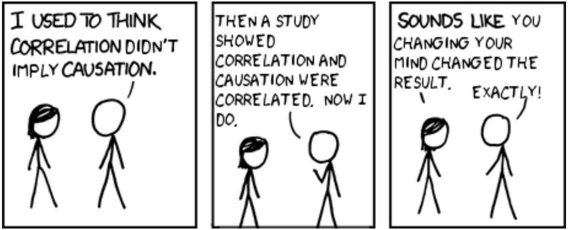
Causal Inference using Invariant Prediction

Jonas Peters
ETH Zürich

UCL, London

28th January 2015



ETH zürich

MARIE CURIE ACTIONS

contains joint work with ...

- ETH Zürich: Peter Bühlmann, Jan Ernest, Nicolai Meinshausen
- Max-Planck-Institute Tübingen: Dominik Janzing, Bernhard Schölkopf
- University of Amsterdam: Joris Mooij
- UC Berkeley: Sivaraman Balakrishnan, Martin Wainwright

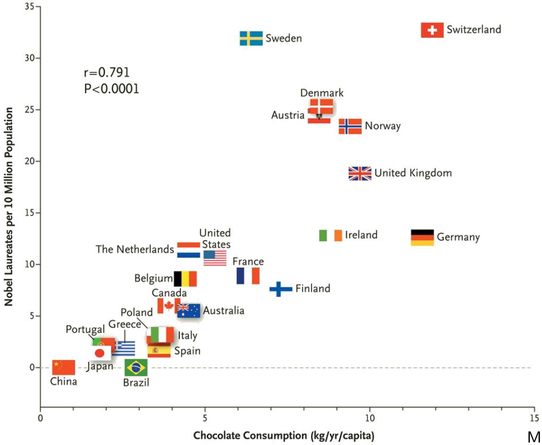
We often work with i.i.d. samples from a joint distribution P. Claim:

We often work with i.i.d. samples from a joint distribution P. Claim:

Many times, we are interested in a different distribution $\tilde{P} \neq P$.

We often work with i.i.d. samples from a joint distribution P. Claim:

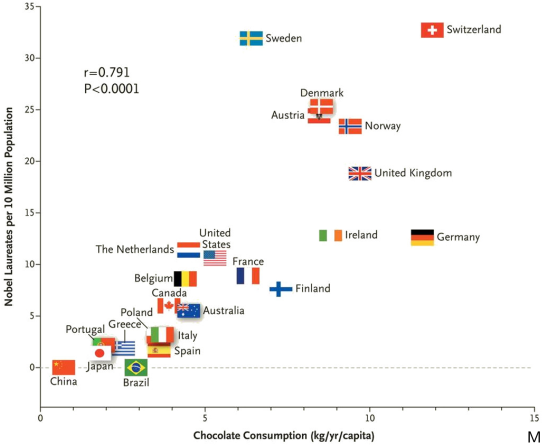
Many times, we are interested in a different distribution $\tilde{P} \neq P$.



Messerli, N Eng J Med 2012

We often work with i.i.d. samples from a joint distribution P. Claim:

Many times, we are interested in a different distribution $\tilde{P} \neq P$.

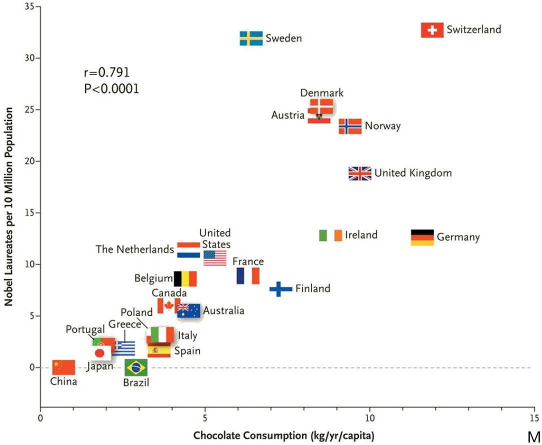


Messerli, N Eng J Med 2012

How do we relate P and $\tilde{P}$?

We often work with i.i.d. samples from a joint distribution P. Claim:

Many times, we are interested in a different distribution $\tilde{P} \neq P$.



