Causality

Joris Mooij and Jonas Peters

Informatics Institute
University of Amsterdam

Department for Mathem. Sciences
University of Copenhagen

MSR, Cambridge
3 July 2018

Die Junge Akademie
www.diejungeakademie.de

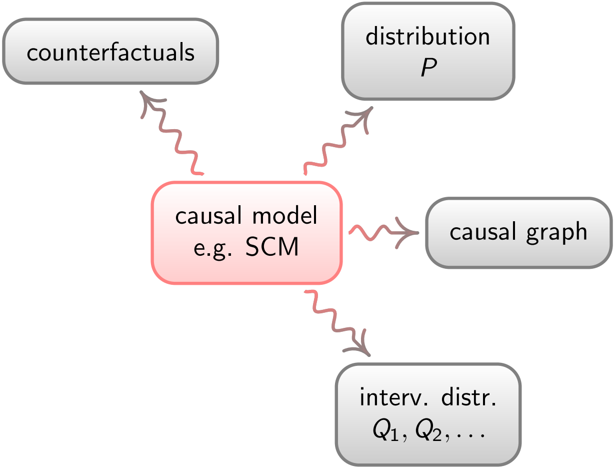
UNIVERSITY OF COPENHAGEN

is based on work by ...

- UCLA: J. Pearl
- CMU: P. Spirtes, C. Glymour, R. Scheines
- Harvard University: D. Rubin, J. Robins
- ETH Zürich: P. Bühlmann, N. Meinshausen, N. Pfister, D. Rothenhäusler
- Max-Planck-Institute Tübingen: D. Janzing, B. Schölkopf
- University of Amsterdam: J. Mooij
- P. Hoyer
- ... and many others

Recall Part I

What is a causal model?



Recall Part I

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

Recall Part I

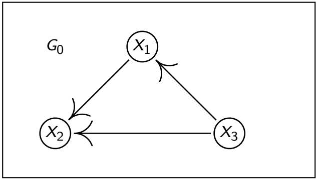
- What if interested in iid prediction, i.e., observational data? Don't worry (too much) about causality!
- But often, we are interested in a system's behaviour under intervention.

Recall 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

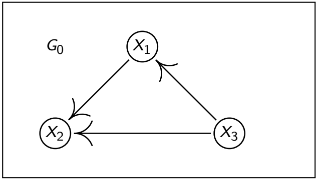


Recall Part I

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

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

- N_i jointly independent
- G_0 has no cycles



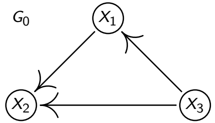
- adjusting: graph + observational distribution $\rightsquigarrow$ interventions

Recall Part I

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

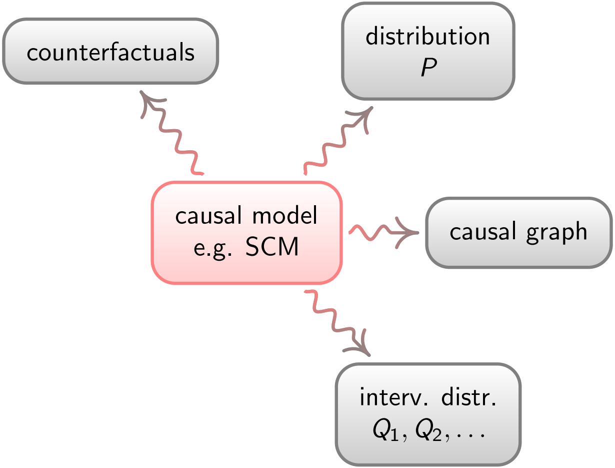
$$X_1 := f_1(X_3, N_1)$$
$$X_2 := f_2(X_1, X_3, N_2)$$
$$X_3 := f_3(N_3)$$

- N_i jointly independent
- G_0 has no cycles



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

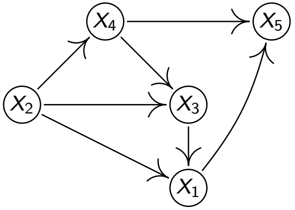
Part II: Causal Discovery


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

The Problem of Causal Discovery:

observed iid data from $P(X_1, \ldots, X_5)$ $\xrightarrow{?}$ causal model, e.g. DAG $\mathcal{G}$

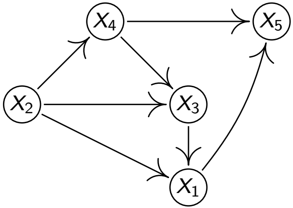
X_1	X_2	X_3	X_4	X_5
3.4	−0.3	5.8	−2.1	2.2
1.7	−0.2	7.0	−1.2	0.4
−2.4	−0.1	4.3	−0.7	3.5
2.3	−0.3	5.5	−1.1	−4.4
3.5	−0.2	3.9	−0.9	−3.9
⋮	⋮	⋮	⋮	⋮



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
⋮	⋮	⋮	⋮	⋮



Correlation (dependence) does not imply causation ... but RECALL:

Definition

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

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

Definition

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

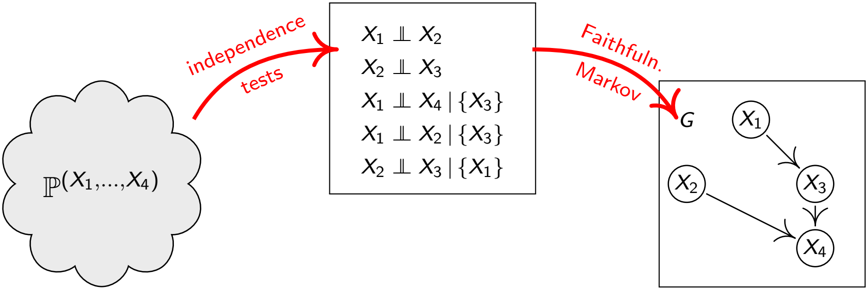
$$\underbrace{X \text{ and } Y \text{ are } d\text{-separated by } \mathcal{S} \text{ in } G}_{\text{properties of graph}} \quad \Rightarrow \quad \underbrace{X \perp\!\!\!\perp Y \,|\, \mathcal{S}}_{\text{properties in } P}$$

Definition

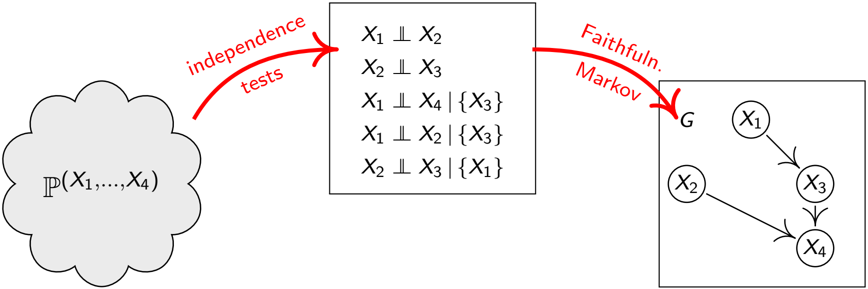
P satisfies faithfulness w.r.t. G if

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

Idea 1: independence-based methods



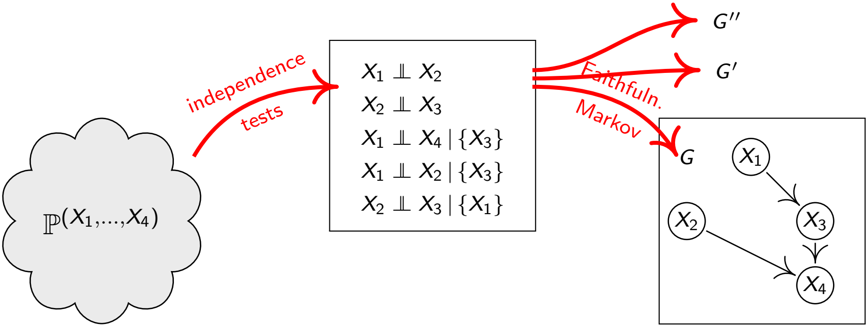
Idea 1: independence-based methods



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

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

Idea 1: independence-based methods



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

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

Idea 1: independence-based methods

Exercise 1: Assume that in $P^{(A,B,C)}$ we find

- $A \perp\!\!\!\perp C \mid B$
- $C \perp\!\!\!\perp A \mid B$
- no other (cond.) independence

Find all DAGs G s.t. $P^{(A,B,C)}$ is Markov and faithful wrt G.

