3: Blackjack

How to exploit causal structure (state simplification):

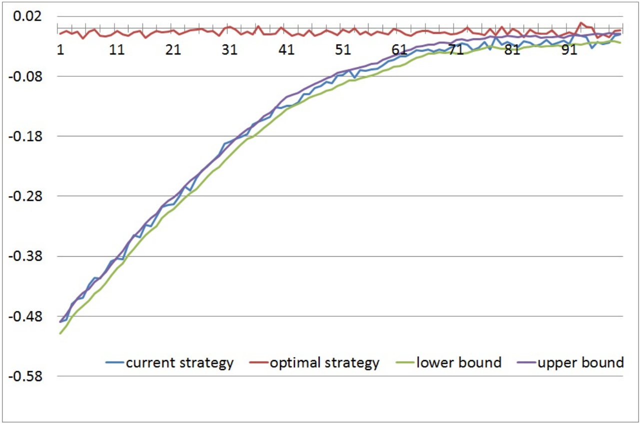


Idea 3: Blackjack

How to exploit causal structure (state simplification):

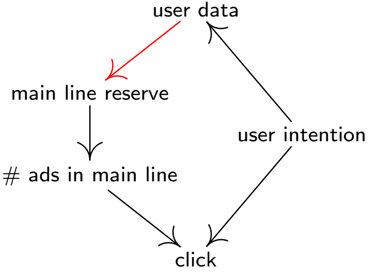


Idea 3: Blackjack

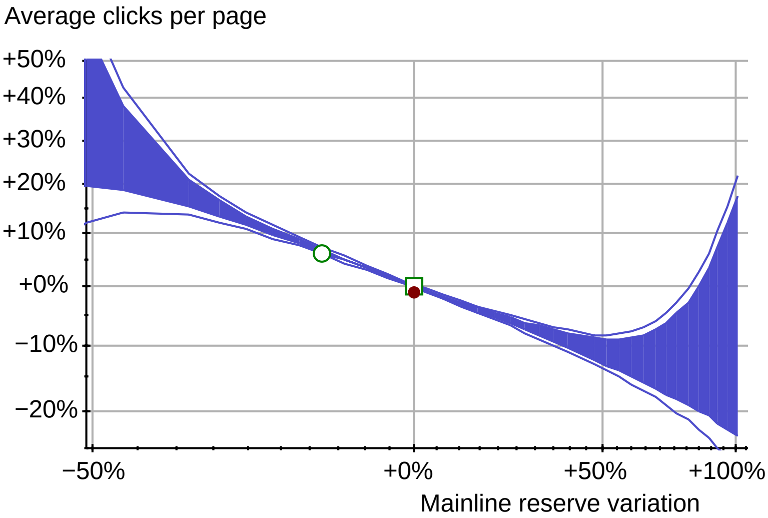


Idea 3: advertisement

How to exploit causal structure (improved weighting):

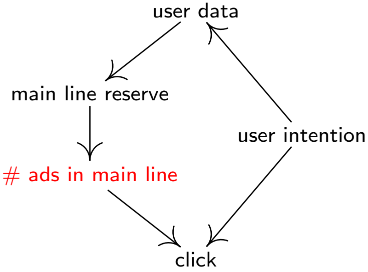


Idea 3: advertisement



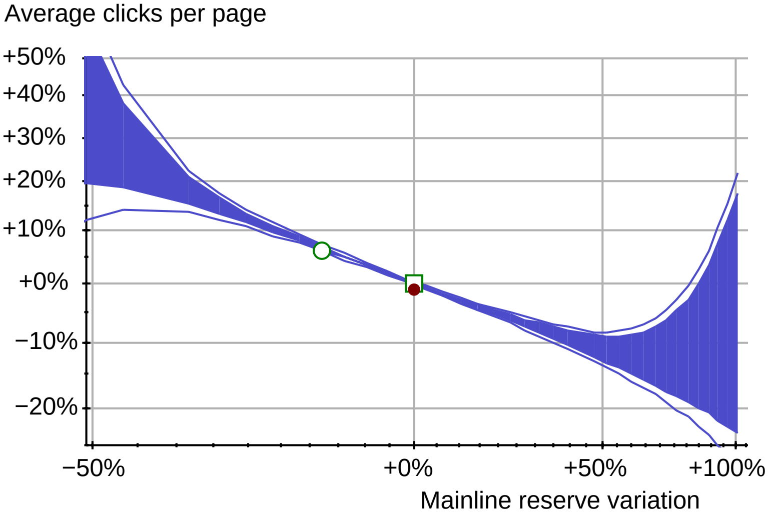
Idea 3: advertisement

How to exploit causal structure (improved weighting):