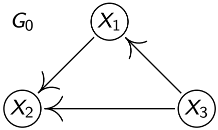
Messerli, N Eng J Med 2012

How do we relate P and $\tilde{P}$? **CAUSALITY!**

$P(X_1, \ldots, X_3)$ has been generated by a **structural equation model** if

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

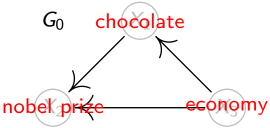
- N_i jointly independent
- G_0 has no cycles



Assume that we know the graph and P.

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

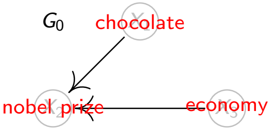
- N_i jointly independent
- G_0 has no cycles



Assume that we know the graph and P. We can then compute $\tilde{P}$.

$$X_1 = f_1(X_3, N_1) \quad X_1 = \tilde{f}_1(\tilde{N}_1)$$
$$X_2 = f_2(X_1, X_3, N_2)$$
$$X_3 = f_3(N_3)$$

- N_i jointly independent
- G_0 has no cycles



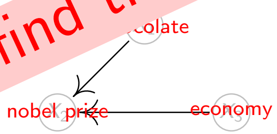
Assume that we know the graph and P. We can then compute $\tilde{P}$.

$$X_1 = f_1(X_3, N_1) \quad X_1 = \tilde{f}_1(\tilde{N}_1)$$

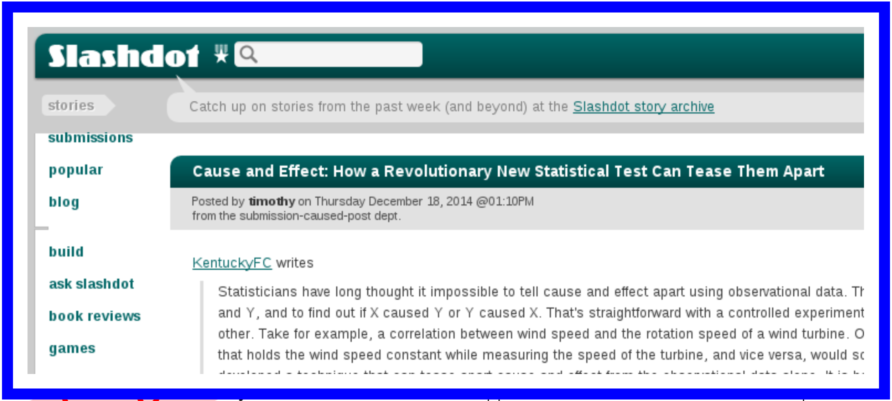
$$X_2 = f_2(X_1, X_3, N_2)$$

$$X_3 = f_3(N_3)$$

- N … ent
… cycles



Ok, but how do we find the graph?



Slashdot

stories

Catch up on stories from the past week (and beyond) at the Slashdot story archive

submissions

popular

blog

build

ask slashdot

book reviews

games

Cause and Effect: How a Revolutionary New Statistical Test Can Tease Them Apart

Posted by **timothy** on Thursday December 18, 2014 @01:10PM
from the submission-caused-post dept.

KentuckyFC writes

> Statisticians have long thought it impossible to tell cause and effect apart using observational data. Th
> and Y, and to find out if X caused Y or Y caused X. That's straightforward with a controlled experiment
> other. Take for example, a correlation between wind speed and the rotation speed of a wind turbine. O
> that holds the wind speed constant while measuring the speed of the turbine, and vice versa, would sc

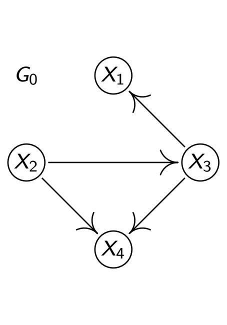
Ok,

Idea 1: Additive Noise

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



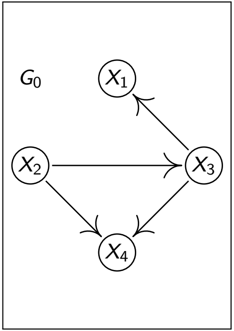
Structural equation model.
Can the DAG be recovered from $P(X_1, \ldots, X_4)$?