Exercise 2: Assume that in $P^{(A,B,C,D)}$ we find

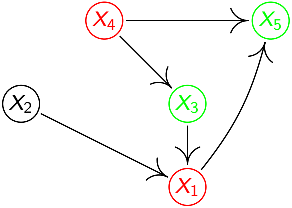
- $A \perp\!\!\!\perp B$
- $B \perp\!\!\!\perp A$
- no other (cond.) independence

Find all DAGs G s.t. $P^{(A,B,C,D)}$ is Markov and faithful wrt G.

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 $\mathcal{S}$.

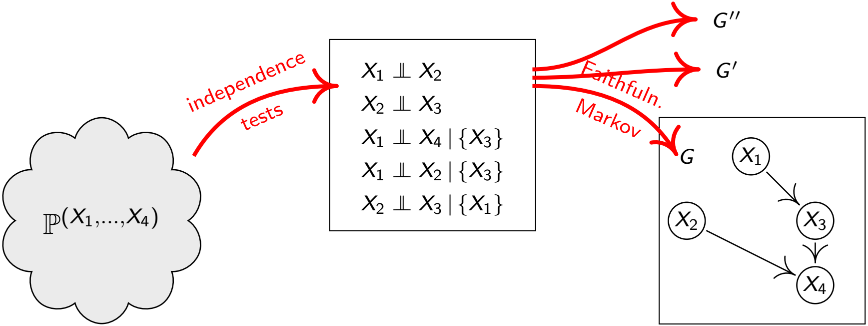
Check, whether all paths blocked!!



$\circ \cdots \rightarrow \circ \rightarrow \cdots \circ$	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 $\{X_3, X_5\}$

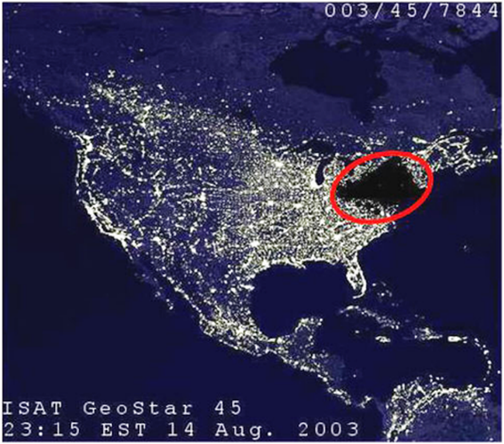
Idea 1: independence-based methods



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

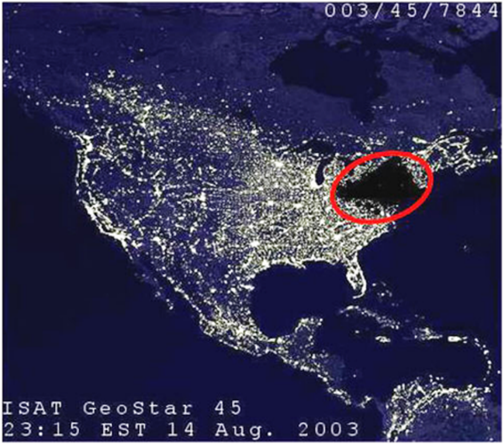
1. ~~Find all (cond.) independences from the data.~~ Be smart.
2. Select the DAG(s) that corresponds to these independences.

Watch out!



https://www.zdnet.com/pictures/no-april-fools-the-top-10-it-fiascos-of-all-time/10/

Watch out!



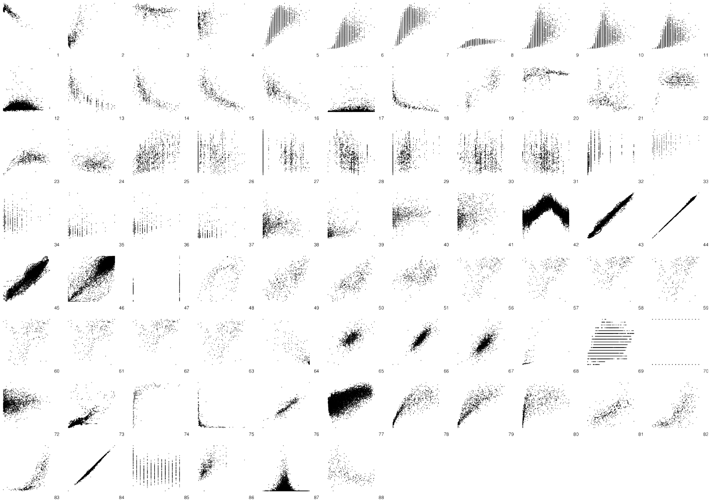
https://www.zdnet.com/pictures/no-april-fools-the-top-10-it-fiascos-of-all-time/10/

Any conditional independence test with correct level has NO POWER
(... if the cond. variable is continuous)

R. Shah and JP: "The hardness of Cond. Ind. Testing and the GCM", arxiv:1804.07203

https://www.visitdenmark.dk/da/kobenhavn/transport/sadan-kommer-du-rundt-i-kobenhavn

Idea 2: restricted structural causal models



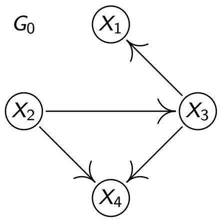
Mooij, JP, Janzing, Zscheischler, Schölkopf: *Disting. cause from effect using obs. data: methods and benchm.*, JMLR 2016

Idea 2: restricted structural causal models

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

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

- N_i jointly independent
- G_0 has no cycles



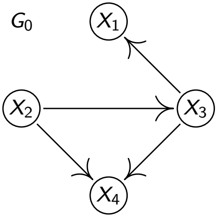
Structural equation model.
Can the DAG be recovered from $P(X_1, \ldots, X_4)$?

Idea 2: restricted structural causal models

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

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

- N_i jointly independent
- G_0 has no cycles



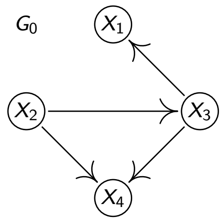
Structural equation model.
Can the DAG be recovered from $P(X_1, \ldots, X_4)$? **No.**

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



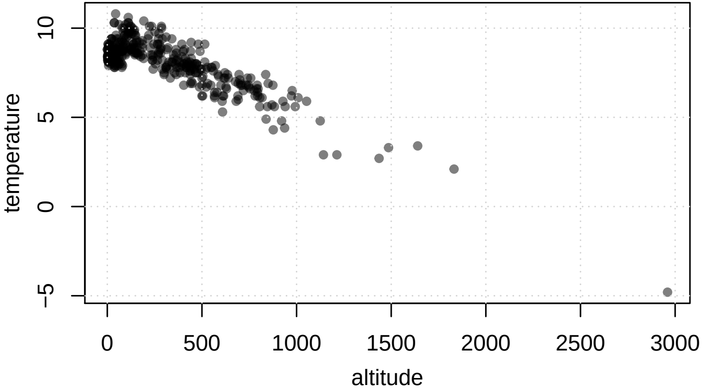
Additive noise model with Gaussian noise.

