Causality – in a wide sense
Lecture III

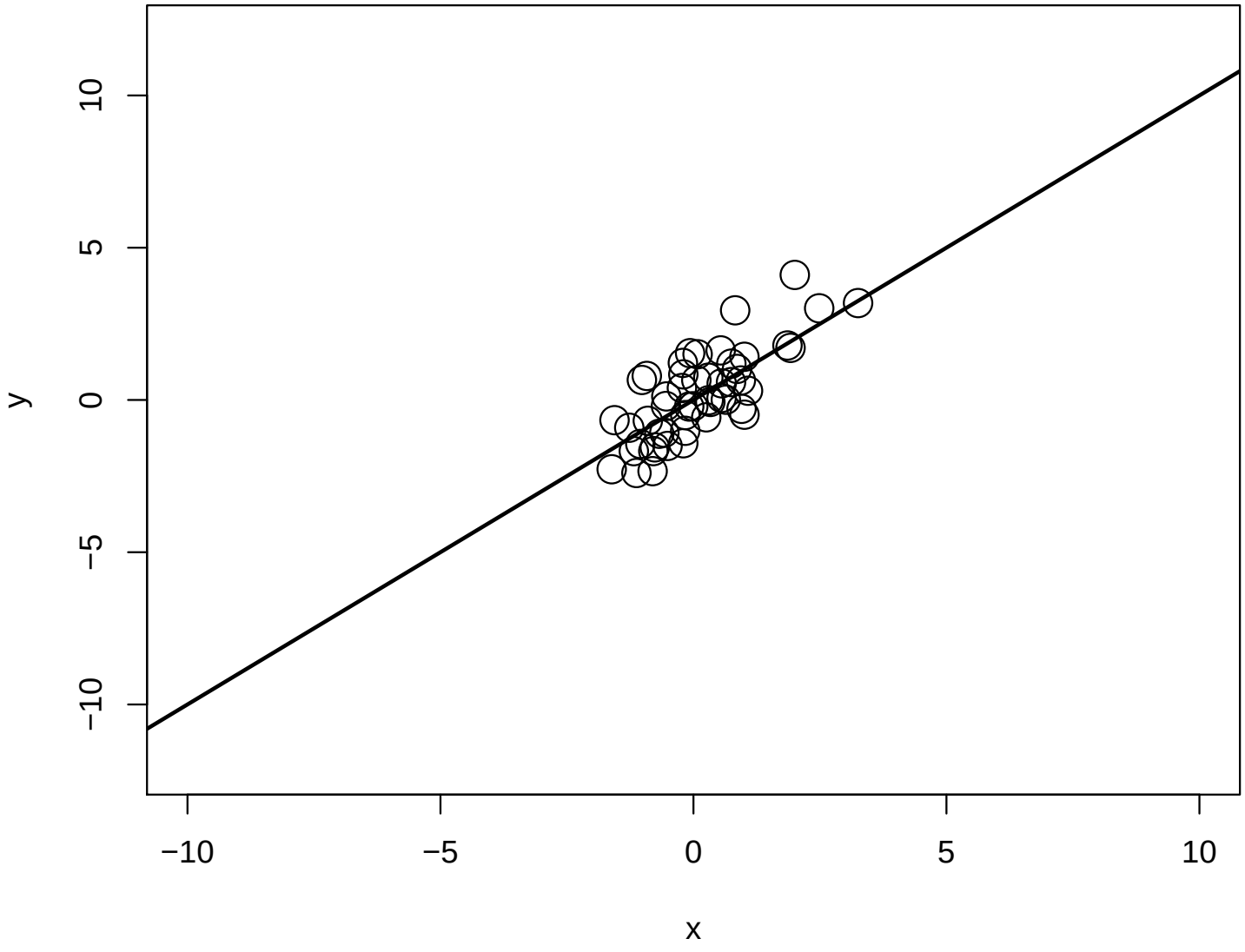
Peter Bühlmann

Seminar for Statistics
ETH Zürich

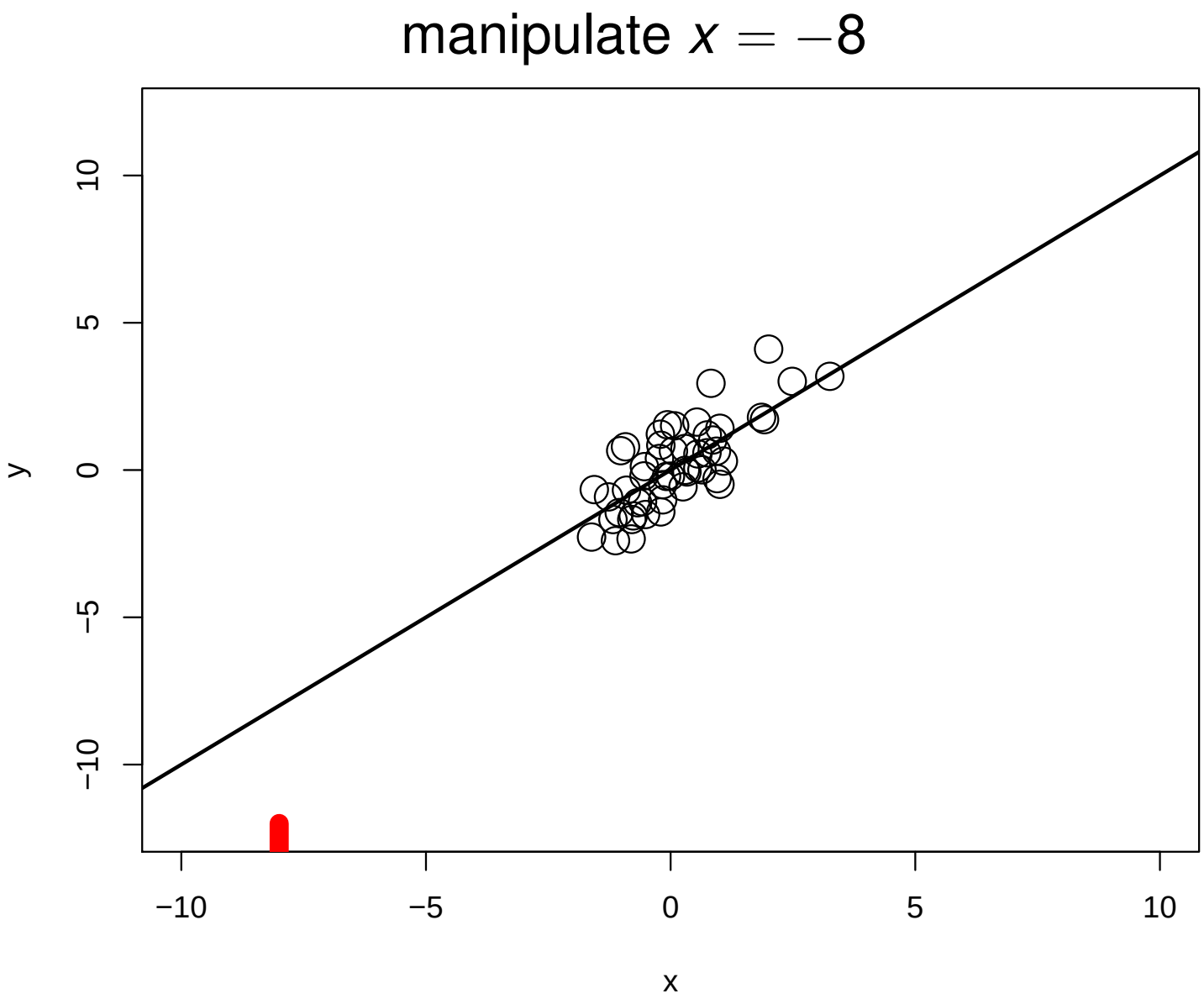
Recap from yesterday

- causality is giving a prediction to an intervention/manipulation

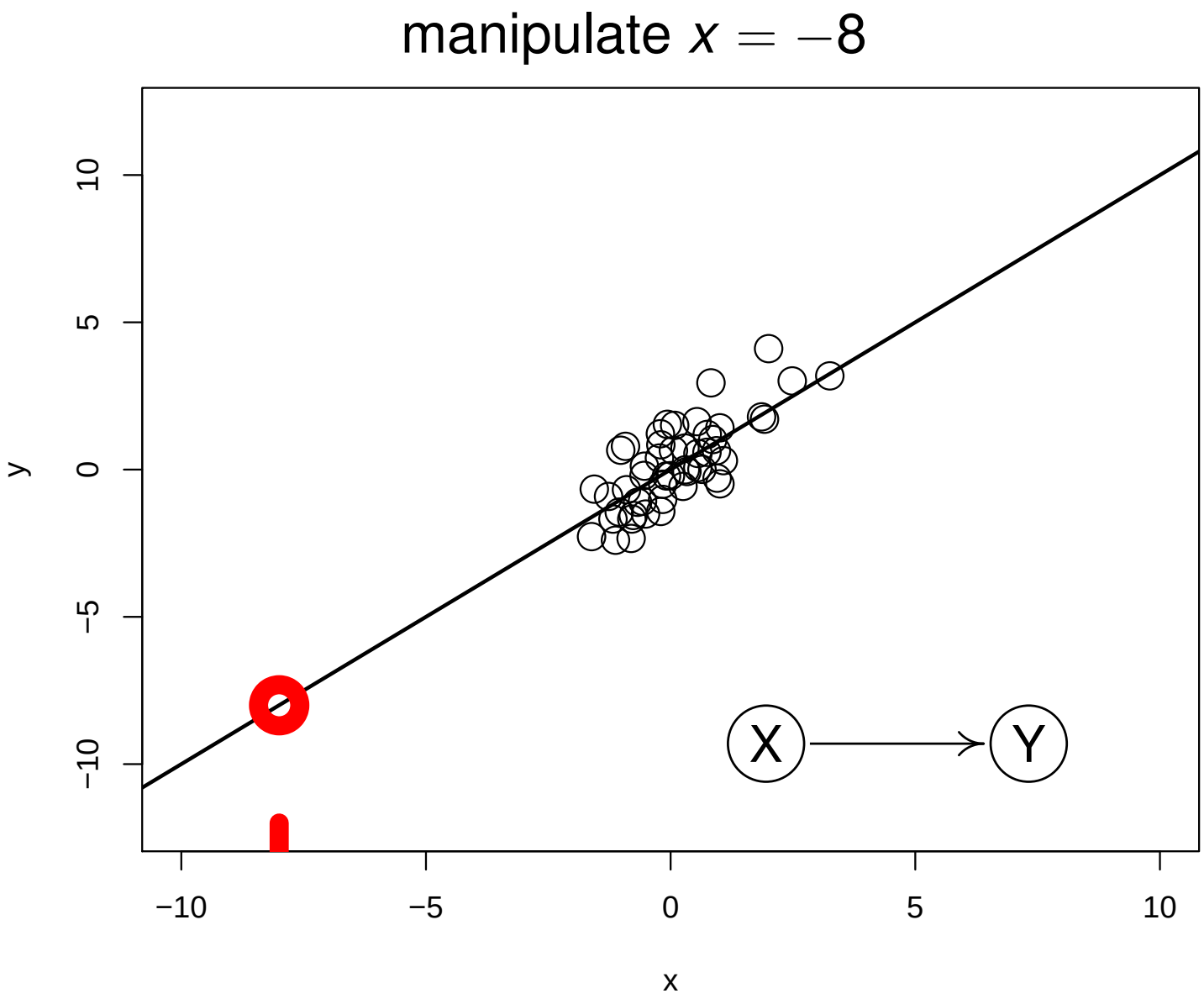
Predicting a potential outcome



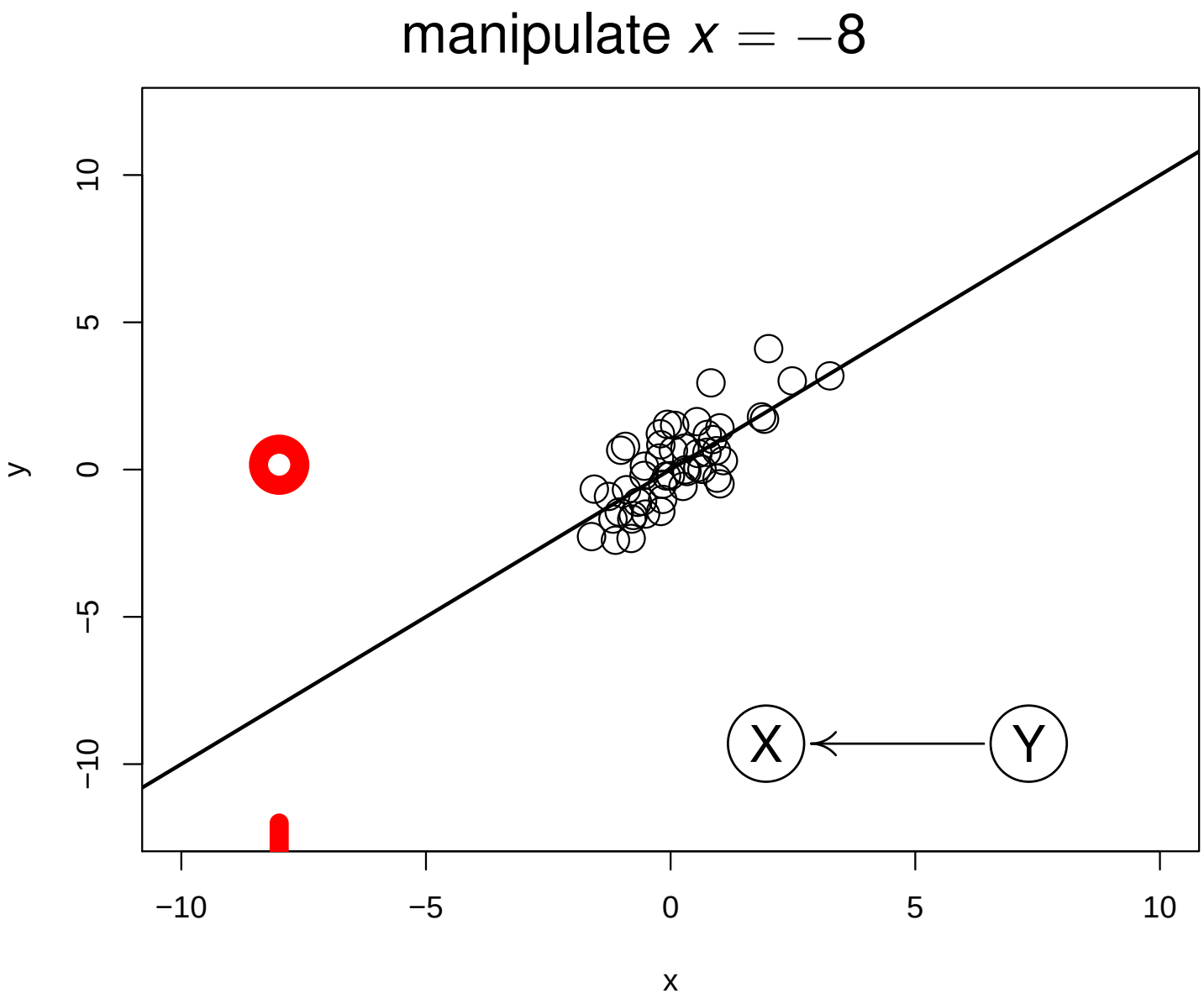
Predicting a potential outcome



It's an ambitious problem



It's an ambitious problem



- observational data plus interventional data is much more informative than observational data alone
- do-intervention model is simple, easy to understand but often too specific: we often cannot intervene precisely at single variables