Idea 1: Additive Noise

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



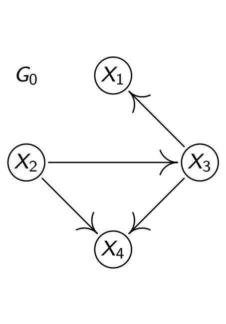
Structural equation model.
Can the DAG be recovered from $P(X_1, \ldots, X_4)$? **No.**

Idea 1: Additive Noise

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



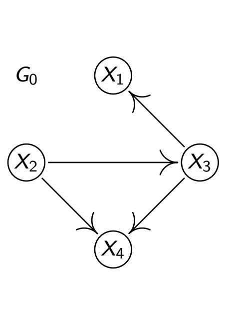
Additive noise model with Gaussian noise.

Idea 1: Additive Noise

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



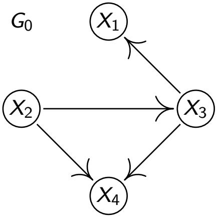
Additive noise model with Gaussian noise.
Can the DAG be recovered from $P(X_1, \ldots, X_4)$?

Idea 1: Additive Noise

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



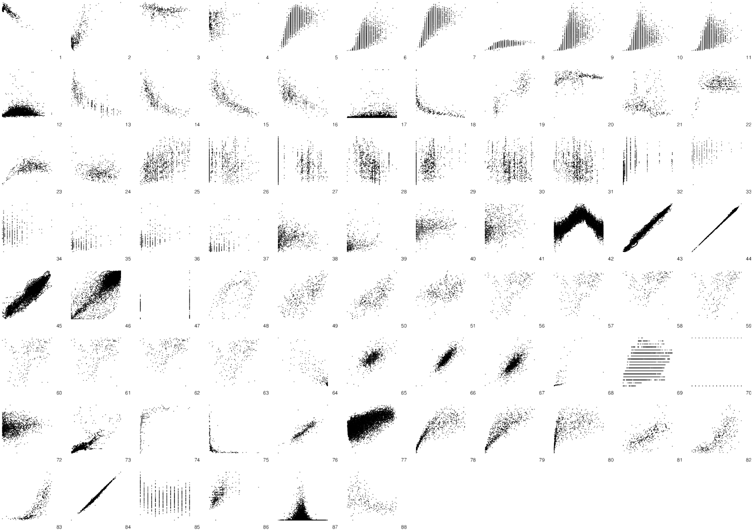
Additive noise model with Gaussian noise.

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

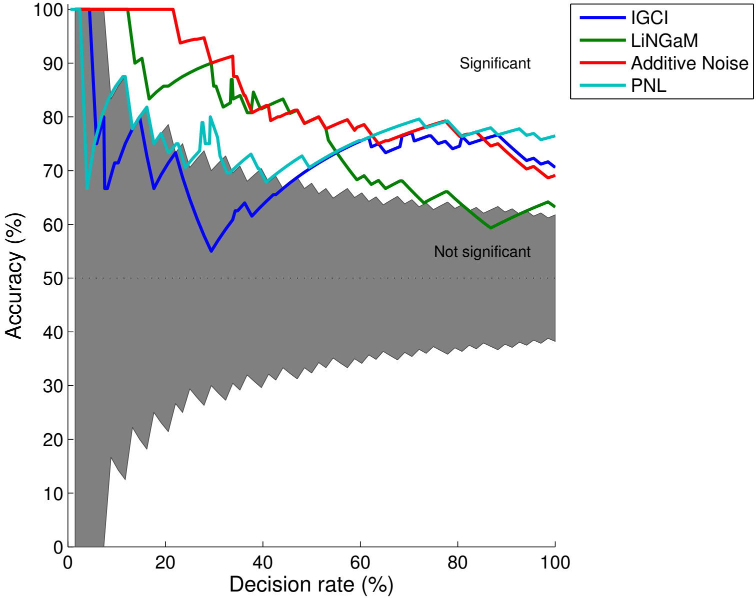
JP, J. Mooij, D. Janzing and B. Schölkopf: *Causal Discovery with Continuous Additive Noise Models, JMLR 2014*

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

Idea 1: Additive Noise



Idea 1: Additive Noise



Idea 1: Additive Noise


Massachusetts Institute of Technology

NEWS | VIDEO | SOCIAL | FOLLOW MIT

MIT News

ON CAMPUS AND AROUND THE WORLD

Browse or Search

FULL SCREEN

Illustration: Jose-Luis Olivares/MIT

Can we see the arrow of time?