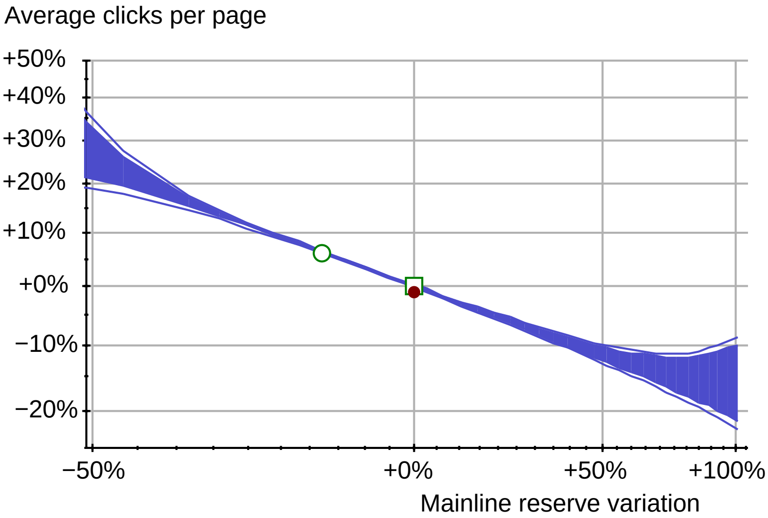
Idea 3: advertisement

Old:



Idea 3: advertisement

Using discrete variable (ads shown in mainline):



Idea 4: domain adaptation

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

Idea 4: domain adaptation

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

Assumption:

$$Y^1 | \mathbf{X}^1_{S^*} \stackrel{\mathcal{L}}{=} Y^2 | \mathbf{X}^2_{S^*} \stackrel{\mathcal{L}}{=} \ldots \stackrel{\mathcal{L}}{=} Y^D | \mathbf{X}^D_{S^*}$$

Idea 4: domain adaptation

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

Assumption:

$$Y^1 | \mathbf{X}^1_{S^*} \stackrel{\mathcal{L}}{=} Y^2 | \mathbf{X}^2_{S^*} \stackrel{\mathcal{L}}{=} \ldots \stackrel{\mathcal{L}}{=} Y^D | \mathbf{X}^D_{S^*}$$

(relaxation of covariate shift).

Idea 4: domain adaptation

How to transfer the knowledge? Assume $S^* = \{1, 3, 4\}$. Suppose you know $\alpha \in \mathbb{R}^3$ in

$$Y = \alpha^t X_{1,3,4} + N, \quad N \perp\!\!\!\perp X_{1,3,4}.$$

Idea 4: domain adaptation

How to transfer the knowledge? Assume $S^* = \{1, 3, 4\}$. Suppose you know $\alpha \in \mathbb{R}^3$ in

$$Y = \alpha^t X_{1,3,4} + N, \quad N \perp\!\!\!\perp X_{1,3,4}.$$

How does it help you to find a good estimator for $\beta \in \mathbb{R}^5$

$$Y = \beta^t X_{1,2,3,4,5} + M, \quad M \perp\!\!\!\perp X_{1,2,3,4,5}?$$

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:

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

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

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

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)

- J. Pearl: *Causality*
- P. Spirtes, C. Glymour, R. Scheines: *Causation, Prediction and Search*
- J. Peters, D. Janzing, B. Schölkopf: *Elements of Causal Inference: Foundations and Learning Algorithms*

  http://www.math.ku.dk/~peters/jonas_files/bookDRAFT5-online-2017-02-27.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)