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

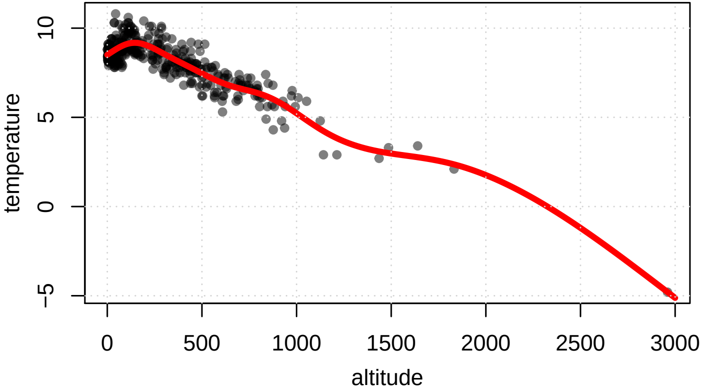
JP, J. Mooij, D. Janzing and B. Schölkopf: *Causal Discovery with Continuous Additive Noise Models, JMLR 2014*

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

Example: altitude and temperature



Example: altitude and temperature



p-value forward: 0.024

p-value backward: 0.0000000000019 (show code)

Idea 2: restricted structural causal models

d	number of DAGs with d nodes
1	1
2	3
3	25
4	543
5	29281
6	3781503
7	1138779265
8	783702329343
9	1213442454842881
10	4175098976430598143
11	31603459396418917607425
12	521939651343829405020504063
13	18676600744432035186664816926721
14	1439428141044398334941790719839535103
15	237725265553410354992180218286376719253505
16	83756670773733320287699303047996412235223138303
17	62707921196923889899446452602494921906963551482675201
18	99421195322159515895228914592354524516555026878588305014783
19	332771901227107591736177573311261125883583076258421902583546773505
20	2344880451051088988152559855229099188899081192234291298795803236068491263
21	34698768283588750028759328430181088222313944540438601719027559113446586077675521
22	1075822921725761493652956179327624326573727662809185218104090000500559527511693495107583

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

Idea 2: restricted structural causal models

Peters et al. (2009), Bauer et al. (2016):

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 structural causal models

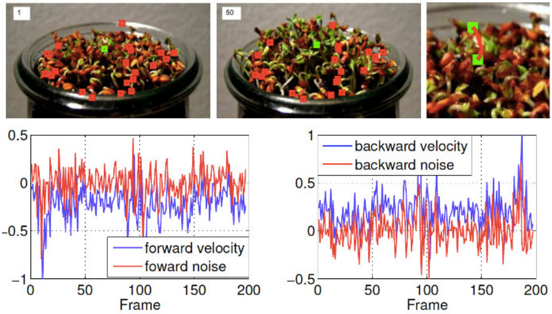
Pickup et al. (2014):

Idea 2: restricted structural causal models

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.

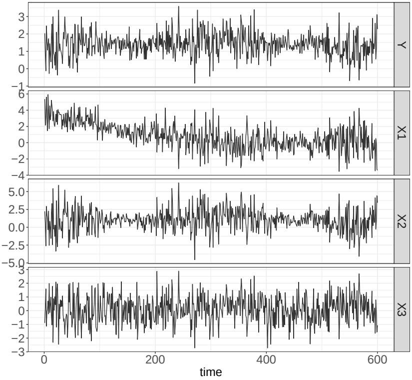
https://edition.cnn.com/travel/article/noma-under-the-bridge/index.html

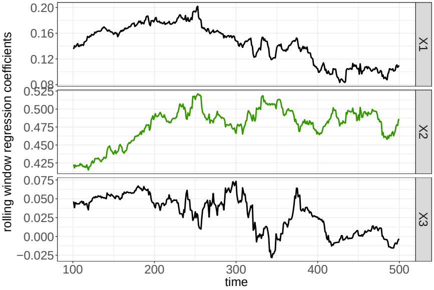
Idea 3: Invariant Causal Prediction

Suppose there is a target Y.

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

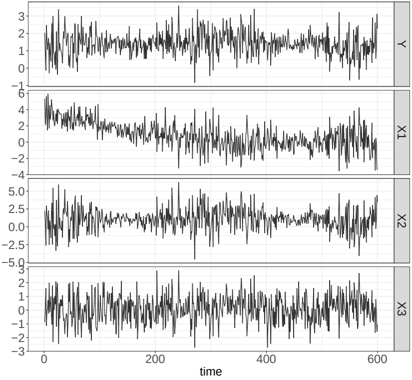
Regressing on (X_1, X_2, X_3):

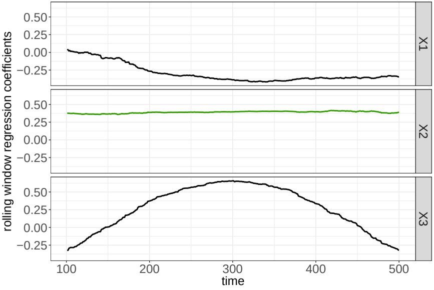


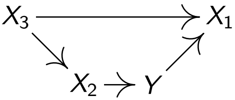


The coefficients change.

Regressing on X_1, X_2, and X_3:

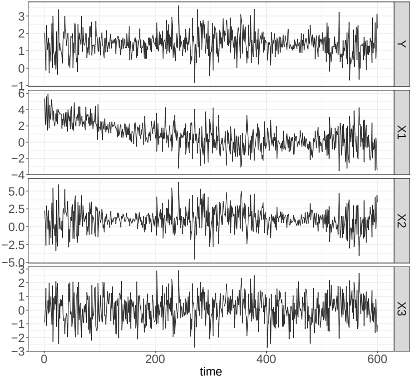


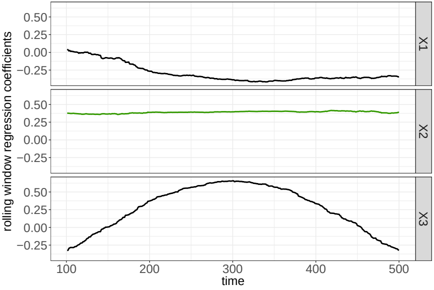




X_2 yields an invariant model. Ground truth:

Regressing on X_1, X_2, and X_3:





$$
\begin{tikzcd}
X_3 \arrow[rrr] \arrow[dr] & & & X_1 \\
& X_2 \arrow[r] & Y \arrow[ur] &
\end{tikzcd}
$$

X_2 yields an invariant model. Ground truth:

Relation invariance and causality: N. Pfister, P. Bühlmann, JP: *Invariant causal prediction for sequential data*, JASA (to appear)

Non-invariant models can be rejected if $\sqrt{\log n/n} = o(a_n)$.

Theorem (PBM 2016)

Assume invariance is satisfied for some S^*. For any test level α we obtain

$$P(\cap_{S:S \text{ invariant}} \subseteq S^*) \geq 1 - \alpha \,.$$

Theorem (PBM 2016)

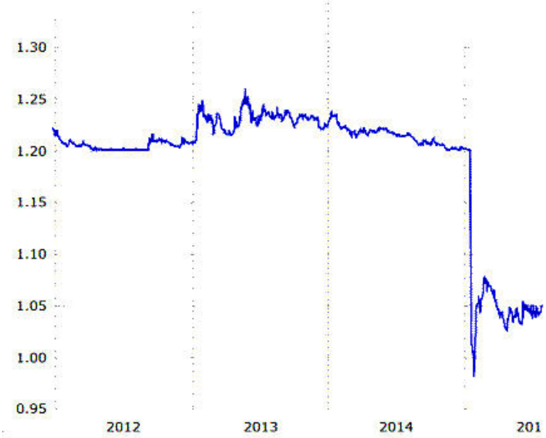
Assume invariance is satisfied for some S^*. For any test level α we obtain

$$P(\cap_{S:S \text{ invariant}} \subseteq S^*) \geq 1 - \alpha \,.$$

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: identification and confidence intervals*, JRSS-B 2016 (with discussion).