Invariant Causal Prediction

Invariance Assumption (w.r.t. $\mathcal{E}$)

there exists $S^* \subseteq \{1, \ldots, d\}$ such that:

$$\mathcal{L}(Y^e | X^e_{S^*}) \text{ is invariant across } e \in \mathcal{E}$$

for linear model setting:
there exists a vector γ^* with $\text{supp}(\gamma^*) = S^* = \{j;\ \gamma^*_j \neq 0\}$
such that:

$$\forall e \in \mathcal{E}: \quad Y^e = X^e \gamma^* + \varepsilon^e,\ \varepsilon^e \perp X^e_{S^*}$$

$\varepsilon^e \sim F_\varepsilon$ the same for all e

X^e has an arbitrary distribution, different across e

Invariant Causal Prediction

Invariance Assumption (w.r.t. $\mathcal{F}$)

there exists $S^* \subseteq \{1, \ldots, d\}$ such that:

$$\mathcal{L}(Y^e | X^e_{S^*}) \text{ is invariant across } e \in \mathcal{F}$$

for linear model setting:
there exists a vector γ^* with $\text{supp}(\gamma^*) = S^* = \{j;\ \gamma^*_j \neq 0\}$
such that:

$$\forall e \in \mathcal{F}: \quad Y^e = X^e \gamma^* + \varepsilon^e,\ \varepsilon^e \perp X^e_{S^*}$$

$\varepsilon^e \sim F_\varepsilon$ the same for all e

X^e has an arbitrary distribution, different across e

if $e \in \mathcal{F}$

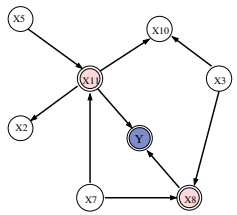
- does not directly affect Y
- does not change the relation between X and Y

then:
$S_{\text{causal}} = \text{pa}(Y)$ satisfy Invariance Assumption w.r.t. $\mathcal{F}$

$$\text{causal structure/variables} \implies \text{invariance}$$

The search for invariance and causality (Peters, PB & Meinshausen, 2016)

causal structure/variables $\Longleftarrow$ invariance



severe issues of identifiability !

can perform statistical test whether a subset S of covariates satisfies the invariance assumption

$H_0 - \mathrm{InvA}(\mathcal{E})$: $\mathcal{L}(Y^e | X_S^e)$ is invariant across $e \in \underbrace{\mathcal{E}}_{\text{observed environments}}$

in a linear model $\rightsquigarrow$ Chow (1960)

$\rightsquigarrow$ sets $S_1, \ldots, S_k$ which are statistically compatible with invariance assumption $H_0 - \mathrm{InvA}(\mathcal{E})$

making it identifiable:

$$\hat{S}(\mathcal{E}) = \bigcap \{S;\ S \underbrace{\text{ statistically compatible with } H_0 - \text{InvA}(\mathcal{E})\}}_{\text{no rejection at significance level } \alpha}$$

Theorem: (Peters, PB and Meinshausen, 2016)
assume structural equation model

- linear model for Y versus X, Gaussian errors
- $e \in \mathcal{E}$ does not act directly on Y and
  does not change the relation between X and Y

Then:

$$\mathbb{P}[\hat{S}(\mathcal{E}) \subseteq \underbrace{S_{\text{causal}}}_{\text{pa}(Y)}] \geq 1 - \alpha$$

confidence guarantee against false positive causal selection

making it identifiable:

$$\hat{S}(\mathcal{E}) = \bigcap \{S;\ S \underbrace{\text{ statistically compatible with } H_0 - \text{InvA}(\mathcal{E})}_{\text{no rejection at significance level } \alpha}\}$$

Theorem: (Peters, PB and Meinshausen, 2016)
assume structural equation model

- linear model for Y versus X, Gaussian errors
- $e \in \mathcal{E}$ does not act directly on Y and
  does not change the relation between X and Y

Then:

$$\mathbb{P}[\hat{S}(\mathcal{E}) \subseteq \underbrace{S_{\text{causal}}}_{\text{pa}(Y)}] \geq 1 - \alpha$$

confidence guarantee against false positive causal selection

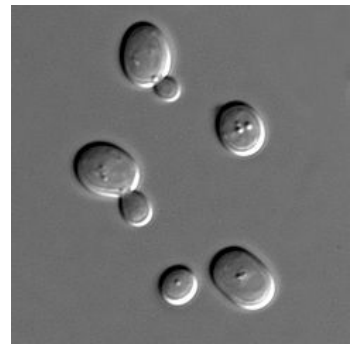
ICP = Invariant Causal Prediction

Proof:
note that the causal set S_{causal} leads to invariance

$$\begin{aligned}
\mathbb{P}[\hat{S}(\mathcal{E}) \subseteq S_{\text{causal}}] &= \mathbb{P}[\bigcap\{S; H_{0,S} \text{ not rejected}\} \subseteq S_{\text{causal}}] \\
&\geq \mathbb{P}[H_{0,S_{\text{causal}}} \text{ not rejected}] \geq 1 - \alpha
\end{aligned}$$

□

Single gene deletion experiments in yeast



$d = 6170$ genes
response of interest: $Y =$ expression of first gene
"covariates" $X =$ gene expressions from all other genes

and then
response of interest: $Y =$ expression of second gene
"covariates" $X =$ gene expressions from all other genes

and so on

infer/predict the effects of unseen/new single gene deletions on all other genes

Kemmeren et al. (2014):
genome-wide mRNA expressions in yeast: $d = 6170$ genes

- $n_{obs} = 160$ "observational" samples of wild-types
- $n_{int} = 1479$ "interventional" samples
  each of them corresponds to a single gene deletion strain

for our method: we use $|\mathcal{E}| = 2$
(observational and interventional data)

training-test data splitting:

- training set: all observational and 2/3 of interventional data
- test set: other 1/3 of gene deletion interventions
  $\rightsquigarrow$ can validate predicted effects of these interventions
- repeat this for the three blocks of interventional test data

multiplicity adjustment:

since ICP is used 6170 times (once for every response var.) we use coverage $1 - \alpha/6170$ with $\alpha = 0.05$

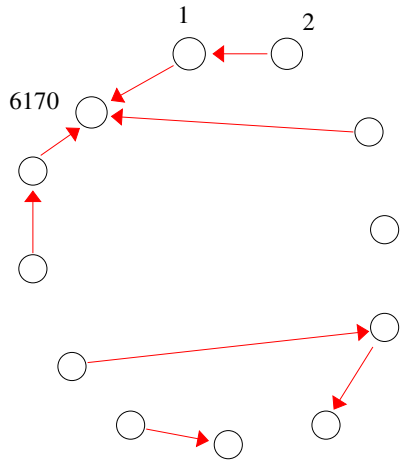
Results for inferring causal variables on a single training-test split

8 genes are "significant" ($\alpha = 0.05$ level) causal variables
(each of the 8 genes "causes" one other gene)

Results for inferring causal variables on a single training-test split

8 genes are "significant" ($\alpha = 0.05$ level) causal variables
(each of the 8 genes "causes" one other gene)

not many findings...



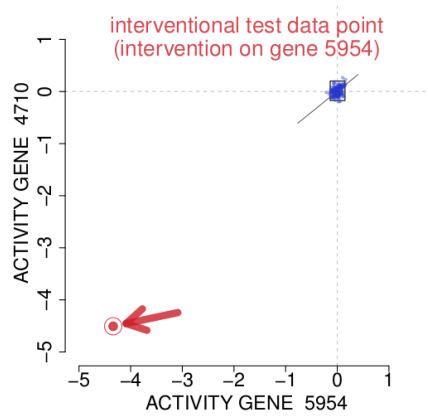
but we use a stringent criterion with Bonferroni corrected
$\alpha/6170 = 0.05/6170$ to control the familywise error rate

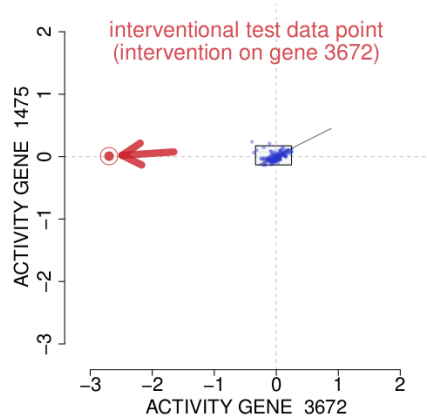
8 genes are “significant” ($\alpha = 0.05$ level) causal variables

validation:
thanks to the intervention experiments (in the test data) we can validate the method(s)

we only consider true Strong Intervention Effects (SIEs)

SIE = the observed response value associated to an intervention is in the 1%- or 99% tail of the observational data





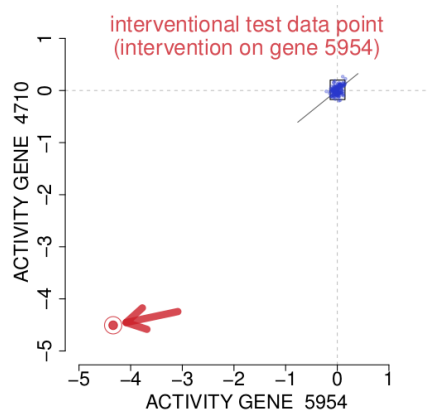
8 genes are “significant” ($\alpha = 0.05$ level) causal variables

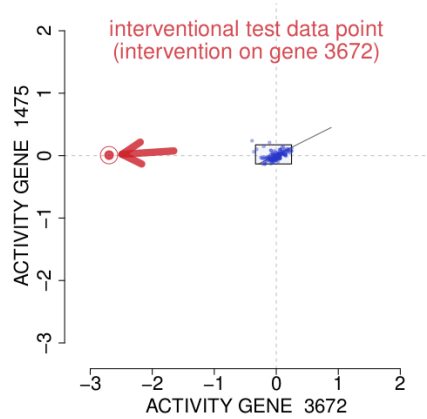
validation:
thanks to the intervention experiments (in the test data) we can validate the method(s)

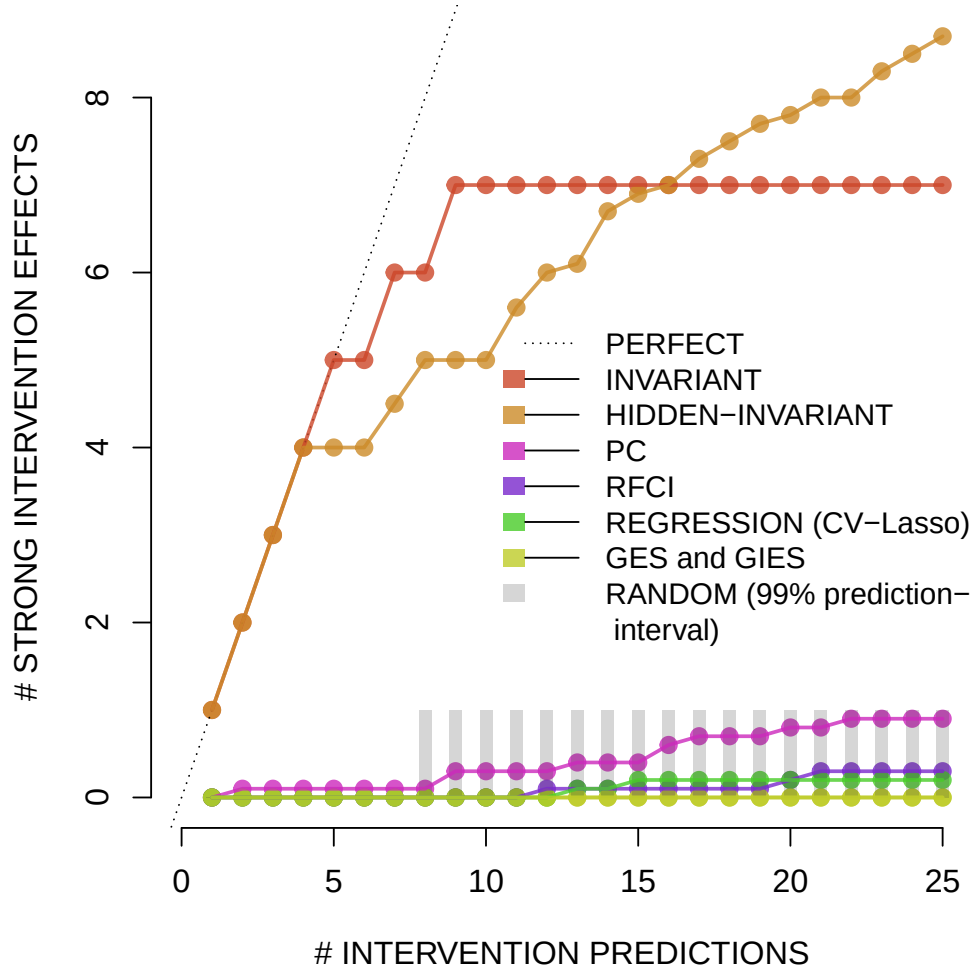
we only consider true Strong Intervention Effects (SIEs)

6 out of the 8 “significant” genes are true SIEs!