- J. Pearl: *Causality*
- P. Spirtes, C. Glymour, R. Scheines: *Causation, Prediction and Search*
- J. Peters, D. Janzing, B. Schölkopf: *Elements of Causal Inference: Foundations and Learning Algorithms*

  `http://www.math.ku.dk/~peters/jonas_files/bookDRAFT5-online-2017-02-27.pdf`

Dankeschön!

nachdenken • klimabewusst reisen
atmosfair

References I

J. Aldrich. Autonomy. Oxford Economic Papers, 41:15–34, 1989.

E. Bareinboim and J. Pearl. Transportability from multiple environments with limited experiments: Completeness results. In Advances in Neural Information Processing Systems 27 (NIPS), pages 280–288, 2014.

S. Bauer, B. Schölkopf, and J. Peters. The arrow of time in multivariate time series. In Proceedings of the 33rd International Conference on Machine Learning (ICML), pages 2043–2051, 2016.

L. Bottou, J. Peters, J. Quiñonero-Candela, D. X. Charles, D. M. Chickering, E. Portugualy, D. Ray, P. Simard, and E. Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. Journal of Machine Learning Research, 14:3207–3260, 2013.

P. Bühlmann, J. Peters, and J. Ernest. CAM: Causal additive models, high-dimensional order search and penalized regression. The Annals of Statistics, 42(6):2526–2556, 2014.

C. R. Charig, D. R. Webb, S. R. Payne, and J. E. A. Wickham. Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy. British Medical Journal (Clin Res Ed), 292(6254): 879–882, 1986.

D. M. Chickering. Optimal structure identification with greedy search. Journal of Machine Learning Research, 3:507–554, 2002.

R. Doll and A. B. Hill. Smoking and carcinoma of the lung. British Medical Journal, 2(4682):739–748, 1950.

A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. Smola. A kernel statistical test of independence. In Advances in Neural Information Processing Systems 20 (NIPS), pages 585–592, 2008.

J. Gwiazda, E. Ong, R. Held, and F. Thorn. Vision: Myopia and ambient night-time lighting. Nature, 404:144, 2000.

T. Haavelmo. The statistical implications of a system of simultaneous equations. Econometrica, 11(1), 1943.

T. Haavelmo. The probability approach in econometrics. Econometrica, 12:S1–S115 (supplement), 1944.

A. Hauser and P. Bühlmann. Jointly interventional and observational data: estimation of interventional Markov equivalence classes of directed acyclic graphs. Journal of the Royal Statistical Society, Series B: Statistical Methodology, 77:291–318, 2015.

D. Janzing, D. Balduzzi, M. Grosse-Wentrup, and B. Schölkopf. Quantifying causal influences. The Annals of Statistics, 41(5): 2324–2358, 2013.

References II

P. Kemmeren, K. Sameith, L. A. van de Pasch, J. J. Benschop, T. L. Lenstra, T. Margaritis, E. O'Duibhir, E. Apweiler, S. van Wageningen, C. W. Ko, S. van Heesch, M. M.. Kashani, G. Ampatziadis-Michailidis, M. O. Brok, N. A. Brabers, A. J. Miles, D. Bouwmeester, S. R. van Hooff, H. van Bakel, E. Sluiters, L. V. Bakker, B. Snel, P. Lijnzaad, D. van Leenen, M. J. Groot Koerkamp, and F. C. Holstege. Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. Cell, 157(3):740–752, 2014.

S. L. Lauritzen. Graphical Models. Oxford University Press, New York, NY, 1996.

N. Meinshausen, A. Hauser, J. Mooij, P. Versteeg, J. Peters, and P. B/"uhlmann. Causal inference from gene perturbationexperiments: methods, software and validation. Proceedings of the National Academy of Sciences, 113(27): 7361–7368, 2016.

F. H. Messerli. Chocolate consumption, cognitive function, and nobel laureates. New England Journal of Medicine, 367(16): 1562–1564, 2012.

J. M. Mooij, D. Janzing, T. Heskes, and B. Schölkopf. On causal discovery with cyclic additive noise models. In Advances in Neural Information Processing Systems 24 (NIPS), 2011.