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



http://www.fremdwaehrungskonto.info/wp-content/uploads/2015/07/CHF-EUR-Kursentwicklung-2011-2015.gif monthly data Swiss National Bank Jan 1999 - Jan 2017

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

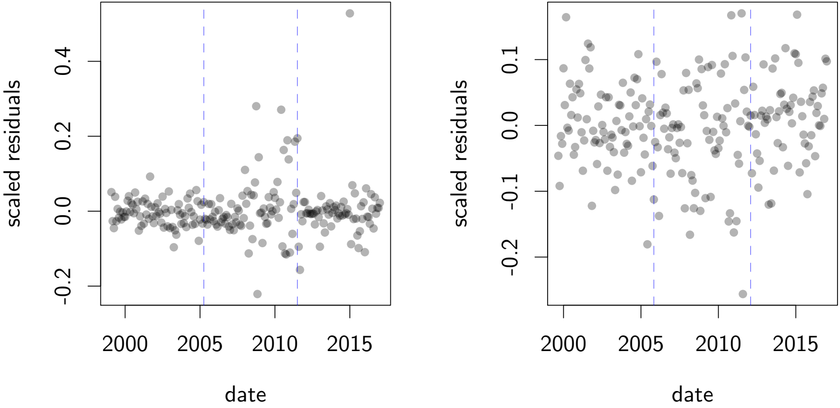
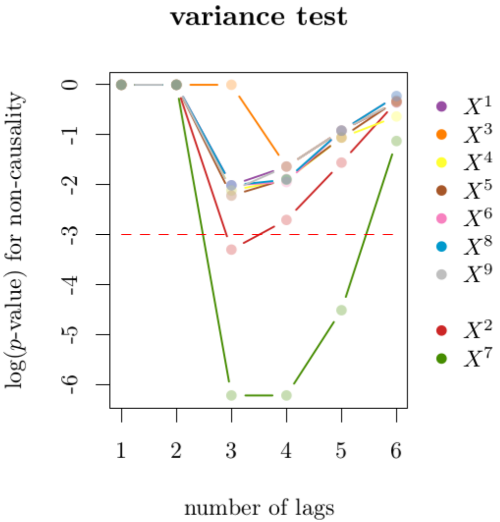
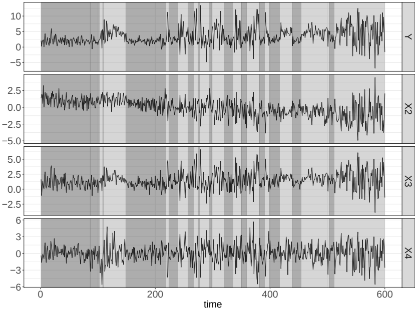


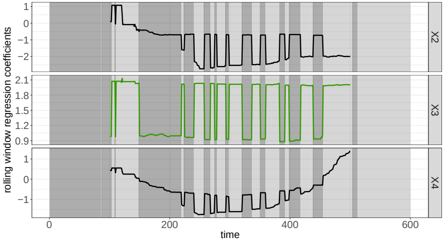
Figure: left plot (lowest p-value) and right plot (highest p-value) monthly data Swiss National Bank Jan 1999 - Jan 2017

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

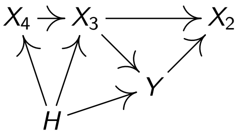

variance test
log(p-value) for non-causality
0
-1
-2
-3
-4
-5
-6
1
2
3
4
5
6
number of lags
X^1
X^3
X^4
X^5
X^6
X^8
X^9
X^2
X^7

This works for more complicated models, too (e.g., mixture models).





Here, X_3 yields an invariant model. Ground truth:



Relation invariance and causality: R. Christiansen, JP (in preparation)

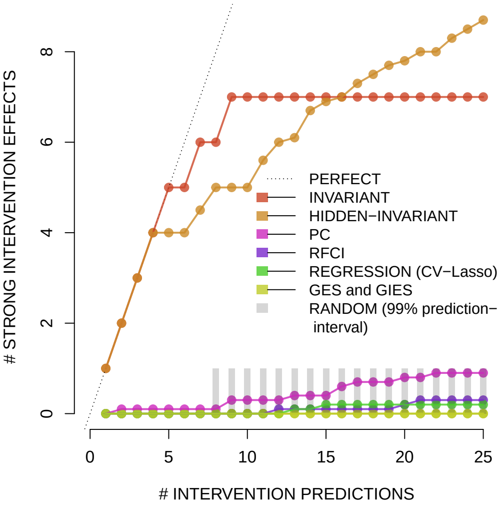
Here,

 $=$ time

Also possible:

 $=$ environments

Can we find causes of genes? (We have access to gene deletions.)



Peters et al. 2016, Kemmeren et al. 2014

ICP (R-package InvariantCausalPrediction)

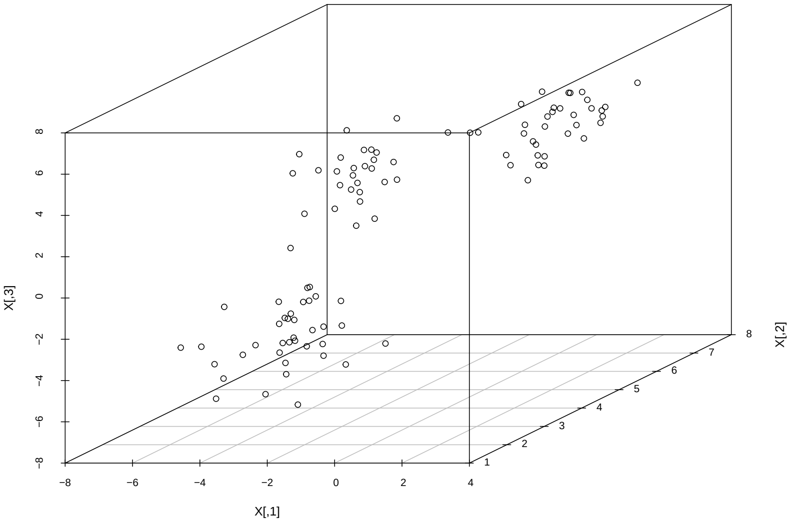
```
> ExpInd
```

[1]11111111111111111111111111111111...22222222222222...

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


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

What if we have no specific environment variable or time?



 = constructed by clustering?


DIE KUNST DES BRAUENS
BLONDE
BIERE
alc. 4.8% vol
50 cl
Mikis Wesensbitter

construc

 = non-descendant of target

Real data: Fertility Data UN Data. World population prospects, 2013

$Y \in \mathbb{R}$: total fertility rate in a country in a given year

$X \in \mathbb{R}^9$:

- IMR – infant mortality rate
- Q5 – under-five mortality rate
- Education expenditure (% of GNI)
- Exports of goods and services (% of GDP)
- GDP per capita (constant 2005 US$)
- GDP per capita growth (annual %)
- Imports of goods and services (% of GDP)
- Primary education (% female)
- Urban population (% of total)

$E \in \mathbb{R}$: continent of the country

Given $(X_1, Y_1, E_1), \ldots, (X_n, Y_n, E_n)$.

Invariance assumption $H_{0,S}$:

- $E \perp\!\!\!\perp Y \mid X_S$.
- $(X_1, Y_1, E_1), \ldots, (X_n, Y_n, E_n)$ are i.i.d.