SIE = the observed response value associated to an intervention is in the 1%- or 99% tail of the observational data



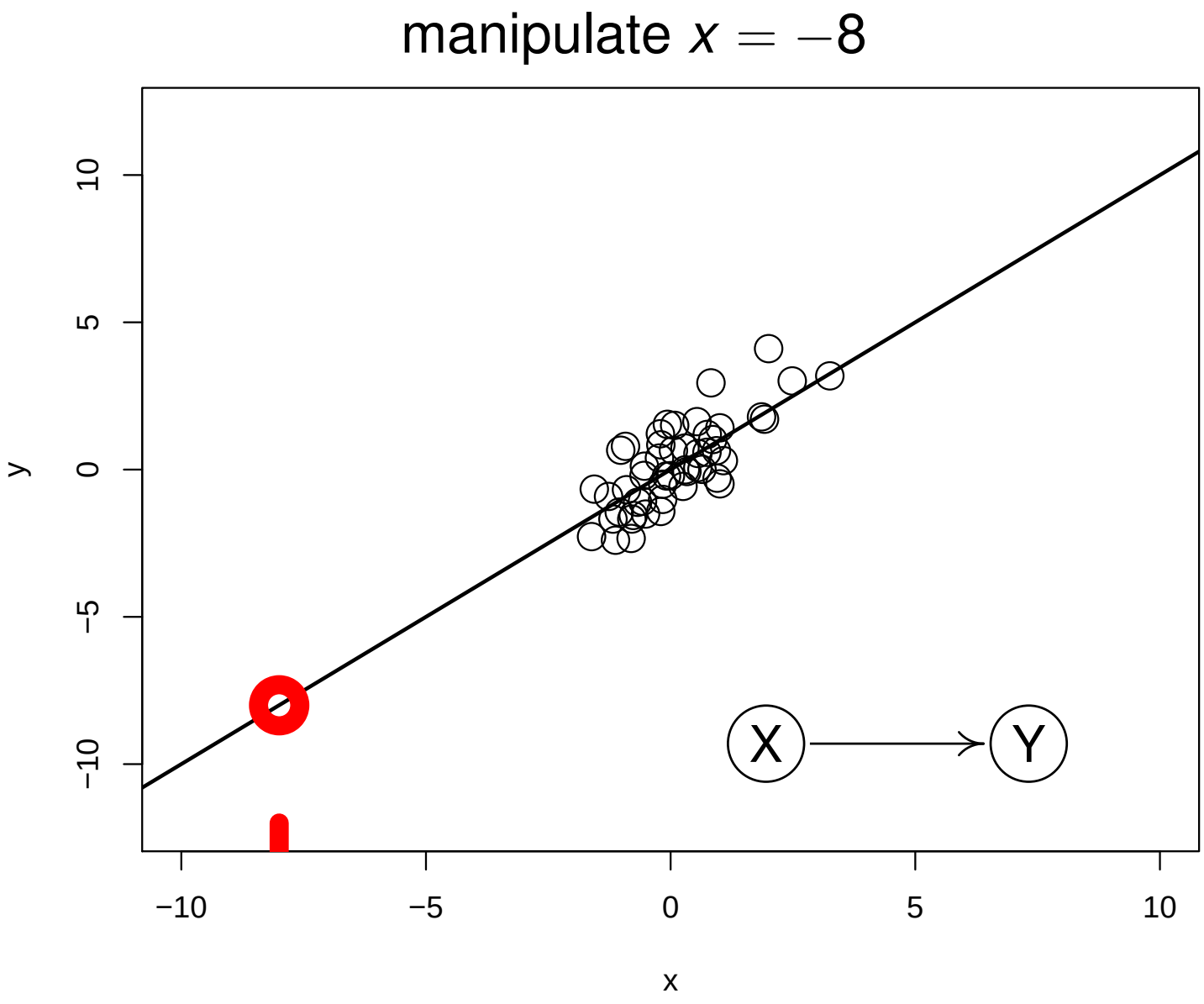




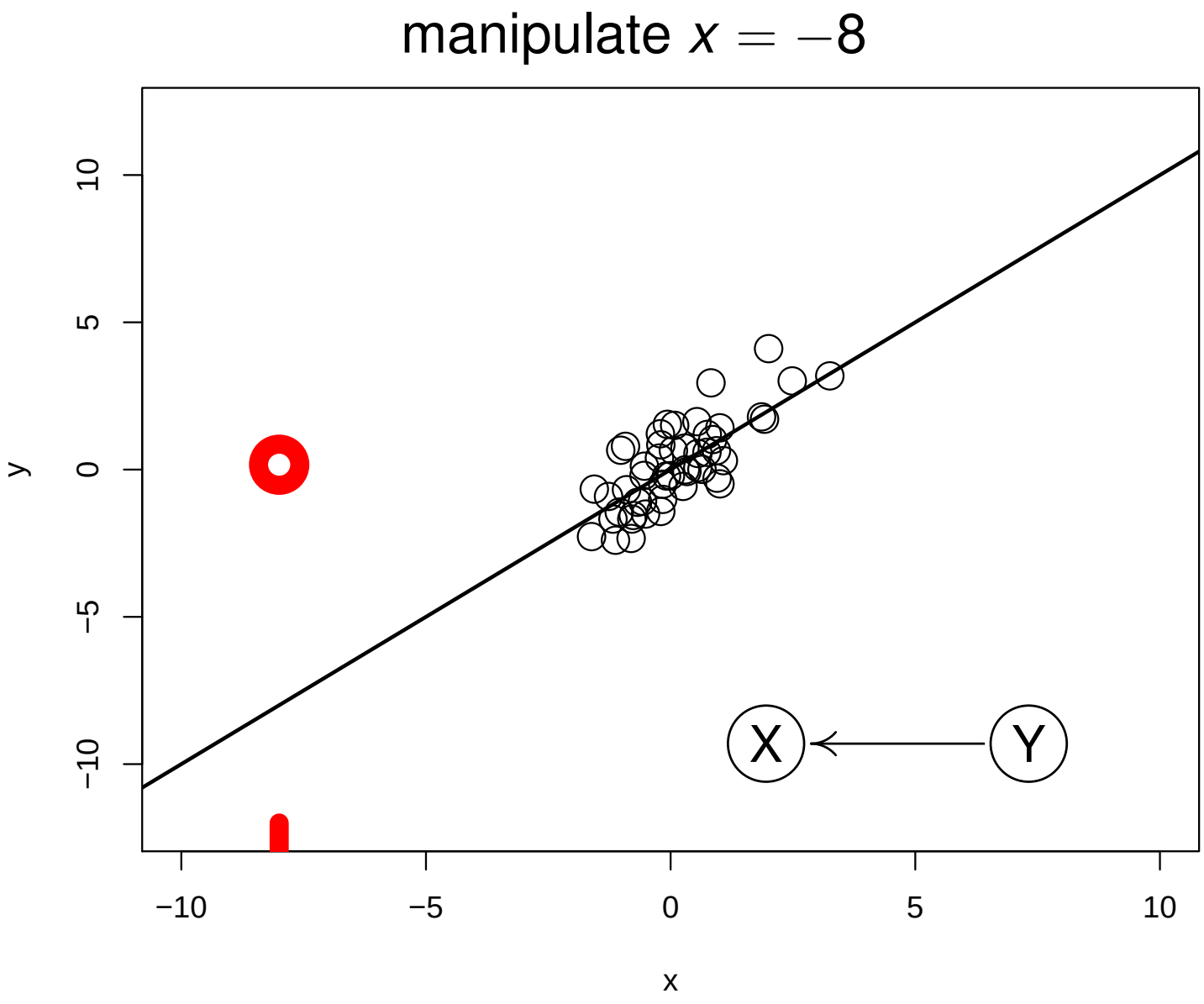
I : invariant prediction method

H: invariant prediction with some hidden variables

Well... it's an ambitious problem



Well... it's an ambitious problem



The Causal Dantzig estimator to deal with hidden variables

(Rothenhäusler, PB & Meinshausen, 2019)

ICP (Invariant Causal Prediction)

- requires an all subset selection search
- does not allow for hidden confounding variables
- is rather general in terms of interventions/perturbations

we develop a methodology and algorithm which

- is computationally efficient (convex optimization)
- allows for hidden confounding
- is more restrictive w.r.t. interventions/perturbations

$\rightsquigarrow$ Causal Dantzig estimator/algorithm

instead of invariance of conditional distributions, require

Assumption: inner product invariance under β^*

$$\mathbb{E}[X_j^e(Y^e - X^e\beta^*)] = \mathbb{E}[X_j^{e'}(Y^{e'} - X^{e'}\beta^*)] \;\; \forall\, e, e' \in \mathcal{E}, \forall\, j$$

Theorem:
Consider

$$X \leftarrow BX + \varepsilon^0$$
$$\rightsquigarrow \quad Y = X_{p+1} = X^T\beta_{\text{causal}} + \varepsilon_Y$$

Inner product invariance holds under the causal coefficient vector β_{causal} if

- the interventions/environments do not act directly on Y
- the interventions are additive noise interventions:

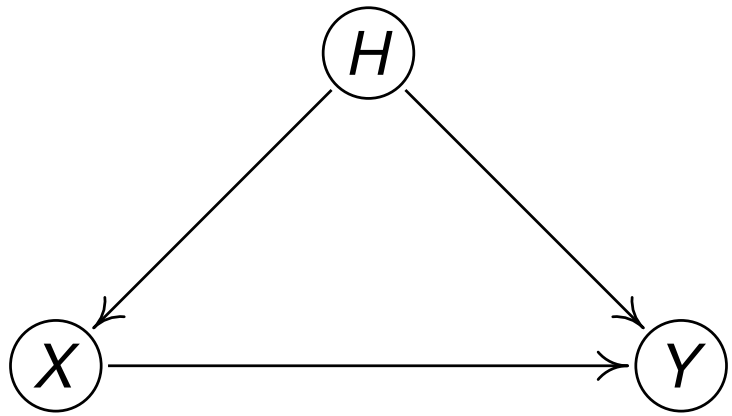
$$\varepsilon^e = \varepsilon^0 + \delta^e$$
$$\mathbb{E}[\varepsilon^0] = 0, \mathrm{Cov}(\varepsilon^0, \delta^e) = 0, \;\; \delta_Y^e \equiv 0$$

and the theorem extends to SEMs with measurement errors

$$\varepsilon^e = \varepsilon^0 + \delta^e$$
$$\mathbb{E}[\varepsilon^0] = 0, \mathrm{Cov}(\varepsilon^0, \delta^e) = 0,\ \delta^e_Y \equiv 0$$

ε^0 and δ^e can have dependent components $\rightsquigarrow$ hidden variables are covered

"reason":



$$Y \leftarrow X\beta + H\delta + \varepsilon_Y = X\beta + \eta_Y$$
$$X \leftarrow H\gamma + \varepsilon_X = \eta_X$$

the η error terms are now dependent!

Causal Dantzig without regularization for low-dimensional settings

consider two environments $e = 1$ and $e' = 2$
differences of Gram matrices:

$$\hat{\mathbf{Z}} = n_1^{-1}(\mathbf{X}^1)^T\mathbf{Y}^1 - n_2^{-1}(\mathbf{X}^2)^T\mathbf{Y}^2,$$
$$\hat{\mathbf{G}} = n_1^{-1}(\mathbf{X}^1)^T\mathbf{X}^1 - n_2^{-1}(\mathbf{X}^2)^T\mathbf{X}^2$$

under inner product invariance with β^*:

$$\mathbb{E}[\hat{\mathbf{Z}} - \hat{\mathbf{G}}\beta^*] = 0$$

$\rightsquigarrow \hat{\beta} = \text{argmin}_\beta \|\hat{\mathbf{Z}} - \hat{\mathbf{G}}\beta\|_\infty$
asymptotic Gaussian distribution with explicit estimable covariance matrix Γ
if β_{causal} is non-identifiable:
the covariance matrix Γ is singular in certain directions
$\rightsquigarrow$ infinite marginal confidence intervals for non-identifiable coefficients $\beta_{\text{causal},k}$

Regularized Causal Dantzig

$$\hat{\beta} = \text{argmin}_{\beta} \|\beta\|_1$$
$$\text{such that } \|\hat{\mathbf{Z}} - \hat{\mathbf{G}}\beta\|_\infty \leq \lambda$$

in analogy to the classical Dantzig selector (Candes & Tao, 2007) which uses

$$\tilde{\mathbf{Z}} = n^{-1}\mathbf{X}^T\mathbf{Y},\ \tilde{\mathbf{G}} = n^{-1}\mathbf{X}^T\mathbf{X}$$

using the machinery of high-dimensional statistics and assuming identifiability (e.g. $\delta^{e'} \neq 0$ except for $\delta^{e'}_Y = 0$) ...