Algorithm can determine, with 80 percent accuracy, whether video is running forward or backward.

Larry Hardesty | MIT News Office
June 20, 2014

Press Inquiries

RELATED

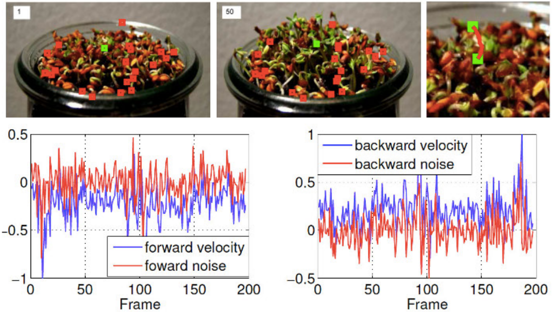
PAPER: "Seeing the arrow of time."

Einstein's theory of relativity envisions time as a spatial dimension, like height, width, and

Idea 1: Additive Noise

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.



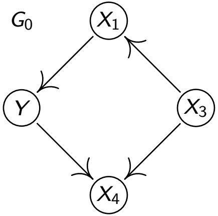
Top: tracked points from a sequence, and an example track. Bottom: Forward-time (left) and backward-time (right) vertical trajectory components, and the corresponding model residuals. Trajectories should be independent from model residuals (noise) in the forward-time direction only. For the example track 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: Observing data from different environments

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

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

- N_i jointly independent
- G_0 has no cycles

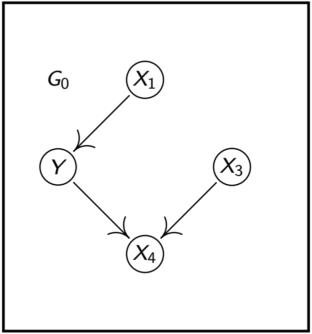


Idea 2: Observing data from different environments

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

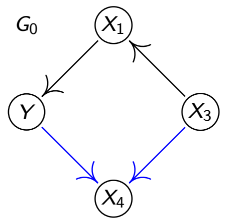


Idea 2: Observing data from different environments

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

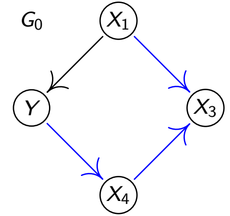


Idea 2: Observing data from different environments

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



Idea 2: Observing data from different environments

Observe target Y and p covariates X in different "environments" $e \in \mathcal{E}$:

$$(X^e, Y^e) \sim P^e \, .$$

Assumption

There exists (S^*, γ^*) *that satisfies property* $H_{0,\gamma,S}(\mathcal{E})$ $:\Leftrightarrow$
γ *vanishes outside* S *and*

$$Y^e = X^e\gamma + \varepsilon, \quad \varepsilon \perp\!\!\!\perp X^e_S.$$

Idea 2: Observing data from different environments

Observe target Y and p covariates X in different "environments" $e \in \mathcal{E}$:

$$(X^e, Y^e) \sim P^e .$$

Assumption

There exists (S^*, γ^*) *that satisfies property* $H_{0,\gamma,S}(\mathcal{E})$ $:\Leftrightarrow$
γ *vanishes outside* S *and*

$$Y^e = X^e \gamma + \varepsilon, \quad \varepsilon \perp\!\!\!\perp X^e_S.$$

Definition

- **Good set**: any set S such that $\exists \gamma$ with $H_{0,\gamma,S}$ is true.

Idea 2: Observing data from different environments

Observe target Y and p covariates X in different "environments" $e \in \mathcal{E}$:

$$(X^e, Y^e) \sim P^e\,.$$

Assumption

There exists (S^*, γ^*) *that satisfies property* $H_{0,\gamma,S}(\mathcal{E})$ $:\Leftrightarrow$
γ *vanishes outside* S *and*

$$Y^e = X^e \gamma + \varepsilon, \quad \varepsilon \perp\!\!\!\perp X_S^e.$$

Definition

- **Good set**: any set S such that $\exists\gamma$ with $H_{0,\gamma,S}$ is true.
- **Identifiable causal predictors** $S(\mathcal{E})$: vars appearing in all good sets.