Real data: Fertility Data UN Data. World population prospects, 2013

$Y \in \mathbb{R}$: total fertility rate in a country in a given year

$X \in \mathbb{R}^9$:

- IMR – infant mortality rate
- Q5 – under-five mortality rate
- Education expenditure (% of GNI)
- Exports of goods and services (% of GDP)
- GDP per capita (constant 2005 US$)
- GDP per capita growth (annual %)
- Imports of goods and services (% of GDP)
- Primary education (% female)
- Urban population (% of total)

$E \in \mathbb{R}$: continent of the country

Real data: Fertility Data UN Data. World population prospects, 2013

$Y \in \mathbb{R}$: total fertility rate in a country in a given year

$X \in \mathbb{R}^9$:

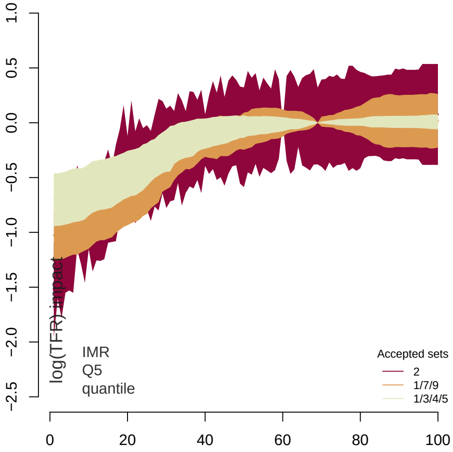
- IMR – infant mortality rate
- Q5 – under-five mortality rate
- Education expenditure (% of GNI)
- Exports of goods and services (% of GDP)
- GDP per capita (constant 2005 US$)
- GDP per capita growth (annual %)
- Imports of goods and services (% of GDP)
- Primary education (% female)
- Urban population (% of total)

$E \in \mathbb{R}$: continent of the country

$S_1 = \{$Q5$\}$
$S_2 = \{$IMR, Imports of goods and services, Urban pop. (% of total)$\}$
$S_3 = \{$IMR, Education expenditure (% of GNI), Exports of goods and services, GDP per capita$\}$

Heinze-Deml, JP and Meinshausen: Predicting the effect of interventions using inv. principles for nonl. models, arxiv:1706.08576

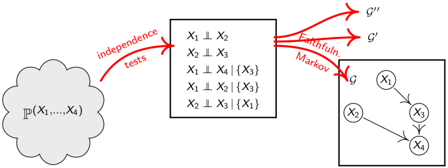
$$E\big[\log(\mathrm{TFR}) \mid do(X = x_{test})\big] - E\big[\log(\mathrm{TFR}) \mid do(X = x_{obs})\big]$$



Nigeria, 1993

Summary Part II:

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

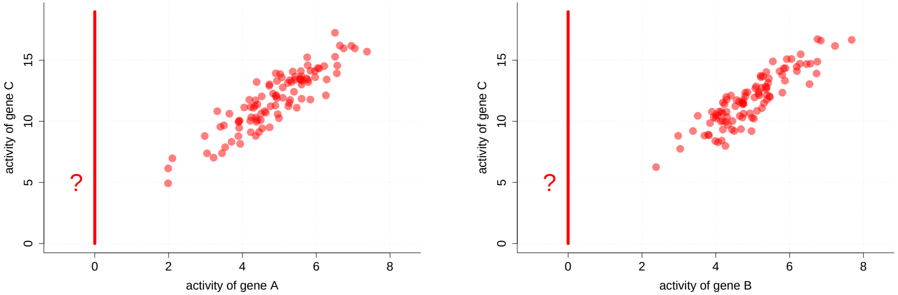


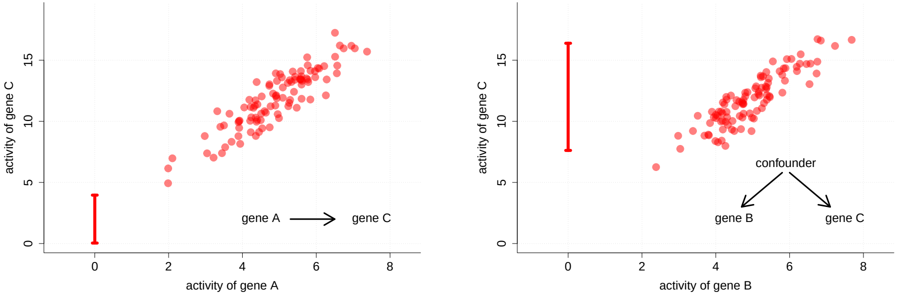
- Idea 2: additive noise (single environment)

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

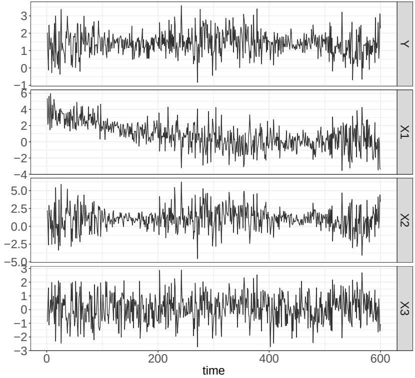
- Idea 3: invariant prediction (the more heterogeneity the better!)

Part III: Relations to Machine Learning


activity of gene C
?
activity of gene A
activity of gene C
?
activity of gene B


activity of gene C
activity of gene A
gene A
gene C
activity of gene C
activity of gene B
confounder
gene B
gene C

Recall:



The coefficients change.

Idea 1: anchor regression

What if there is more than one invariant model? Choose the best predictive model.

Idea 1: anchor regression

What if there is more than one invariant model? Choose the best predictive model.

anchor Regression

Idea 1: anchor regression

What if there is more than one invariant model? Choose the best predictive model.

anchor Regression

Find a trade-off between

- invariance with respect to 

AND

- predictive power

Idea 1: anchor regression

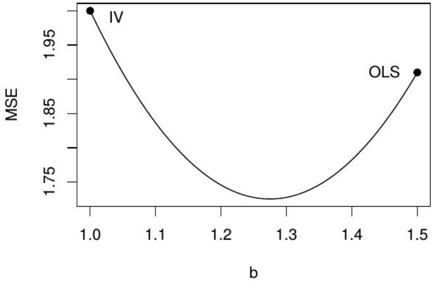
$$b^\gamma := \underset{b}{\operatorname{argmin}} \; \mathbf{E}(Y - Xb)^2 + \gamma \|\mathbf{E}A^t(Y - Xb)\|_2^2$$

Idea 1: anchor regression

$$b^\gamma := \underset{b}{\operatorname{argmin}} \; \mathbf{E}(Y - Xb)^2 + \gamma \|\mathbf{E}A^t(Y - Xb)\|_2^2$$

$$\begin{aligned} \gamma \to \infty: &\quad \text{IV} \\ \gamma \to 0: &\quad \text{OLS} \end{aligned}$$

MSE under a shift of X:

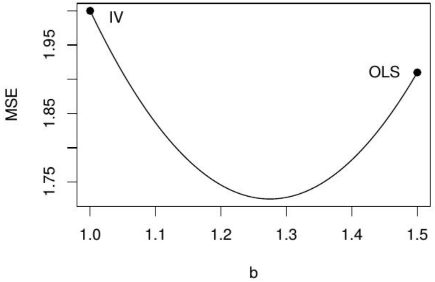


Idea 1: anchor regression

$$b^\gamma := \underset{b}{\operatorname{argmin}} \ \mathbf{E}(Y - Xb)^2 + \gamma \|\mathbf{E}A^t(Y - Xb)\|_2^2$$

$$\begin{array}{rc} \gamma \to \infty: & \text{IV} \\ \gamma \to 0: & \text{OLS} \end{array}$$

MSE under a shift of X:



Idea 1: anchor regression

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

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

Idea 1: anchor regression

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

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

Theorem

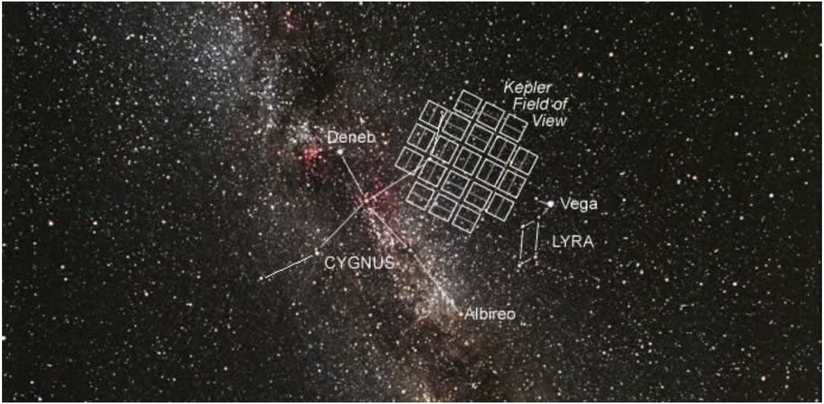
For any $b \in \mathbb{R}^d$ we have

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

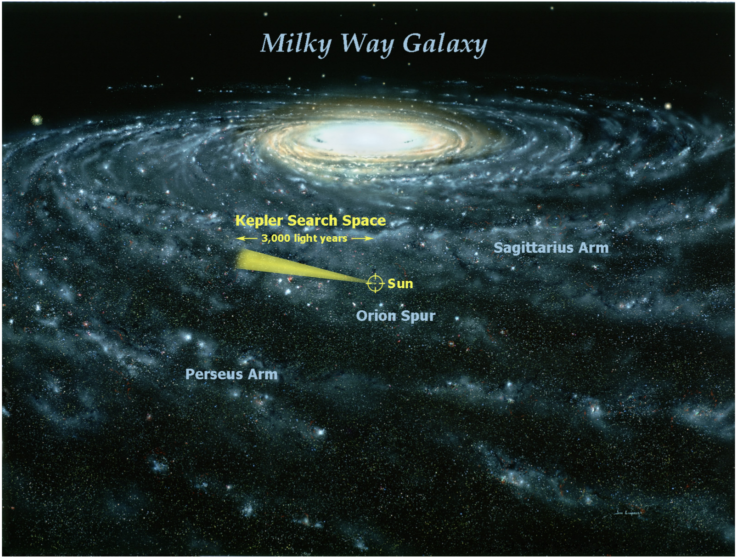
where

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

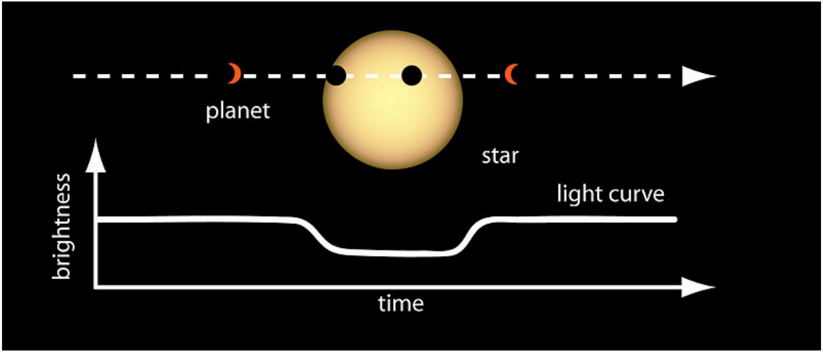
Idea 2: half-sibling regression



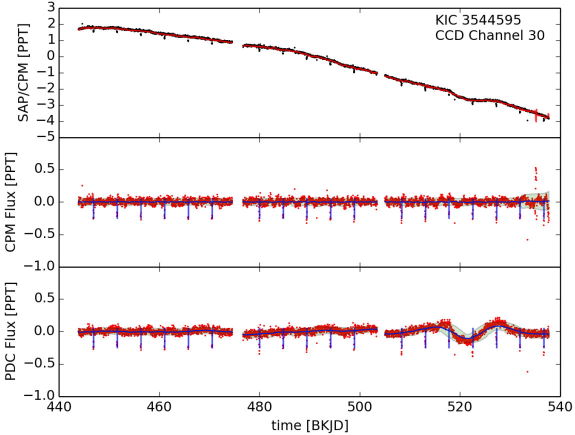
Idea 2: half-sibling regression



Idea 2: half-sibling regression



Idea 2: half-sibling regression

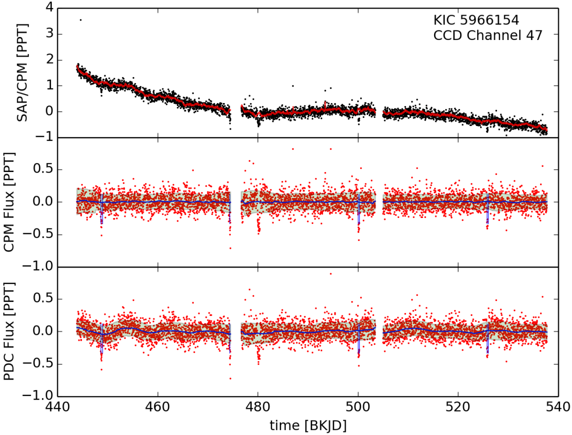


`http://archive.stsci.edu/index.html`

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 2: half-sibling regression

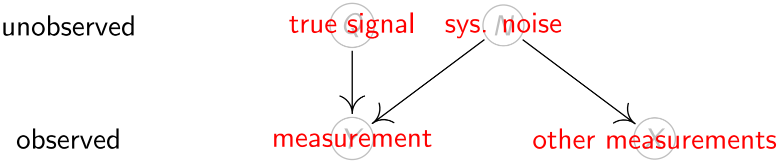


`http://archive.stsci.edu/index.html`

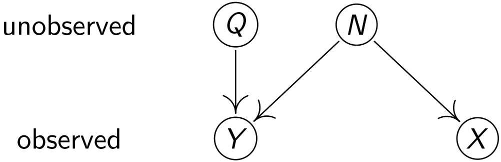
Schölkopf et al.: *Removing systematic errors for exoplanet search via latent causes*, ICML 2015

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

Idea 2: half-sibling regression

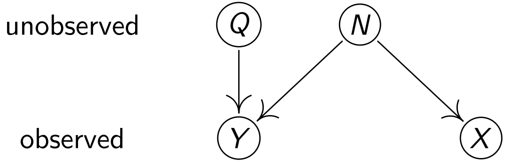


Idea 2: half-sibling regression



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

Idea 2: half-sibling regression

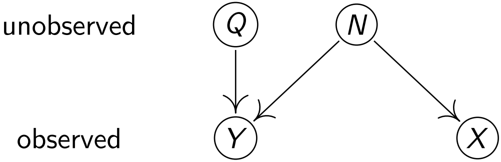


Assume $Y = f(N) + Q$.

Proposed idea:

Remove everything from Y explained by X.

Idea 2: half-sibling regression



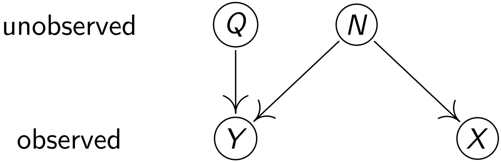
Assume $Y = f(N) + Q$.

Proposed idea:

Remove everything from Y explained by X.

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

Idea 2: half-sibling regression



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

Proposed idea:

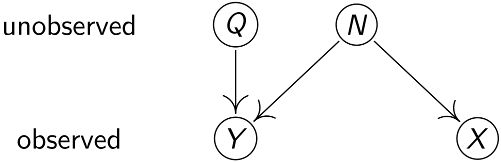
Remove everything from Y explained by X. Or: $\hat{Q} := Y - \mathbf{E}[Y \mid X]$.

Proposition

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

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

Idea 2: half-sibling regression



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

Proposed idea:

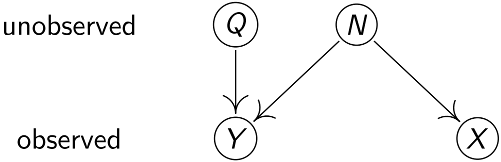
Remove everything from Y explained by X. Or: $\hat{Q} := Y - \mathbf{E}[Y \mid X]$.