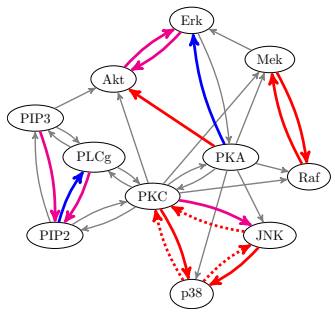
$$\|\hat{\beta} - \beta_{\text{causal}}\|_q \leq O(s^{1/q}\sqrt{\log(p)/\min(n_1, n_2)}) \text{ for } q \geq 1$$

various options to deal with more than two environments:
e.g. all pairs and aggregation

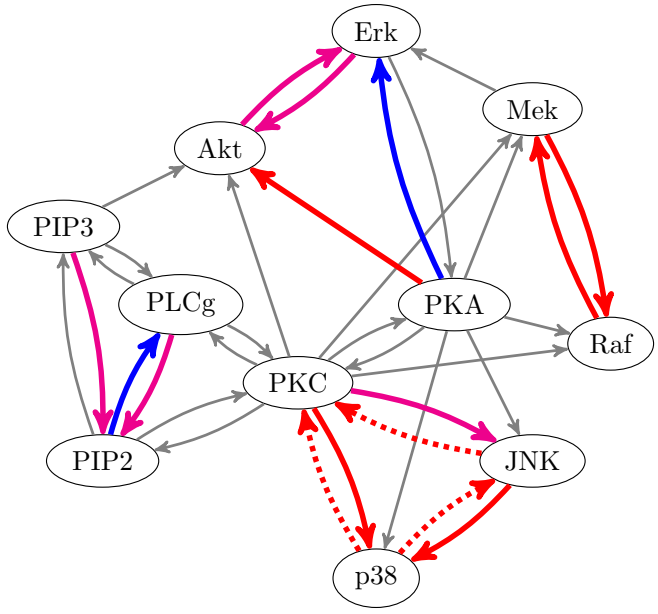
Flow cytometry data (Sachs et al., 2005)

- $p = 11$ abundances of chemical reagents
- 8 different environments (not "well-defined" interventions)
  (one of them observational; 7 different reagents added)
- each environment contains $n_e \approx 700 - 1'000$ samples

goal:
recover network of causal relations (linear SEM)



approach: "pairwise" invariant causal prediction
(one variable the response Y; the other 10 the covariates X;
do this 11 times with every variable once the response)



blue edges: only invariant causal prediction approach (ICP)
red: only ICP allowing hidden variables and feedback
purple: both ICP with and without hidden variables
solid: all relations that have been reported in literature
broken: new findings not reported in the literature

$\rightsquigarrow$ reasonable consensus with existing results
but no real ground-truth available
serves as an illustration that we can work with "vaguely defined interventions"

Causal Regularization

the causal parameter optimizes a worst case risk:

$$\text{argmin}_\beta \max_{e \in \{\mathcal{F}} \mathbb{E}[(Y^e - (X^e)^T \beta)^2] \ni \beta_{\text{causal}}$$

if $\mathcal{F} = \{$arbitrarily strong perturbations not acting directly on $Y\}$

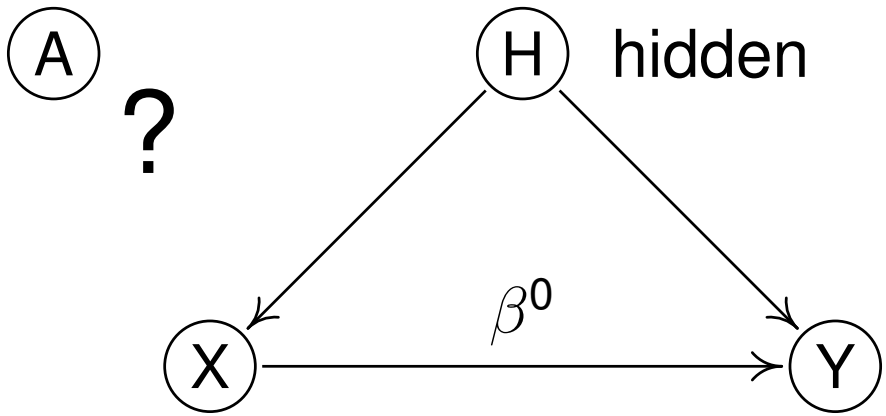
agenda for today: consider other classes $\mathcal{F}$

... and give up on causality

Anchor regression: as a way to formalize the extrapolation from $\mathcal{E}$ to $\mathcal{F}$

(Rothenhäusler, Meinshausen, PB & Peters, 2018)

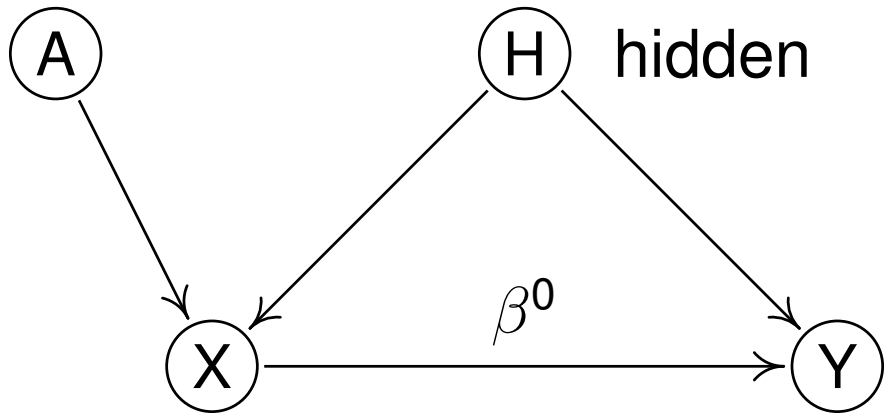
the environments from before, denoted as e:
they are now outcomes of a variable $\underbrace{A}_{\text{anchor}}$



Anchor regression and causal regularization

(Rothenhäusler, Meinshausen, PB & Peters, 2018)

the environments from before, denoted as e:
they are now outcomes of a variable $\underbrace{A}_{\text{anchor}}$



$$Y \leftarrow X\beta^0 + \varepsilon_Y + H\delta,$$
$$X \leftarrow A\alpha^0 + \varepsilon_X + H\gamma,$$

Instrumental variables regression model

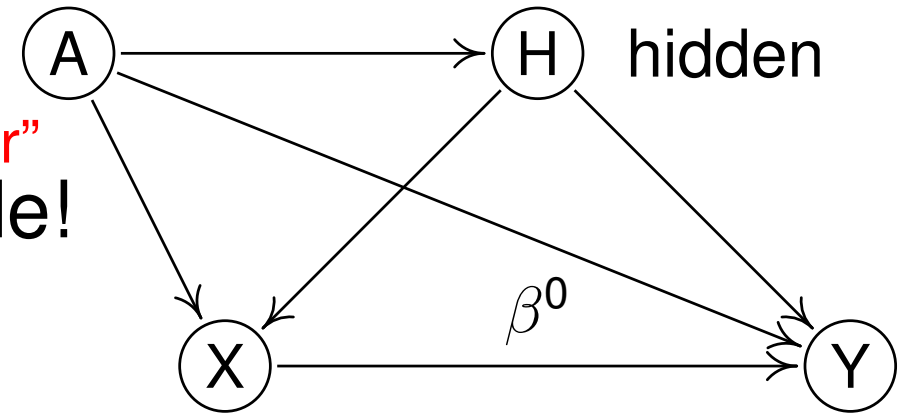
(cf. Angrist, Imbens, Lemieux, Newey, Rosenbaum, Rubin,...)

Anchor regression and causal regularization

(Rothenhäusler, Meinshausen, PB & Peters, 2018)

the environments from before, denoted as e:
they are now outcomes of a variable $\underbrace{A}_{\text{anchor}}$

A is an "anchor"
source node!



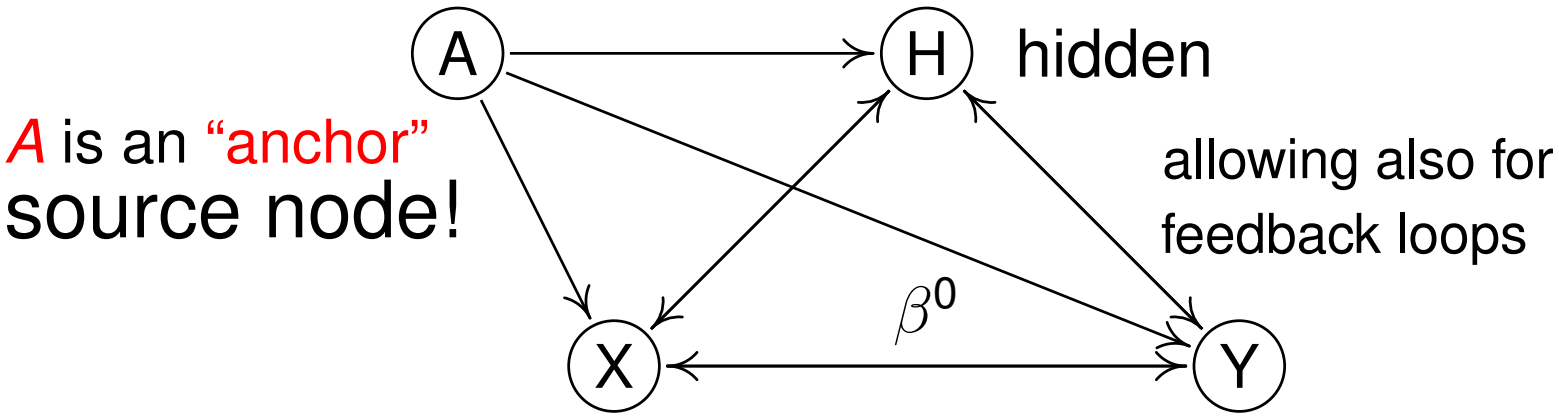
$\rightsquigarrow$ Anchor regression

$$\begin{pmatrix} X \\ Y \\ H \end{pmatrix} \leftarrow B \begin{pmatrix} X \\ Y \\ H \end{pmatrix} + \varepsilon + MA$$

Anchor regression and causal regularization

(Rothenhäusler, Meinshausen, PB & Peters, 2018)

the environments from before, denoted as e:
they are now outcomes of a variable $\underbrace{A}_{\text{anchor}}$



$\rightsquigarrow$ Anchor regression

$$\begin{pmatrix} X \\ Y \\ H \end{pmatrix} \leftarrow B \begin{pmatrix} X \\ Y \\ H \end{pmatrix} + \varepsilon + MA$$

allow that A acts on Y and H

$\rightsquigarrow$ there is a fundamental identifiability problem

cannot identify β^0

this is the price for more realistic assumptions than IV model

... but “Causal Regularization” offers something

find a parameter vector β such that the residuals

$(Y - X\beta)$ stabilize, have the same distribution

across perturbations of A = environments/sub-populations

we want to encourage orthogonality of residuals with A something like

$$\tilde{\beta} = \text{argmin}_\beta \|Y - X\beta\|_2^2/n + \xi\|A^T(Y - X\beta)/n\|_2^2$$

$$\tilde{\beta} = \text{argmin}_\beta \|Y - X\beta\|_2^2/n + \xi\|A^T(Y - X\beta)/n\|_2^2$$

causal regularization:

$$\hat{\beta} = \text{argmin}_\beta \|(I - \Pi_A)(Y - X\beta)\|_2^2/n + \gamma\|\Pi_A(Y - X\beta)\|_2^2/n$$

$\Pi_A = A(A^TA)^{-1}A^T$ (projection onto column space of A)

- for $\gamma = 1$: least squares
- for $\gamma = 0$: adjusting for heterogeneity due to A
- for $0 \leq \gamma < \infty$: general causal regularization

$$\tilde{\beta} = \text{argmin}_\beta \|Y - X\beta\|_2^2/n + \xi \|A^T(Y - X\beta)/n\|_2^2$$

causal regularization:

$$\hat{\beta} = \text{argmin}_\beta \|(I - \Pi_A)(Y - X\beta)\|_2^2/n + \gamma \|\Pi_A(Y - X\beta)\|_2^2/n + \lambda\|\beta\|_1$$

$\Pi_A = A(A^TA)^{-1}A^T$ (projection onto column space of A)

- for $\gamma = 1$: least squares + ℓ_1-penalty
- for $\gamma = 0$: adjusting for heterogeneity due to A + ℓ_1-penalty
- for $0 \leq \gamma < \infty$: general causal regularization + ℓ_1-penalty

convex optimization problem

It's simply linear transformation

consider

$$W_\gamma = I - (1 - \sqrt{\gamma})\Pi_A,$$
$$\tilde{X} = W_\gamma X, \ \tilde{Y} = W_\gamma Y$$

then:
(ℓ_1-regularized) anchor regression is (Lasso-penalized) least squares of $\tilde{Y}$ versus $\tilde{X}$
$\rightsquigarrow$ super-easy (but have to choose a tuning parameter γ)

... there is a fundamental identifiability problem...

but causal regularization solves for

$$\text{argmin}_\beta \max_{e \in \mathcal{F}} \mathbb{E}|Y^e - X^e\beta|^2$$

for a certain class of shift perturbations $\mathcal{F}$

recap: causal parameter solves for
$\text{argmin}_\beta \max_{e \in \mathcal{F}} \mathbb{E}|Y^e - X^e\beta|^2$ for $\mathcal{F} =$ "essentially all" perturbations

Model for $\mathcal{F}$: shift perturbations

model for observed heterogeneous data ("corresponding to $\mathcal{E}$")

$$\begin{pmatrix} X \\ Y \\ H \end{pmatrix} = B \begin{pmatrix} X \\ Y \\ H \end{pmatrix} + \varepsilon + MA$$

model for unobserved perturbations $\mathcal{F}$ (in test data)
shift vectors v acting on (components of) X, Y, H

$$\begin{pmatrix} X^v \\ Y^v \\ H^v \end{pmatrix} = B \begin{pmatrix} X^v \\ Y^v \\ H^v \end{pmatrix} + \varepsilon + v$$