Idea 2: Observing data from different environments

Observe target Y and p covariates X in different "environments" $e \in \mathcal{E}$:

$$(X^e, Y^e) \sim P^e \, .$$

Assumption

There exists (S^*, γ^*) *that satisfies property* $H_{0,\gamma,S}(\mathcal{E})$ $:\Leftrightarrow$
γ *vanishes outside* S *and*

$$Y^e = X^e \gamma + \varepsilon, \quad \varepsilon \perp\!\!\!\perp X^e_S.$$

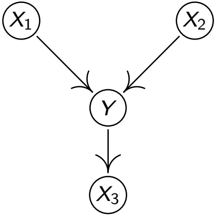
Definition

- **G** [illegible] ny set S such that $\exists \gamma$ with $H_{0,\gamma,S}$ is true.
- [illegible] **ble causal predictors** $S(\mathcal{E})$: vars appearing in all good sets.

Idea 2: Observing data from different environments

Example 1:

$$S^* = \{X_1, X_2\}, \quad \mathcal{E} = \{1\}$$

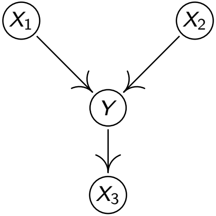


$e = 1$: observational

Idea 2: Observing data from different environments

Example 1:

$$S^* = \{X_1, X_2\}, \quad \mathcal{E} = \{1\}$$



$e = 1$: observational

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

Idea 2: Observing data from different environments

Example 2:

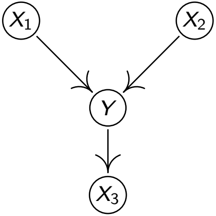
$$S^* = \{X_1, X_2\}, \quad \mathcal{E} = \{1, 2\}$$

$e = 1$: observational $\qquad$ $e = 2$: intervention on X_1

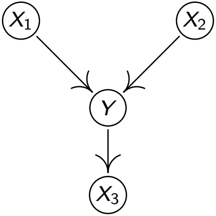
Idea 2: Observing data from different environments

Example 2:

$$S^* = \{X_1, X_2\}, \quad \mathcal{E} = \{1, 2\}$$



$e = 1$: observational



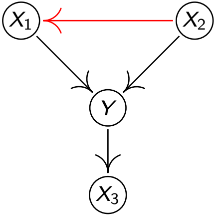
$e = 2$: intervention on X_1

$$S(\mathcal{E}) = \{X_1\}$$

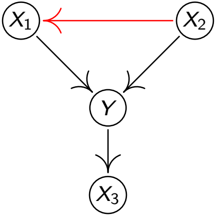
Idea 2: Observing data from different environments

Example 3:

$$S^* = \{X_1, X_2\}, \quad \mathcal{E} = \{1, 2\}$$



$e = 1$: observational



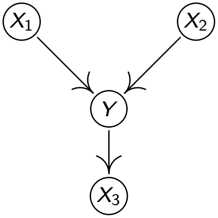
$e = 2$: intervention on X_1

$$S(\mathcal{E}) = \{X_1, X_2\}$$

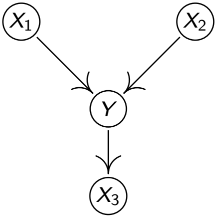
Idea 2: Observing data from different environments

Example 4:

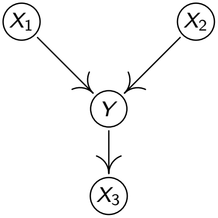
$$S^* = \{X_1, X_2\}, \quad \mathcal{E} = \{1, 2, 3\}$$



$e = 1$: obs.



$e = 2$: interv. on X_1

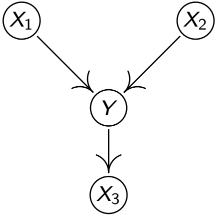


$e = 3$: interv. on X_2

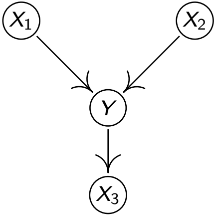
Idea 2: Observing data from different environments

Example 4:

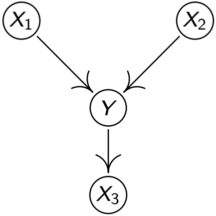
$$S^* = \{X_1, X_2\}, \quad \mathcal{E} = \{1, 2, 3\}$$



$e = 1$: obs.