Proposition

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

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

Idea 2: half-sibling regression



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

Proposed idea:

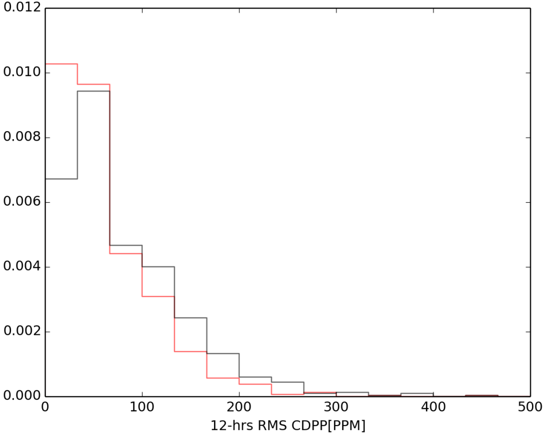
Remove everything from Y explained by X. Or: $\hat{Q} := Y - \mathbf{E}[Y \mid X]$.

Proposition

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

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

Idea 2: half-sibling regression



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

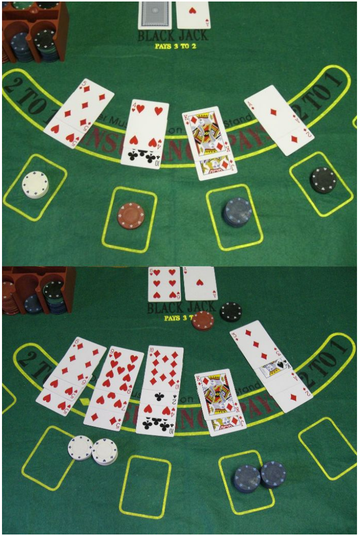
https://upload.wikimedia.org/wikipedia/commons/0/04/Nyhavn_copenhagen.jpg

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 $\rightarrow 1.5 \cdot$bet.

Idea 3: Blackjack



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

Idea 3: Blackjack

When can we learn?

Objects of Interest:

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

Idea 3: Blackjack
When can we learn?

Objects of Interest:

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

Questions:

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

Idea 3: Blackjack

When can we learn?

Objects of Interest:

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

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

Idea 3: Blackjack

Computation: Means

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

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

Idea 3: Blackjack

Computation: Means

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

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

Idea 3: Blackjack

Computation: Means

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

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

Estimate η by

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

Idea 3: Blackjack

Computation: Means

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

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

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

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

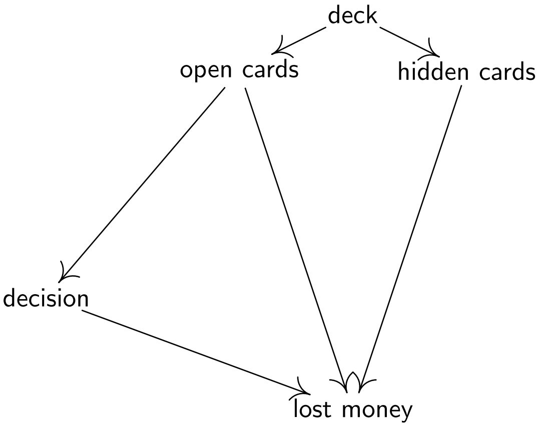
Which p^* is best? Parameterize and estimate

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

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, ...

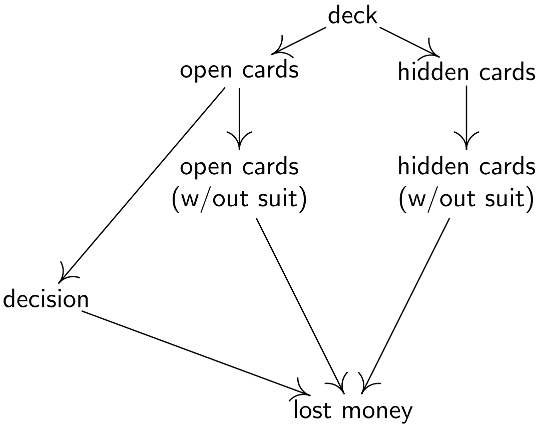
Idea 3: Blackjack

How to exploit causal structure:



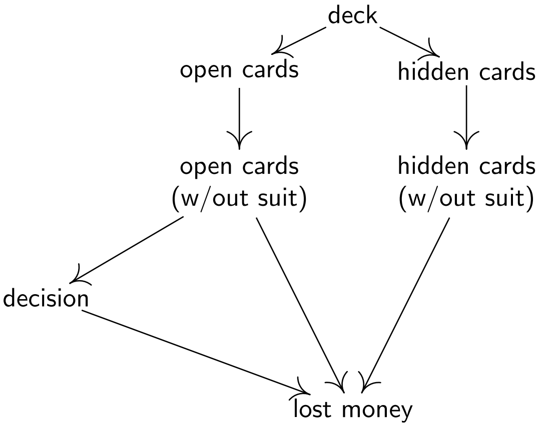
Idea 3: Blackjack

How to exploit causal structure:



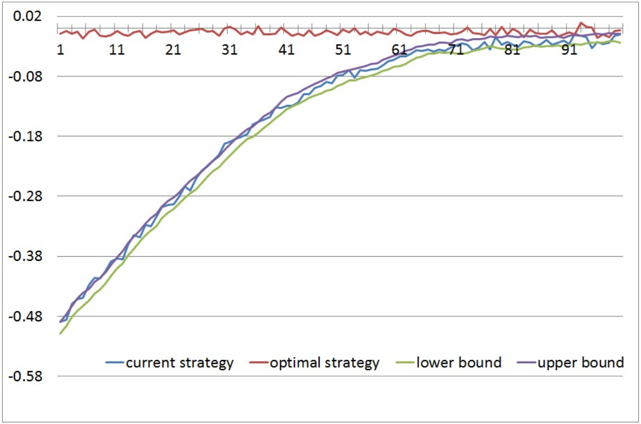
Idea 3: Blackjack

How to exploit causal structure:



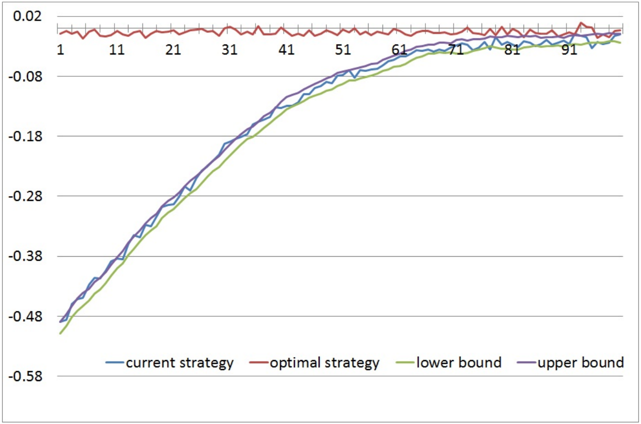
Idea 3: Blackjack

Learning Blackjack.



Idea 3: Blackjack

Learning Blackjack.