$v \in C_\gamma \subset \text{ span}(M)$, γ measuring the size of v

i.e. $v \in C_\gamma = \{v;\ v = Mu \text{ for some } u \text{ with } \mathbb{E}[uu^T] \preceq \gamma \mathbb{E}[AA^T]\}$

A fundamental duality theorem

(Rothenhäusler, Meinshausen, PB & Peters, 2018)

P_A the population projection onto A: $P_A \bullet = \mathbb{E}[\bullet|A]$

For any β

$$\max_{v \in C_\gamma} \mathbb{E}[|Y^v - X^v\beta|^2] = \mathbb{E}\left[\left|(\mathrm{Id} - P_A)(Y - X\beta)\right|^2\right] + \gamma\mathbb{E}\left[\left|P_A(Y - X\beta)\right|^2\right]$$

$$\approx \underbrace{\|(I - \Pi_A)(Y - X\beta)\|_2^2/n + \gamma\|\Pi_A(Y - X\beta)\|_2^2/n}_{\text{objective function on data}}$$

worst case shift interventions $\longleftrightarrow$ regularization!

in the population case

for any β

$$\overbrace{\max_{v \in C_\gamma} \mathbb{E}\left[\left|Y^v - X^v\beta\right|^2\right]}^{\text{worst case test error}}$$

$$= \underbrace{\mathbb{E}\left[\left|(\mathrm{Id} - P_A)(Y - X\beta)\right|^2\right] + \gamma\mathbb{E}\left[\left|P_A(Y - X\beta)\right|^2\right]}_{\text{criterion on training population sample}}$$

$$\begin{aligned} & \overbrace{\text{argmin}_\beta \max_{v \in C_\gamma} \mathbb{E}\left[\left|Y^v - X^v\beta\right|^2\right]}^{\text{worst case test error}} \\ = \ & \text{argmin}_\beta \underbrace{\mathbb{E}\left[\left|(\text{Id} - P_A)(Y - X\beta)\right|^2\right] + \gamma\mathbb{E}\left[\left|P_A(Y - X\beta)\right|^2\right]}_{\text{criterion on training population sample}} \end{aligned}$$

and “therefore” also finite sample guarantee:

$$\hat{\beta} = \text{argmin}_\beta \|(I - \Pi_A)(Y - Xu)\|_2^2/n + \gamma\|\Pi_A(Y - X\beta)\|_2^2 \ \ (+\lambda\|\beta\|_1)$$

leads to predictive stability (i.e. optimizing a worst case risk)

fundamental duality in anchor regression model:

$$\max_{v \in C_\gamma} \mathbb{E}[|Y^v - X^v\beta|^2] = \mathbb{E}\left[\left|(\mathrm{Id} - P_A)(Y - X\beta)\right|^2\right] + \gamma\mathbb{E}\left[\left|P_A(Y - X\beta)\right|^2\right]$$

$\rightsquigarrow$

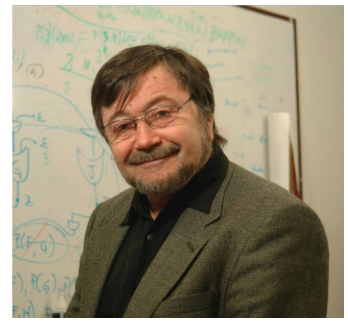
robustness $\longleftrightarrow$ causal regularization

Adversarial Robustness

machine learning, Generative Networks

e.g. Ian Goodfellow

Causality



e.g. Judea Pearl

robustness $\longleftrightarrow$ causal regularization

the languages are rather different:

- metric for robustness
  Wasserstein, f-divergence
- minimax optimality
- inner and outer optimization
- regularization
- ...

- causal graphs
- Markov properties on graphs
- perturbation models
- identifiability of systems
- transferability of systems
- ...

mathematics allows to classify equivalences and differences

$\rightsquigarrow$ can be exploited for better methods and algorithms
taking “the good” from both worlds!

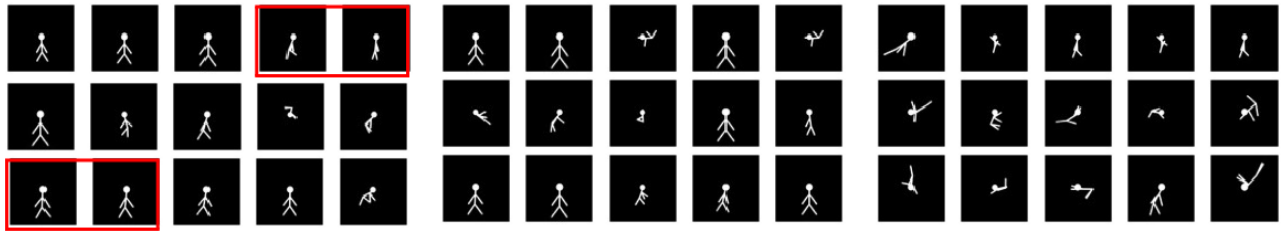
indeed: causal regularization is nowadays used (still a "side-branch") in robust deep learning

Bouttou et al. (2013), ... , Heinze-Deml & Meinshausen (2017), ...

and indeed, we can improve prediction

Stickmen classification (Heinze-Deml & Meinshausen (2017))

Classification into $\{$child, adult$\}$ based on stickmen images



5-layer CNN, training data ($n = 20'000$)

	5-layer CNN	5-layer CNN with some causal regularization
training set	4%	4%
test set 1	3%	4%
test set 2 (domain shift)	41 %	9 %

in training and test set 1: children show stronger movement than adults

in test set 2 data: adults show stronger movement

spurious correlation between age and movement is reversed!

Connection to distributionally robust optimization

(Ben-Tal, El Ghaoui & Nemirovski, 2009; Sinha, Namkoong & Duchi, 2017)

$$\text{argmin}_\beta \max_{P \in \mathcal{P}} \mathbb{E}_P P[(Y - X\beta)^2]$$

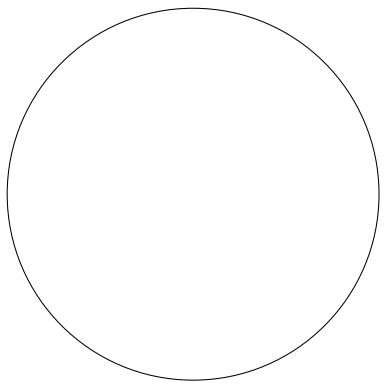
perturbations are within a class of distributions

$$\mathcal{P} = \{P; d(P, \underbrace{P_0}_{\text{emp. distrib.}}) \leq \rho\}$$

the "model" is the metric $d(.,.)$ and is simply postulated

often as Wasserstein distance

Perturbations from distributional robustness

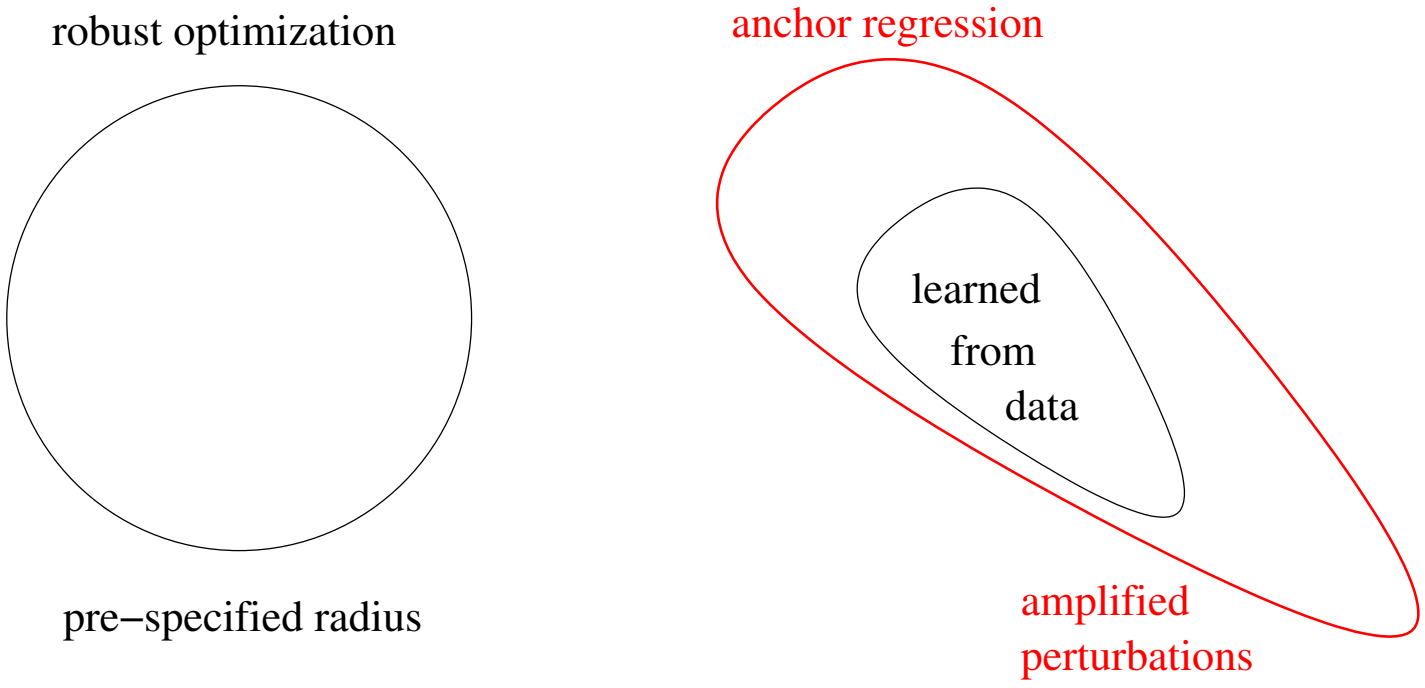


metric d(.,.)
radius rho

our anchor regression approach:

$$b^{\gamma} = \text{argmin}_{\beta} \max_{v \in C_{\gamma}} \mathbb{E}[|Y^{v} - X^{v}\beta|^{2}]$$

perturbations are assumed from a causal-type model
the class of perturbations is learned from data



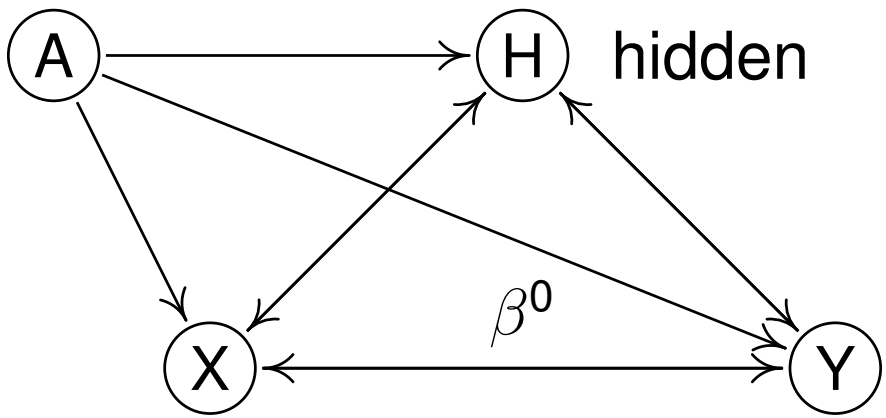
anchor regression: the class of perturbations is an amplification of the observed and learned heterogeneity from $\mathcal{E}$

Science aims for causal understanding

... but this may be a bit ambitious...

in absence of randomized studies, causal inference necessarily requires (often untestable) additional assumptions

in anchor regression model: we cannot find/identify the causal ("systems") parameter β^0



The parameter $b^{\to\infty}$: "diluted causality"

$$b^{\gamma} = \text{argmin}_{\beta} \mathbb{E}\left[\left|(\text{Id} - P_A)(Y - X\beta)\right|^2\right] + \gamma \mathbb{E}\left[\left|P_A(Y - X\beta)\right|^2\right])$$
$$b^{\to\infty} = \lim_{\gamma \to \infty} b^{\gamma}$$

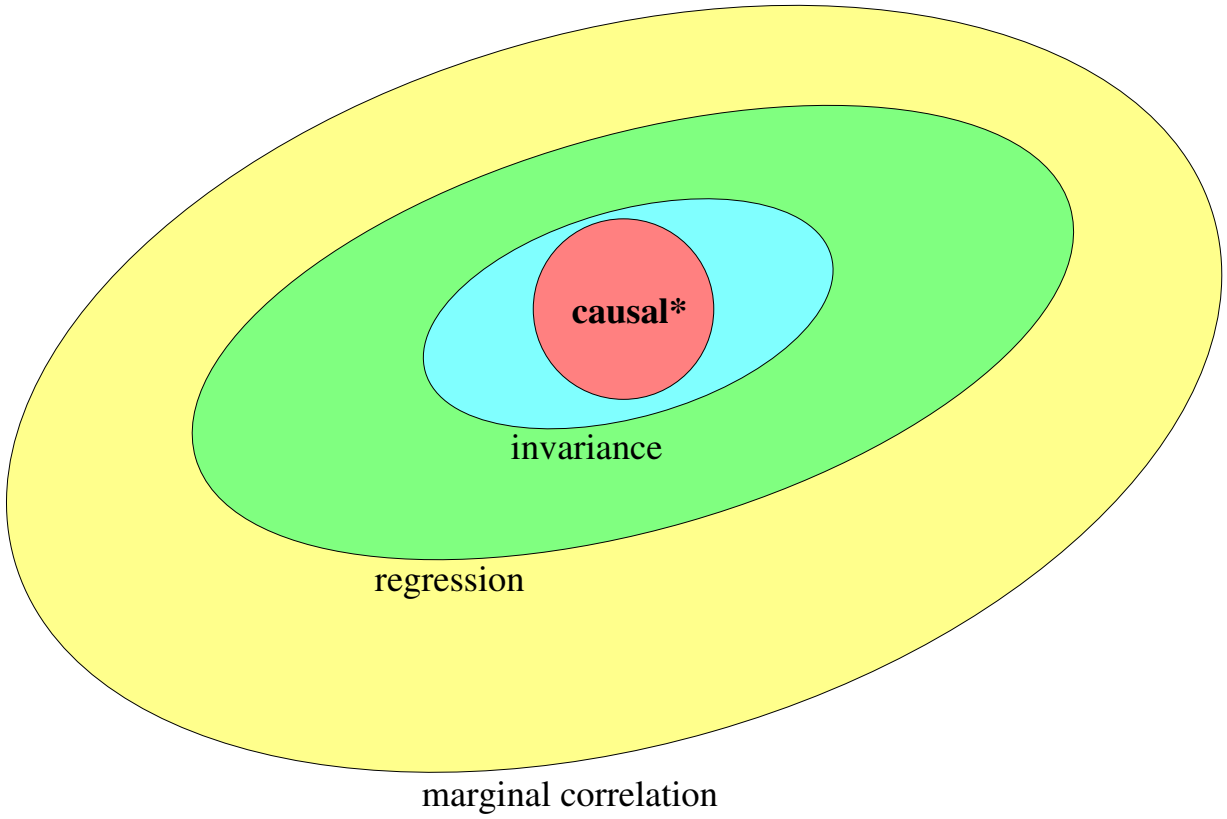
by the fundamental duality: it leads to "invariance"

the parameter which optimizes worst case prediction risk over shift interventions of arbitrary strength

it is generally not the causal parameter
but because of shift invariance: name it "diluted causal"

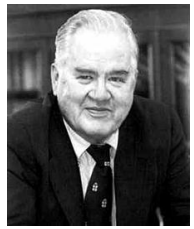
note: causal = invariance w.r.t. very many perturbations

notions of associations



under faithfulness conditions, the figure is valid (causal* are the causal variables as in e.g. large parts of Dawid, Pearl, Robins, Rubin, ...)