$e = 2$: interv. on X_1



$e = 3$: interv. on X_2

$$S(\mathcal{E}) = \{X_1, X_2\}$$

Idea 2: Observing data from different environments

Theorem

1. *No mistakes:*

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

Idea 2: Observing data from different environments

Theorem

1. *No mistakes:*

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

2. *No chances with one environment:*

$$\#\mathcal{E} = 1 \implies S(\mathcal{E}) = \emptyset.$$

Idea 2: Observing data from different environments

Theorem

1. *No mistakes:*
$$S(\mathcal{E}) \subseteq S^*.$$

2. *No chances with one environment:*
$$\#\mathcal{E} = 1 \implies S(\mathcal{E}) = \emptyset.$$

3. *Seeing more environments helps:*
$$S(\mathcal{E}_1) \subseteq S(\mathcal{E}_2) \quad \textit{if} \quad \mathcal{E}_1 \subseteq \mathcal{E}_2$$

Idea 2: Observing data from different environments

Theorem

1. *No mistakes:*
$$S(\mathcal{E}) \subseteq S^*.$$
2. *No chances with one environment:*
$$\#\mathcal{E} = 1 \implies S(\mathcal{E}) = \emptyset.$$
3. *Seeing more environments helps:*
$$S(\mathcal{E}_1) \subseteq S(\mathcal{E}_2) \quad \text{if} \quad \mathcal{E}_1 \subseteq \mathcal{E}_2$$
4. *Sufficient conditions for $S(\mathcal{E}) = S^*$:*
   a) *many "generic" interventions: on each node except Y OR*
   b) *single "generic" intervention: on a "youngest" parent of Y that is directly connected to all other parents of Y.*

JP, P. Bühlmann, N. Meinshausen: *Causal inference using invariant prediction: identification and conf. interv.*, arXiv 1501.01332

Idea 2: Observing data from different environments

Method for finite samples: construct $\hat{\Gamma}(\mathcal{E}), \hat{S}(\mathcal{E})$ by testing for $H_{0,\gamma,S}$.

Idea 2: Observing data from different environments

Method for finite samples: construct $\hat{\Gamma}(\mathcal{E}), \hat{S}(\mathcal{E})$ by testing for $H_{0,\gamma,S}$.

Theorem

Assume that the probability of falsely rejecting H_{0,γ^*,S^*} is less than α. Then

Idea 2: Observing data from different environments

Method for finite samples: construct $\hat{\Gamma}(\mathcal{E}), \hat{S}(\mathcal{E})$ by testing for $H_{0,\gamma,S}$.

Theorem

Assume that the probability of falsely rejecting H_{0,γ^*,S^*} is less than α. Then

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

Idea 2: Observing data from different environments

Method for finite samples: construct $\hat{\Gamma}(\mathcal{E}), \hat{S}(\mathcal{E})$ by testing for $H_{0,\gamma,S}$.

Theorem

Assume that the probability of falsely rejecting H_{0,γ^*,S^*} is less than α. Then

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

- **possible test 1**: test whether the regression models are the same for group e and $-e$ (Chow, 1960) (inversion of cov. matrix of Res_e)

Idea 2: Observing data from different environments

Method for finite samples: construct $\hat{\Gamma}(\mathcal{E}), \hat{S}(\mathcal{E})$ by testing for $H_{0,\gamma,S}$.

Theorem

Assume that the probability of falsely rejecting H_{0,γ^*,S^*} is less than α. Then

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

- **possible test 1**: test whether the regression models are the same for group e and $-e$ (Chow, 1960) (inversion of cov. matrix of Res_e)
- **possible test 2**: linear regression on pooled data; test whether Res_e of group e have same mean and variance as Res_{-e}.

Idea 2: Observing data from different environments

Method for finite samples: construct $\hat{\Gamma}(\mathcal{E}), \hat{S}(\mathcal{E})$ by testing for $H_{0,\gamma,S}$.

Theorem

Assume that the probability of falsely rejecting H_{0,γ^*,S^*} is less than α. Then

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

- **possible test 1**: test whether the regression models are the same for group e and $-e$ (Chow, 1960) (inversion of cov. matrix of Res_e)
- **possible test 2**: linear regression on pooled data; test whether Res_e of group e have same mean and variance as Res_{-e}.
- (easy) tricks for computations and high-dimensions.