These ideas relate to IPW, RL, Horvitz-Thompson, ..., see

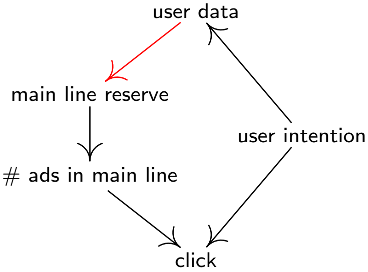
Bottou, JP et al.: *Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising*, JMLR 2013

Idea 3: Blackjack

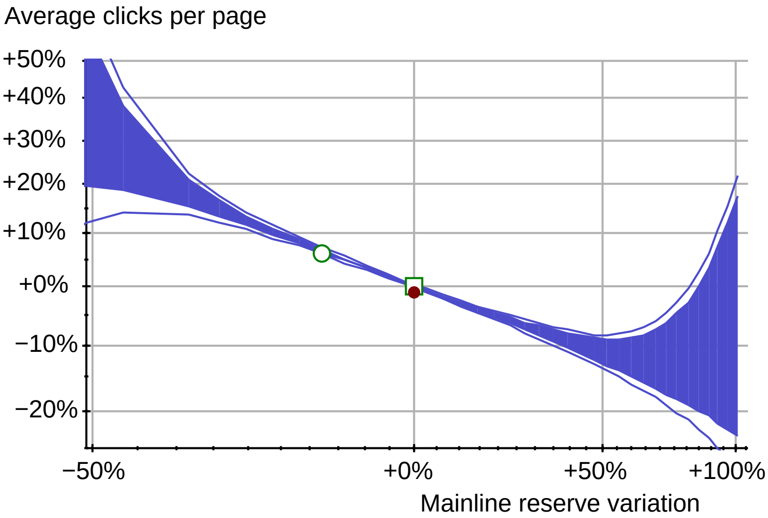
What can we do with $100,000$ samples?

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

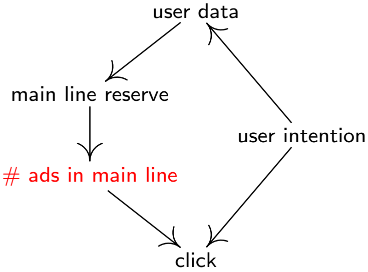
Idea 3: advertisement



Idea 3: advertisement

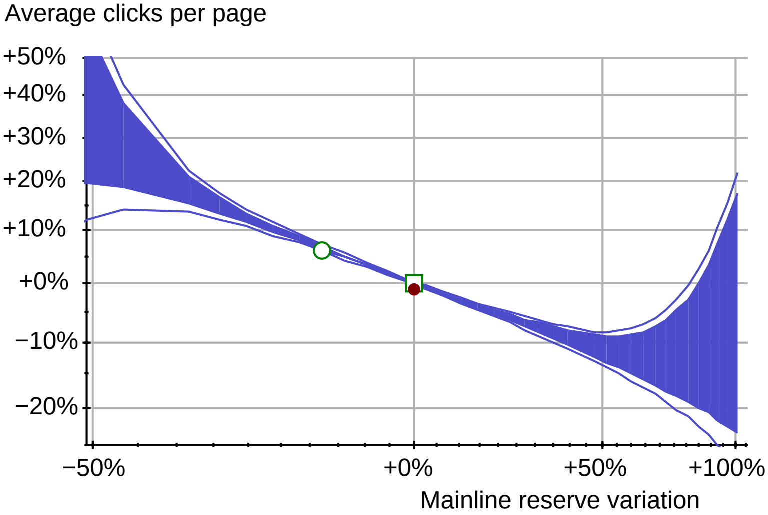


Idea 3: advertisement



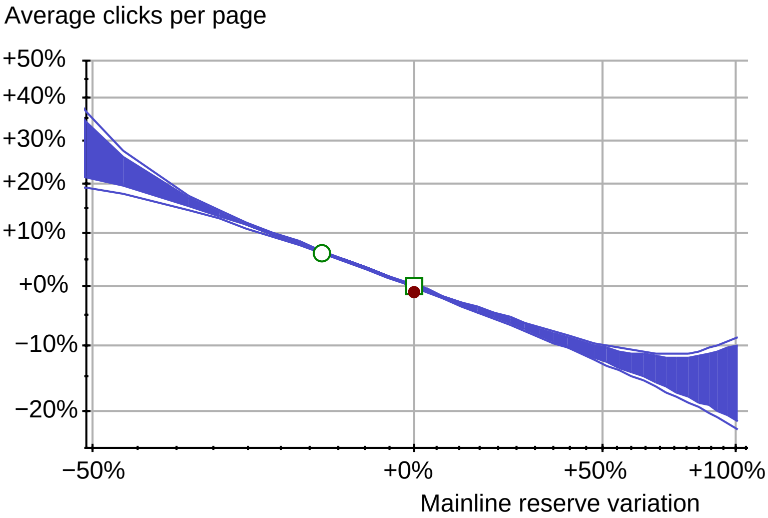
Idea 3: advertisement

Old:

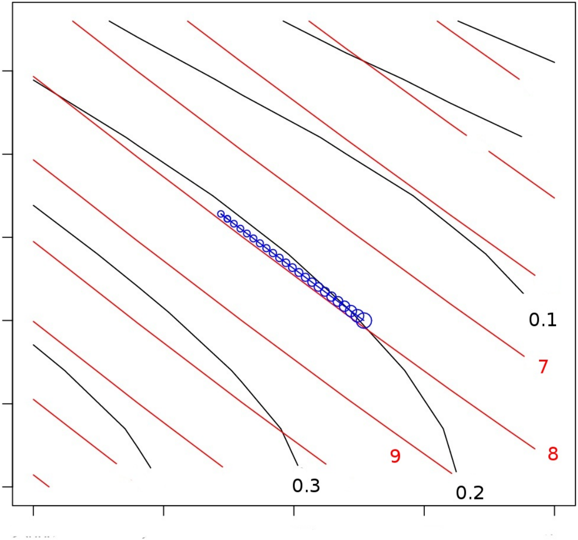


Idea 3: advertisement

Using discrete variable (ads shown in mainline):



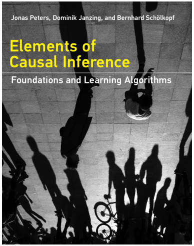
Optimization under constraints is possible, too



Summary Part III:

- Idea 1: anchor regression trades off prediction and invariance
- Idea 2: half-sibling regression
- Idea 3: reformulate reinforcement learning, use causal structure

Tusind tak!



New book: JP, D. Janzing, B. Schölkopf: *Elements of Causal Inference: Foundations and Learning Algorithms*

Check out our jupyter notebooks: `http://web.math.ku.dk/~peters/elements.html`

Idea 1: semi-supervised learning

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

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

Idea 1: semi-supervised learning

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

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

Modularity suggests:

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

Idea 1: semi-supervised learning

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

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

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:

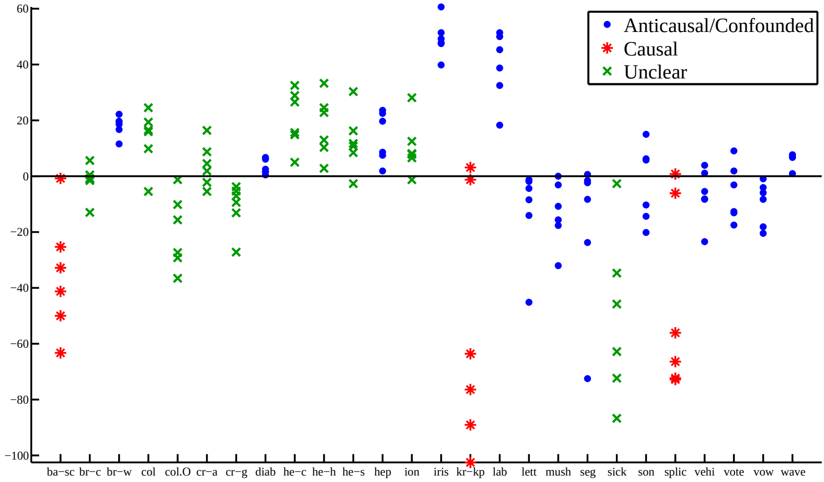
$$p(x_1 \,|\, x_{pa(1)}), \ldots, p(x_d \,|\, x_{pa(d)}) \text{ are "independent"}$$

Special case:

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

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