Stabilizing

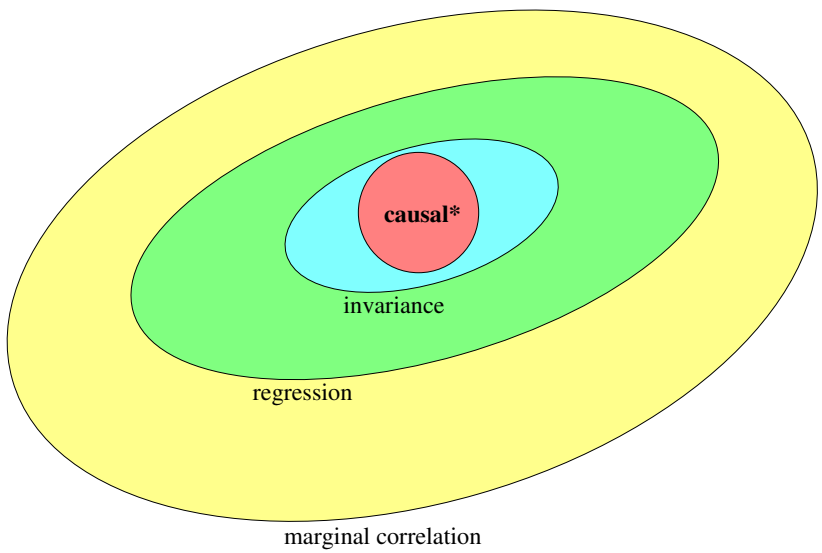


Tukey (1954)

John W. Tukey (1915 – 2000)

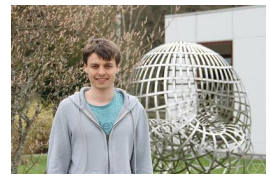
"One of the major arguments for regression instead of correlation is potential stability. We are very sure that the correlation cannot remain the same over a wide range of situations, but it is possible that the regression coefficient might. ...

We are seeking stability of our coefficients so that we can hope to give them theoretical significance."

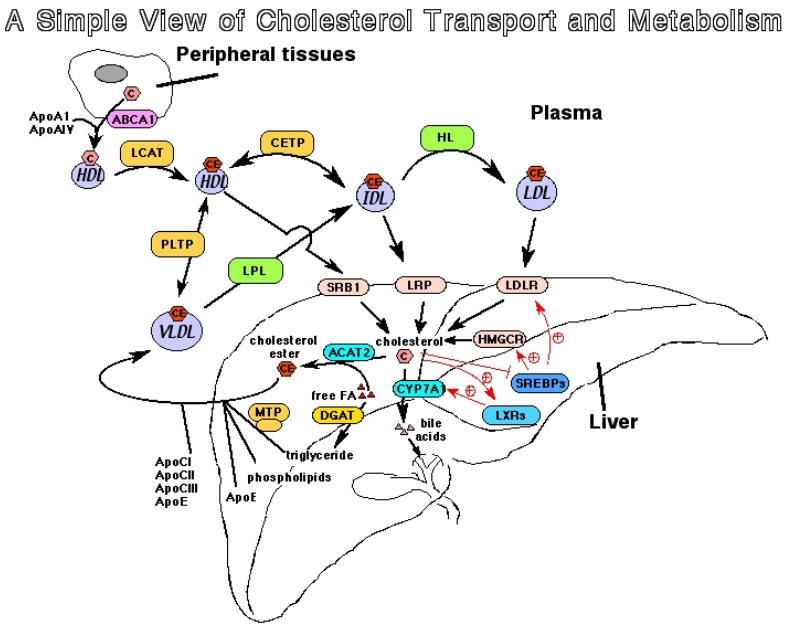


“Diluted causality” and robustness in proteomics

Ruedi Aebersold, ETH Zürich



Niklas Pfister, ETH Zürich



3934 other proteins
which of those are
“diluted causal”
for cholesterol

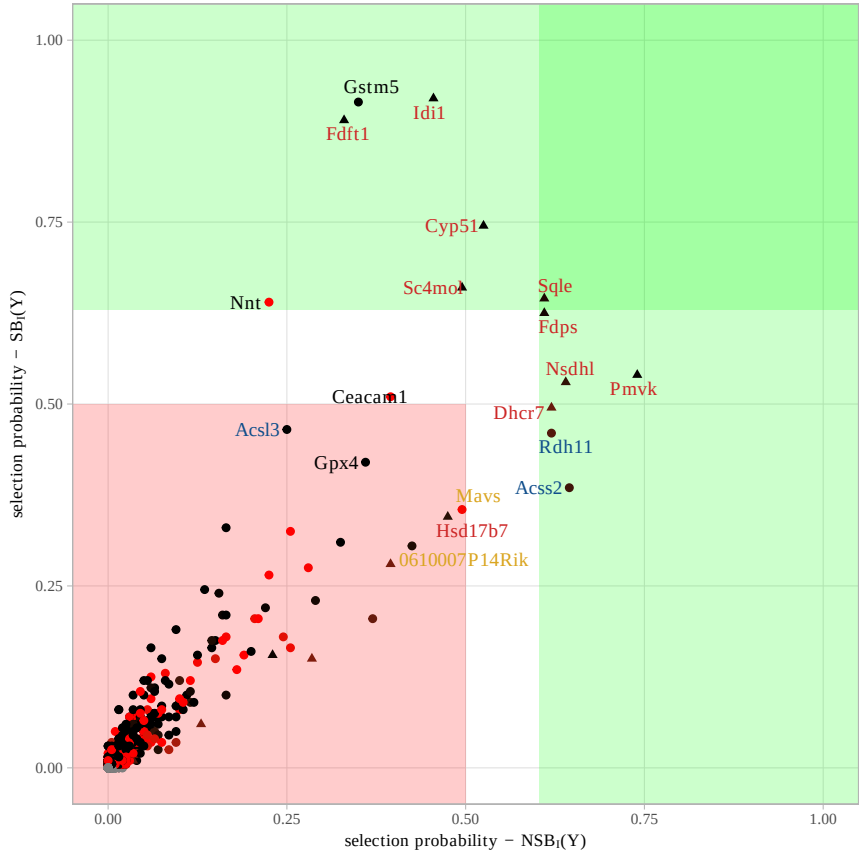
experiments with mice: 2 environments with fat/low fat diet

high-dimensional regression, total sample size $n = 270$

Y = cholesterol pathway activity, X = 3934 protein expressions

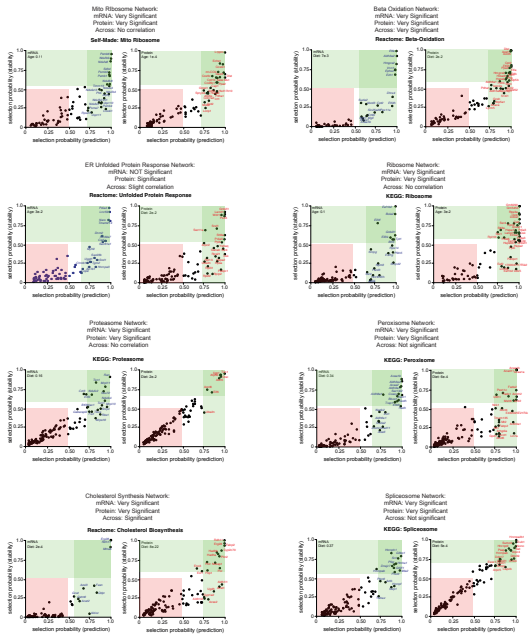
x-axis: importance w.r.t regression but non-invariant

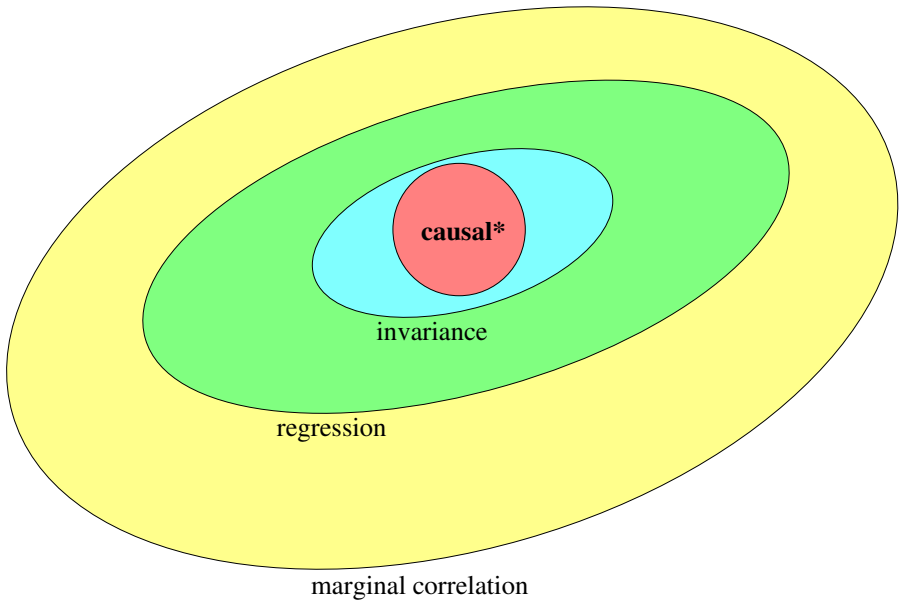
y-axis: importance w.r.t. invariance



beyond cholesterol: with transcriptomics and proteomics

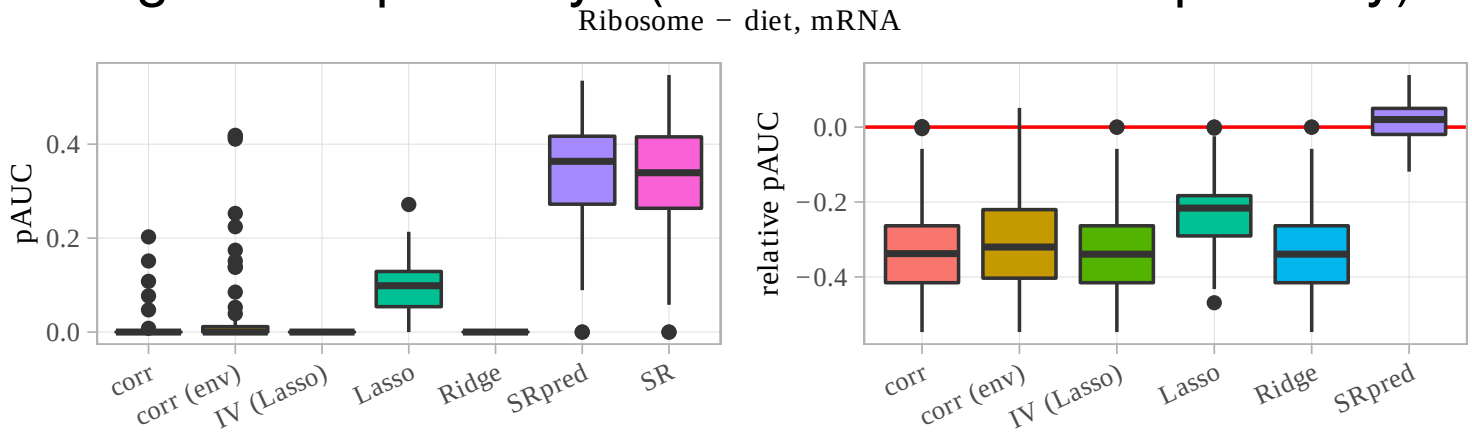
not all of the predictive variables from regression lead to invariance!





"validation" in terms of

- finding known pathways (here for Ribosome pathway)



$\rightsquigarrow$ invariance-type modeling improves over regression!