JP, P. Bühlmann, N. Meinshausen: *Causal inference using invariant prediction: identification and conf. interv.*, arXiv 1501.01332

Idea 2: Observing data from different environments

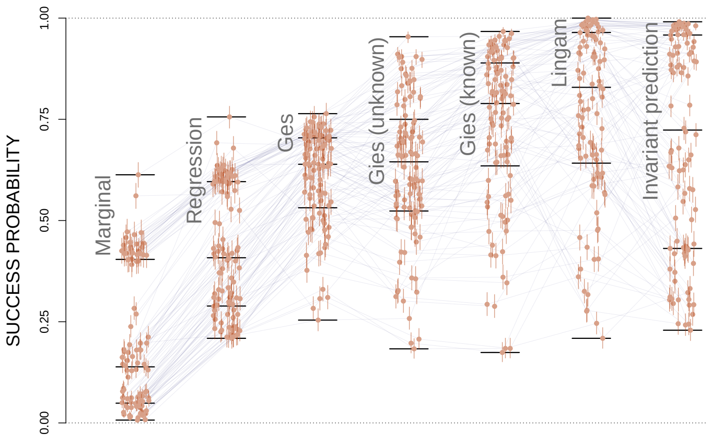
Simulations: 100 settings, 1000 data sets each.

How often do we find $\hat{S}(\mathcal{E}) = S^*$?

Idea 2: Observing data from different environments

Simulations: 100 settings, 1000 data sets each.

How often do we find $\hat{S}(\mathcal{E}) = S^*$?



Idea 2: Observing data from different environments

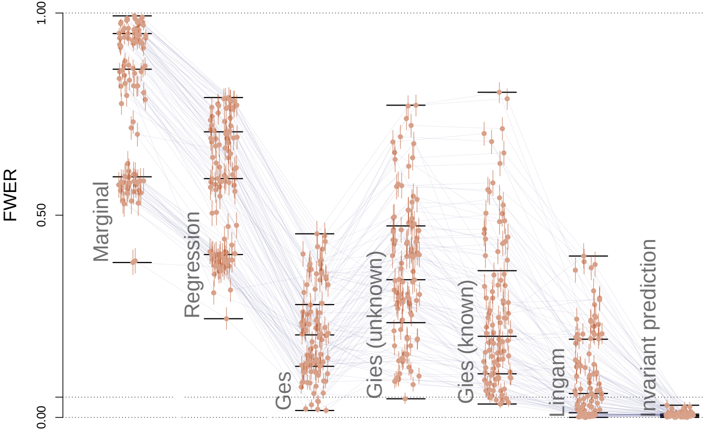
Simulations: 100 settings, 1000 data sets each.

How often do we find $\hat{S}(\mathcal{E}) \not\subseteq S^*$?

Idea 2: Observing data from different environments

Simulations: 100 settings, 1000 data sets each.

How often do we find $\hat{S}(\mathcal{E}) \not\subseteq S^*$?



Idea 2: Observing data from different environments

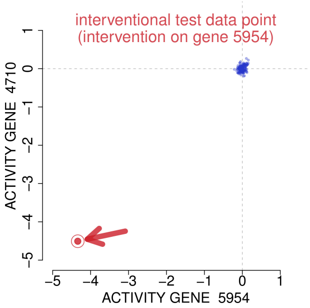
Real data: genetic perturbation experiments for yeast (Kemmeren et al., 2014)

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

Idea 2: Observing data from different environments

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

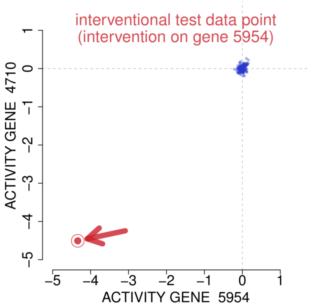
- $p = 6170$ genes
- $n_{obs} = 160$ wild-types
- $n_{int} = 1479$ gene deletions (targets known)
- true hits: $\approx 9\%$ of pairs