J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. Journal of Machine Learning Research, 17:1–102, 2016.

J. Pearl. Causality: Models, Reasoning, and Inference. Cambridge University Press, New York, NY, 2nd edition, 2009.

J. Peters and P. Bühlmann. Identifiability of Gaussian structural equation models with equal error variances. Biometrika, 101(1): 219–228, 2014.

J. Peters and P. Bühlmann. Structural intervention distance (SID) for evaluating causal graphs. Neural Computation, 27: 771–799, 2015.

J. Peters, D. Janzing, A. Gretton, and B. Schölkopf. Detecting the direction of causal time series. In Proceedings of the 26th International Conference on Machine Learning (ICML), pages 801–808, 2009.

J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. Identifiability of causal graphs using functional models. In Proceedings of the 27th Annual Conference on Uncertainty in Artificial Intelligence (UAI), pages 589–598, 2011.

J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. Causal discovery with continuous additive noise models. Journal of Machine Learning Research, 15:2009–2053, 2014.

References III

J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference using invariant prediction: identification and confidence intervals. Journal of the Royal Statistical Society, Series B: Statistical Methodology (with discussion), 78(5):947–1012, 2016.

N. Pfister, P. Bühlmann, B. Schölkopf, and J. Peters. Kernel-based tests for joint independence. ArXiv e-prints (1603.00285), 2016.

L. C. Pickup, Z. Pan, D. Wei, Y. Shih, C. Zhang, A. Zisserman, B. Schölkopf, and W. T. Freeman. Seeing the arrow of time. In IEEE Conference on Computer Vision and Pattern Recognition, 2014.

G. E. Quinn, C. H. Shin, M. G. Maguire, and R. A. Stone. Myopia and ambient lighting at night. Nature, 399:113–114, 1999.

M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters. Causal transfer in machine learning. ArXiv e-prints (1507.05333v3), 2016.

B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. M. Mooij. On causal and anticausal learning. In Proceedings of the 29th International Conference on Machine Learning (ICML), pages 1255–1262, 2012.

B. Schölkopf, D. W. Hogg, D. Wang, D. Foreman-Mackey, D. Janzing, C.-J. Simon-Gabriel, and J. Peters. Removing systematic errors for exoplanet search via latent causes. In Proceedings of the 32nd International Conference on Machine Learning (ICML), pages 2218–2226, 2015.

B. Schölkopf, D. W. Hogg, D. Wang, D. Foreman-Mackey, D. Janzing, C.-J. Simon-Gabriel, and J. Peters. Modeling confounding by half-sibling regression. Proceedings of the National Academy of Sciences, 113(27):7391–7398, 2016.

S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. J. Kerminen. A linear non-Gaussian acyclic model for causal discovery. Journal of Machine Learning Research, 7:2003–2030, 2006.

I. Shpitser, T. J. Van der Weele, and J. M. Robins. On the validity of covariate adjustment for estimating causal effects. In Proceedings of the 26th Annual Conference on Uncertainty in Artificial Intelligence (UAI), pages 527–536, 2010.

P. Spirtes, C. Glymour, and R. Scheines. Causation, Prediction, and Search. MIT Press, Cambridge, MA, 2nd edition, 2000.

C. Uhler, G. Raskutti, P. Bühlmann, and B. Yu. Geometry of the faithfulness assumption in causal inference. The Annals of Statistics, 41(2):436–463, 2013.

K. Zadnik, L. A. Jones, B. C. Irvin, R. N. Kleinstein, R. E. Manny, J. A. Shin, and D. O. Mutti. Vision: Myopia and ambient night-time lighting. Nature, 404:143–144, 2000.

K. Zhang and A. Hyvärinen. On the identifiability of the post-nonlinear causal model. In Proceedings of the 25th Annual Conference on Uncertainty in Artificial Intelligence (UAI), pages 647–655, 2009.

K. Zhang, J. Peters, D. Janzing, and B. Schölkopf. Kernel-based conditional independence test and application in causal discovery. In Proceedings of the 27th Annual Conference on Uncertainty in Artificial Intelligence (UAI), pages 804–813, 2011.