- reported results in the literature

Distributional Replicability

The replicability crisis

... scholars have found that the results of many scientific studies are difficult or impossible to replicate (Wikipedia)

Distributional Replicability

Replicability on new and different data

- regression parameter b is estimated on one (possibly heterogeneous) dataset with distributions P_e, $e \in \mathcal{E}$
- can we see replication for b on another different dataset with distribution $P_{e'}$, $e' \notin \mathcal{E}$?

this is a question of "zero order" replicability
it is a first step before talking about efficient inference
(in an i.i.d. or stationary setting)

it's not about accurate p-values, selective inference, etc.

The projectability condition

$$I = \{\beta; \mathbb{E}[Y - X\beta|A] \equiv 0\} \neq \emptyset$$

it holds iff

$$\text{rank}(\text{Cov}(A, X)) = \text{rank}\,(\text{Cov}(A, X)|\text{Cov}(A, Y))$$

example:
$\text{rank}(\text{Cov}(A, X))$ is full rank and $\dim(A) \leq \dim(X)$
"under- or just-identified case" in IV literature

checkable! in practice

the “diluted causal” parameter $b^{\to\infty}$ is replicable

assume

- new dataset arises from shift perturbations $v \in \text{span(M)}$ (as before)
- projectability condition holds

consider

$b^{\to\infty}$ which is estimated from the first dataset
$b'^{\to\infty}$ which is estimated from the second (new) dataset

Then: $b^{\to\infty}$ is replicable, i.e.,

$$b^{\to\infty} = b'^{\to\infty}$$

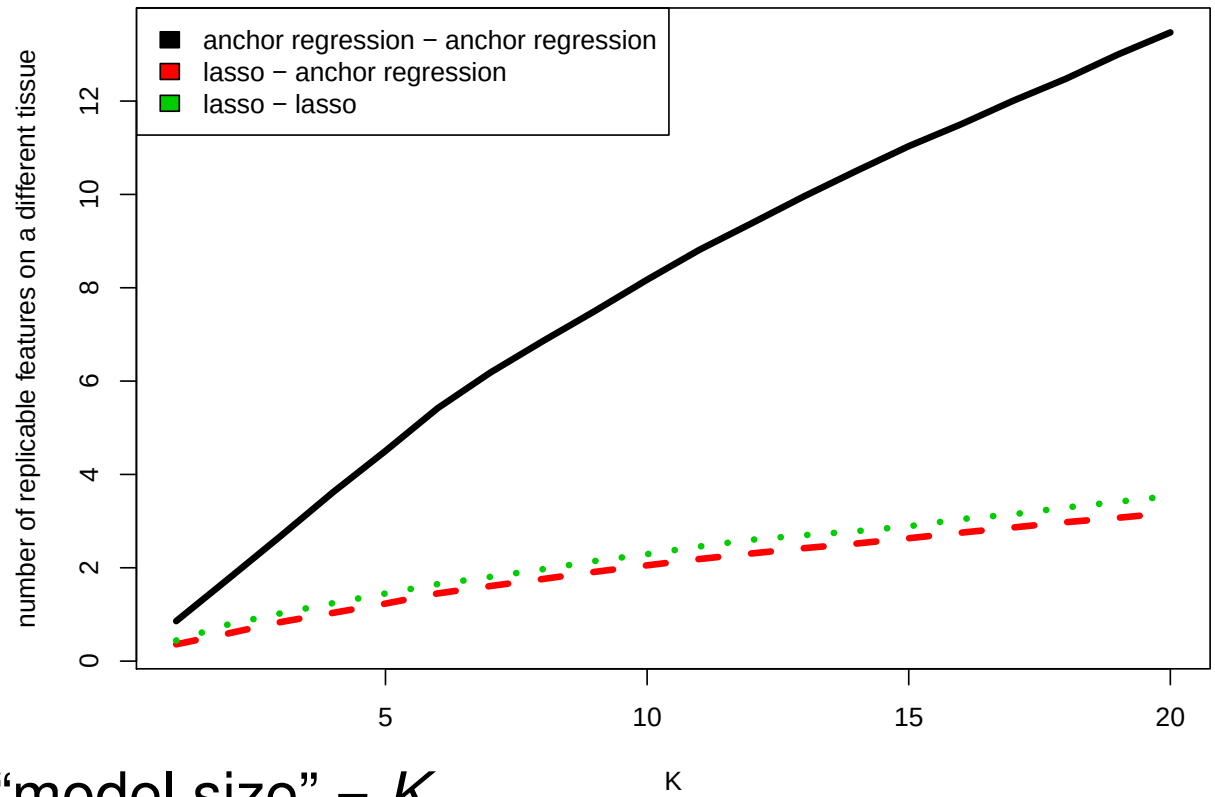
Replicability for $b^{\to\infty}$ in GTEx data across tissues

- 13 tissues
- gene expression measurements for 12'948 genes, sample size between 300 - 700
- Y = expression of a target gene
  X = expressions of all other genes
  A = 65 PEER factors (potential confounders)

estimation and findings on one tissue

$\rightsquigarrow$ are they replicable on other tissues?

Replicability for $b^{\to\infty}$ in GTEx data across tissues



x-axis: "model size" = K

y-axis: how many of the top K ranked associations (found by a method on a tissue t are among the top K on a tissue $t' \neq t$

summed over 12 different tissues $t' \neq t$, averaged over all 13 t and averaged over 1000 random choice of a gene as the response

additional information in anchor regression path!

the anchor regression path:

$$\text{anchor stability: } b^0 = b^{\to\infty} (= b^\gamma \ \forall \gamma \geq 0)$$

checkable!

assume:

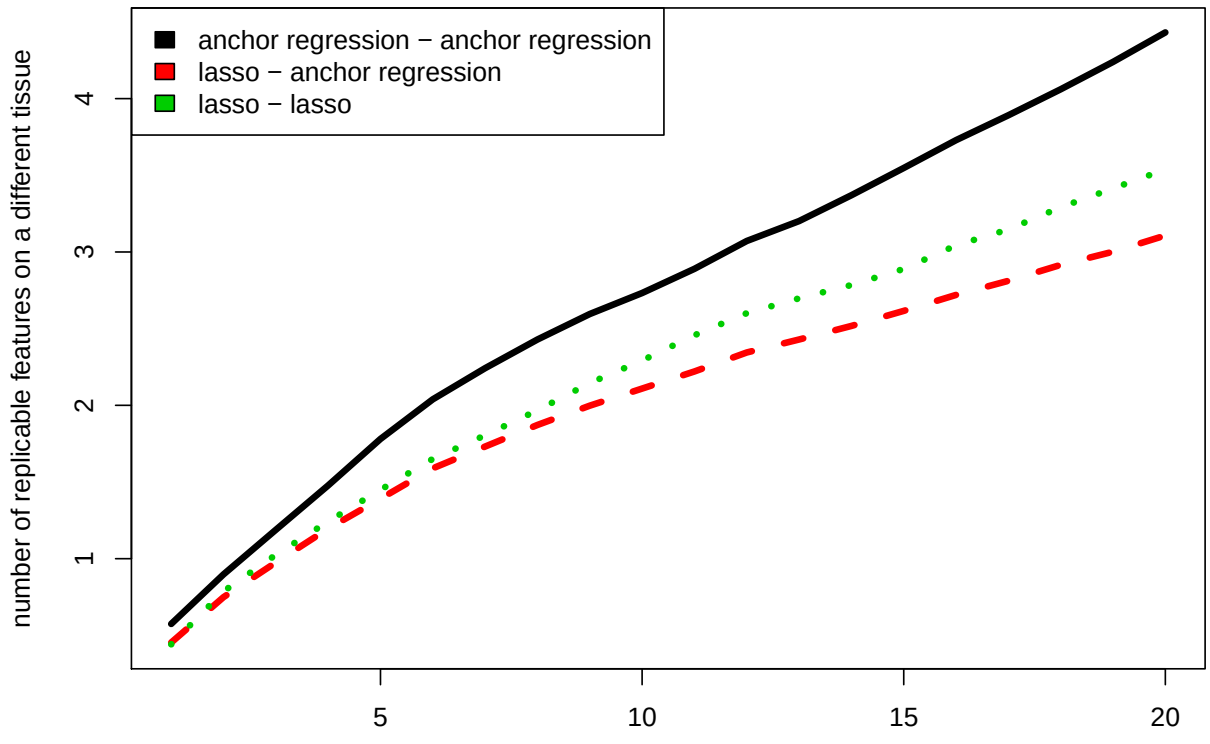
- anchor stability
- projectability condition

$\rightsquigarrow$ the least squares parameter b^1 is replicable!

we can safely use "classical" least squares principle and methods (Lasso/ℓ_1-norm regularization, de-biased Lasso, etc.) for transferability to some class of new data generating distributions $P_{e'}$ $e' \notin \mathcal{E}$

Replicability for least squares par. in GTEx data across tissues

using anchor stability, denoted here as “anchor regression”



x-axis: “model size” = K

y-axis: how many of the top K ranked associations (found by a method on a tissue t are among the top K on a tissue $t' \neq t$

summed over 12 different tissues $t' \neq t$, averaged over all 13 t and averaged over 1000 random choice of a gene as the response

We can make relevant progress by exploiting invariances/stability

- finding more promising proteins and genes: based on high-throughput proteomics
- replicable findings across tissues: based on high-throughput transcriptomics
- prediction of gene knock-downs: based on transcriptomics (Meinshausen, Hauser, Mooij, Peters, Versteeg, and PB, 2016)
- large-scale kinetic systems (not shown): based on metabolomics (Pfister, Bauer and Peters, 2019)

What if there is only observational data with hidden confounding variables?

can lead to spurious associations

number of Nobel prizes vs. chocolate consumption

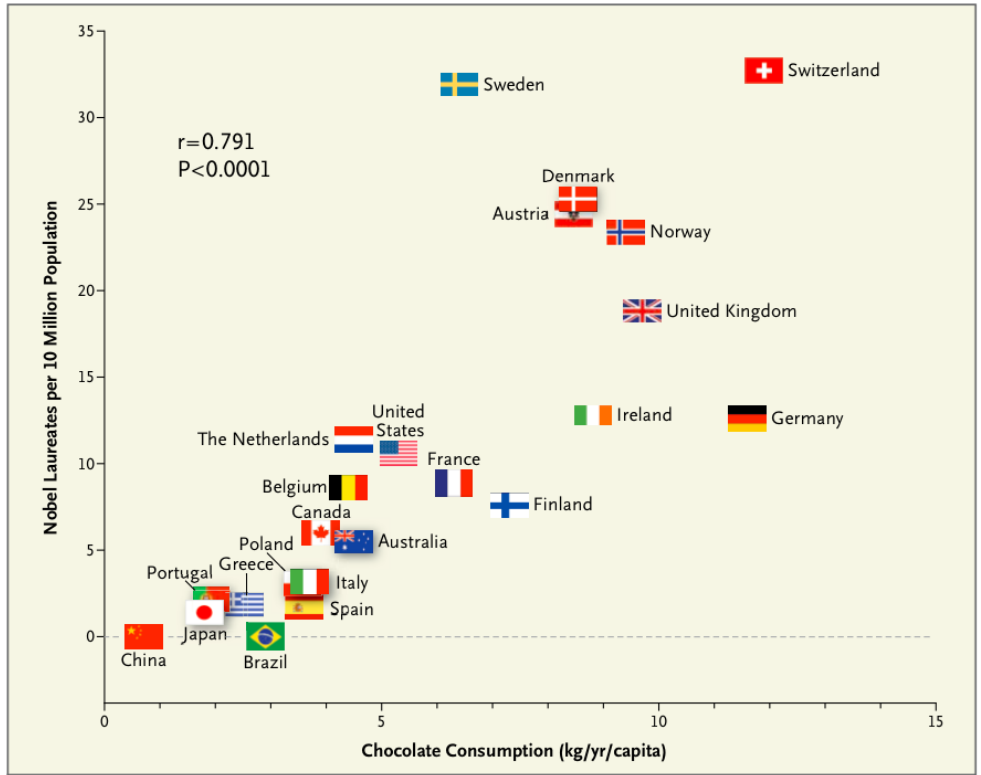


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

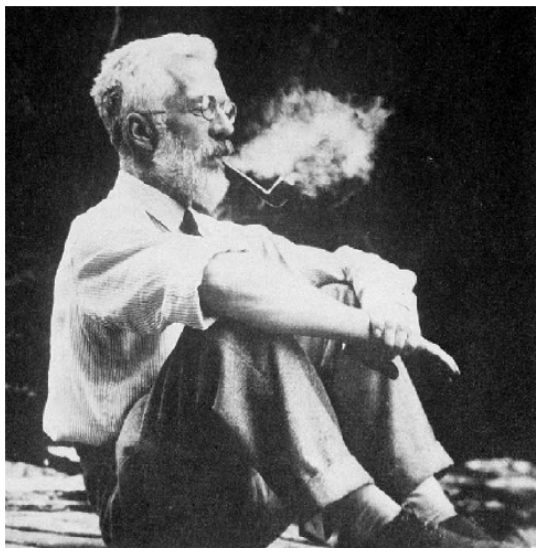
F. H. Messerli: *Chocolate Consumption, Cognitive Function, and Nobel Laureates*, N Engl J Med 2012