Idea 2: Observing data from different environments

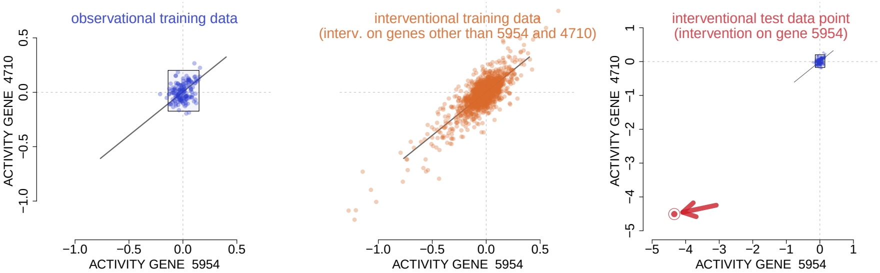
Real data: genetic perturbation experiments for yeast (Kemmeren et al., 2014)

- $p = 6170$ genes
- $n_{obs} = 160$ wild-types
- $n_{int} = 1479$ gene deletions (targets known)
- true hits: $\approx 9\%$ of pairs



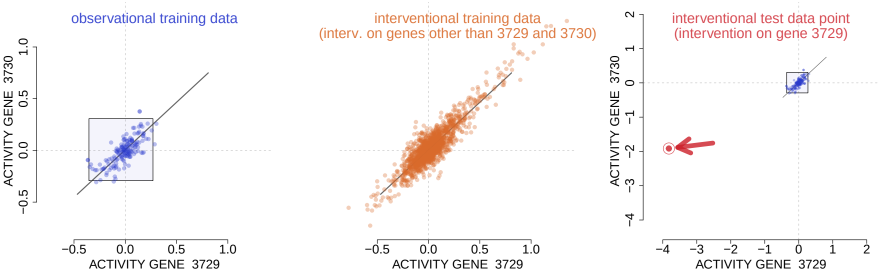
- our method: $\mathcal{E} = \{obs, int\}$

Idea 2: Observing data from different environments



most significant pair

Idea 2: Observing data from different environments



2nd most significant pair

Idea 2: Observing data from different environments

3rd most significant pair

Idea 2: Observing data from different environments

	proposed method	GIES	IDA	marginal corr.		random guessing
				observ.	pooled	
# of true positives (out of 8)	6	2	2	1	2	2 (95% quantile) 3 (99% quantile) 4 (99.9% quantile)

Summary:

- Idea 1: additive noise (single environment)
- Idea 2: invariant prediction (multiple environments); control family wise error rate

Summary:

- Idea 1: additive noise (single environment)
- Idea 2: invariant prediction (multiple environments); control family wise error rate

Future directions (theory):

- extend to estimation of graphs
- combine ideas 1 and 2
- nonlinear models
- finite sample guarantees

Summary:

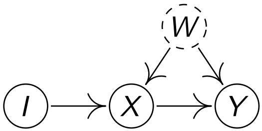
- Idea 1: additive noise (single environment)
- Idea 2: invariant prediction (multiple environments); control family wise error rate

Future directions (theory):

- extend to estimation of graphs
- combine ideas 1 and 2
- nonlinear models
- finite sample guarantees

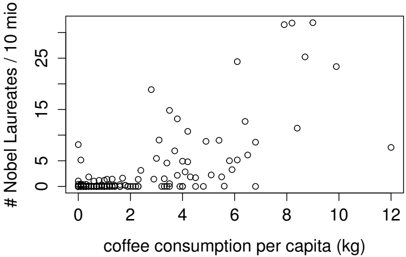
Future work (methodology):

- domain adaption
- instrumental variables



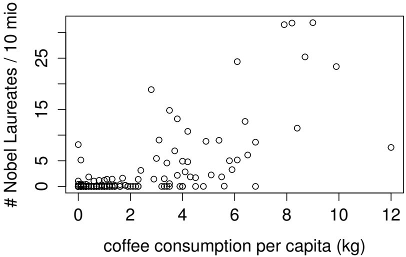
Dankeschön!

Does X cause Y or vice versa?



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

Does X cause Y or vice versa?

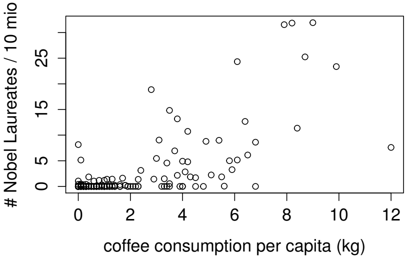


Correlation: 0.698
p-value: $< 2.2 \cdot 10^{-16}$

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?

Does X cause Y or vice versa?



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?