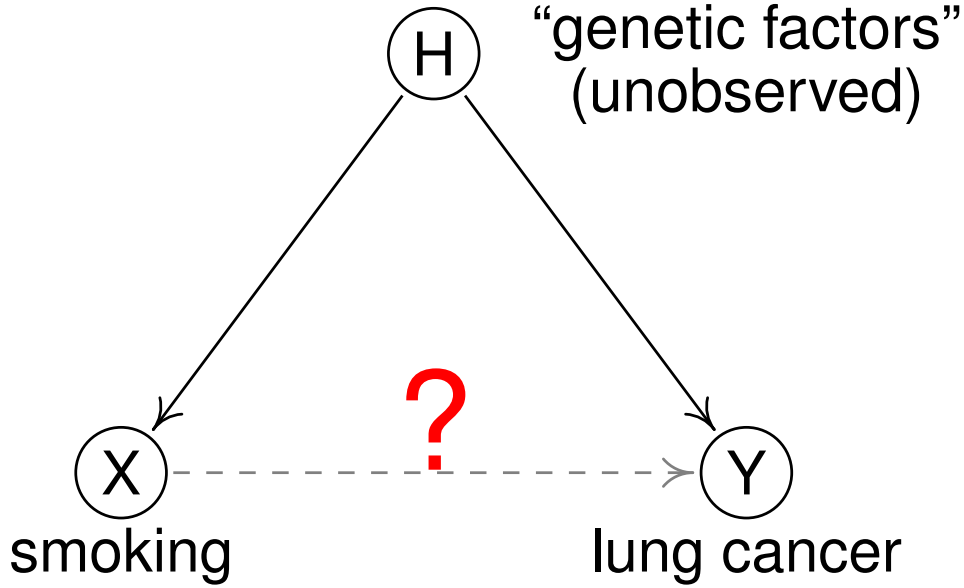
Hidden confounding

can be a major problem

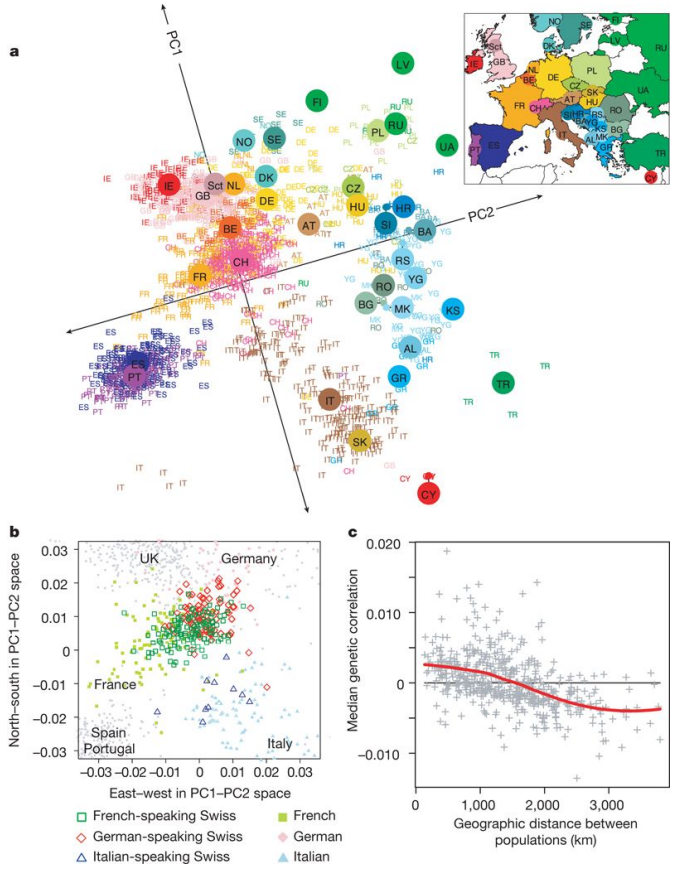
Hidden confounding, causality and perturbation of sparsity

does smoking cause lung cancer?





Genes mirror geography within Europe (Novembre et al., 2008)



confounding effects are found on the first principal components

also for “non-causal” questions:

want to adjust for unobserved confounding

when interpreting regression coefficients, correlations, undirected graphical models, ...

... interpretable AI ...

..., Leek and Storey, 2007; Gagnon-Bartsch and Speed, 2012; Wang, Zhao, Hastie and Owen, 2017; Wang and Blei, 2018;...

in particular: we want to “robustify” the Lasso against hidden confounding variables

also for “non-causal” questions:

want to adjust for unobserved confounding

when interpreting regression coefficients, correlations, undirected graphical models, ...

... interpretable AI ...

..., Leek and Storey, 2007; Gagnon-Bartsch and Speed, 2012; Wang, Zhao, Hastie and Owen, 2017; Wang and Blei, 2018;...

in particular: we want to “robustify” the Lasso against hidden confounding variables

Linear model setting

response Y, covariates X

aim: estimate the regression parameter of Y versus X in presence of hidden confounding

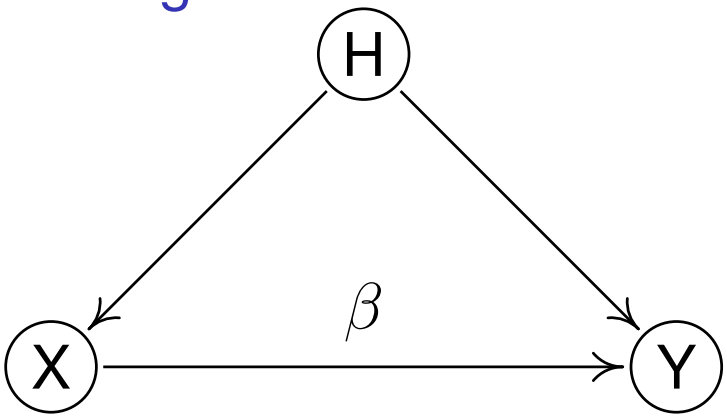
- want to be
  “robust” against unobserved confounding

  we might not completely address the unobserved confounding problem in a particular application

  but we are “essentially always” better than doing nothing against it!
- the procedure should be
  simple with almost zero effort to be used!

  $\rightsquigarrow$ it’s just linearly transforming the data!
- some mathematical guarantees

The setting and a first formula



$$Y = X\beta + H\delta + \eta$$
$$X = H\Gamma + E$$

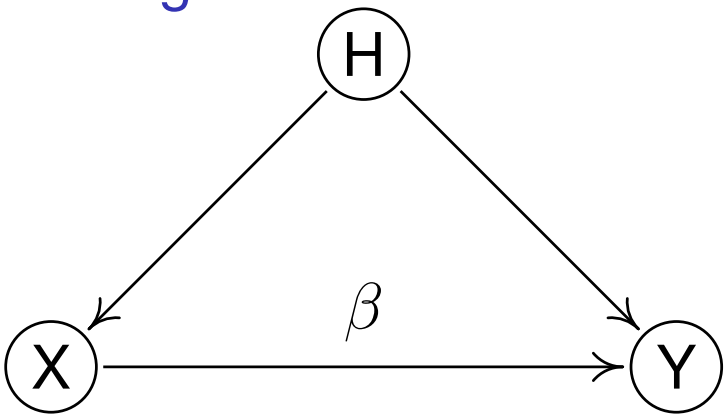
goal: infer β from observations $(X_1, Y_1), \ldots, (X_n, Y_n)$

the population least squares principle leads to the parameter

$$\beta^* = \text{argmin}_u \mathbb{E}[(Y - X^T u)^2],$$
$$\beta^* = \beta + \underbrace{b}_{\text{“bias”/”perturbation”}}$$
$$\|b\|_2 \leq \frac{\|\delta\|_2}{\sqrt{\text{“number of } X\text{-components affected by } H\text{”}}}$$

small “bias”/”perturbation” if confounder has dense effects!

The setting and a first formula



$$Y = X\beta + H\delta + \eta$$
$$X = H\Gamma + E$$

goal: infer β from observations $(X_1, Y_1), \ldots, (X_n, Y_n)$

the population least squares principle leads to the parameter

$$\beta^* = \text{argmin}_u \mathbb{E}[(Y - X^T u)^2],$$
$$\beta^* = \beta + \underbrace{b}_{\text{“bias”/”perturbation”}}$$
$$\|b\|_2 \leq \frac{\|\delta\|_2}{\sqrt{\text{“number of } X\text{-components affected by } H\text{”}}}$$

small “bias”/”perturbation” if confounder has dense effects!

Perturbation of sparsity

the hidden confounding model

$$Y = X\beta + H\delta + \eta$$
$$X = H\Gamma + E$$

can be written as

$$Y = X\beta^* + \varepsilon,$$
$$\beta^* = \underbrace{\beta}_{\text{"sparse"}} + \underbrace{b}_{\text{"dense"}}$$
ε uncorrelated of X, $\mathbb{E}[\varepsilon] = 0$

and $\|b\|_2 \leq \dfrac{\|\delta\|_2}{\sqrt{\text{"number of } X\text{-components affected by } H\text{"}}}$

Perturbation of sparsity

the hidden confounding model

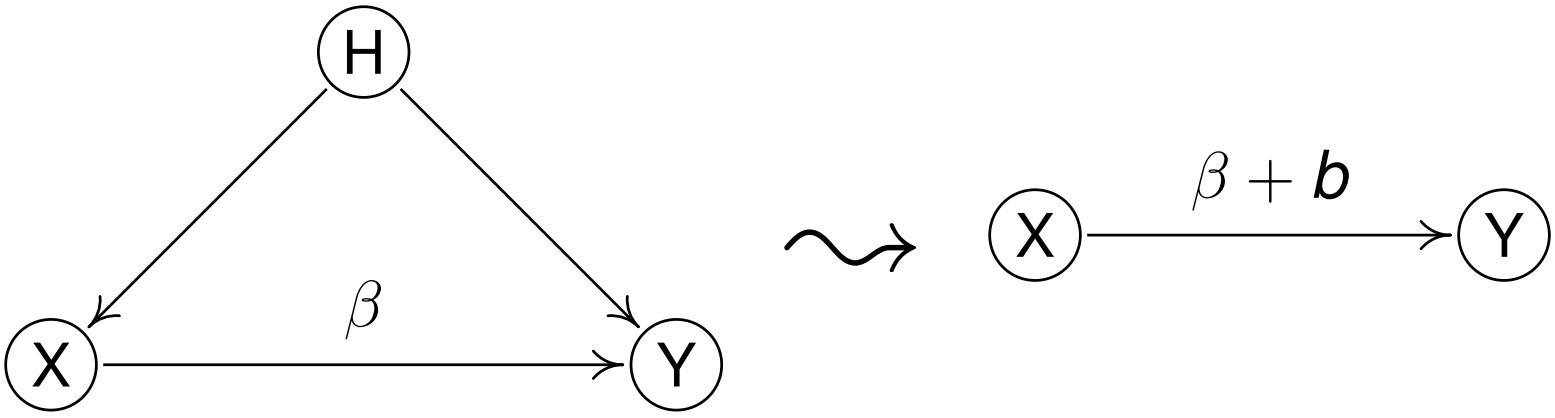
$$Y = X\beta + H\delta + \eta$$
$$X = H\Gamma + E$$

can be written as

$$Y = X\beta^* + \varepsilon,$$
$$\beta^* = \underbrace{\beta}_{\text{"sparse"}} + \underbrace{b}_{\text{"dense"}}$$
ε uncorrelated of X, $\mathbb{E}[\varepsilon] = 0$

and $\|b\|_2 \leq \dfrac{\|\delta\|_2}{\sqrt{\text{"number of } X\text{-components affected by } H\text{"}}}$

hidden confounding is perturbation to sparsity



$$Y = X\beta + H\delta + \eta,$$
$$X = H\Gamma + E$$

$$Y = X(\beta + b) + \varepsilon,$$
$$b = \Sigma^{-1}\Gamma^T\delta \ \text{(”dense”)}$$
$$\Sigma = \Sigma_E + \Gamma^T\Gamma,$$
$$\sigma_\varepsilon^2 = \sigma_\eta^2 + \delta^T(I - \Gamma\Sigma\Gamma^T)\delta$$

and thus $\rightsquigarrow$ consider the more general model

$$Y = X(\beta + b) + \varepsilon,$$
$$\beta \text{ "sparse"}, \; b \text{ "dense"}$$

goal: recover β

Lava method (Chernozhukov, Hansen & Liao, 2017) is considering this model/problem

- with no connection to hidden confounding
- we improve the results and provide a "somewhat simpler" methodology

and thus $\rightsquigarrow$ consider the more general model

$$Y = X(\beta + b) + \varepsilon,$$
$$\beta \text{ "sparse"}, \; b \text{ "dense"}$$

goal: recover β

Lava method (Chernozhukov, Hansen & Liao, 2017) is considering this model/problem

- with no connection to hidden confounding
- we improve the results and provide a "somewhat simpler" methodology

What has been proposed earlier (among many other suggestions)

- adjust for a few first PCA components from X
  motivation: low-rank structure is generated from a few unobserved confounders

  well known among practitioners:
  often pretty reasonable... but we will improve on it
- latent variable models and EM-type or MCMC algorithms (Wang and Blei, 2018)

  need precise knowledge of hidden confounding structure
  cumbersome for fitting to data
- undirected graphical model search with penalization encouraging sparsity plus low-rank (Chandrasekharan et al., 2012)

  two tuning parameters to choose, not so straightforward

..., Leek and Storey, 2007; Gagnon-Bartsch and Speed, 2012; Wang, Zhao, Hastie and Owen, 2017; ... $\rightsquigarrow$ different

motivation: when using Lasso for the non-sparse problem with $\beta^* = \beta + b$

$$\text{a bias term } \|Xb\|_2^2/n \text{ enters}$$

for the bound of $\|X\hat{\beta} - X\beta^*\|_2^2/n + \|\hat{\beta} - \beta^*\|_1$

strategy: linear transformation $F : \mathbb{R}^n \to \mathbb{R}^n$

$$\tilde{Y} = FY,\ \tilde{X} = FX,\ \tilde{\varepsilon} = F\varepsilon,$$
$$\tilde{Y} = \tilde{X}\beta^* + \tilde{\varepsilon}$$

and use Lasso for $\tilde{Y}$ versus $\tilde{X}$ such that

- $\|\tilde{X}b\|_2^2/n$ small
- $\tilde{X}\beta$ "large"
- $\tilde{\varepsilon}$ remains "of order $O(1)$"

Spectral transformations

which transform singular values of X will achieve

- $\|\tilde{X}b\|_2^2/n$ small
- $\tilde{X}\beta$ "large"
- $\tilde{\varepsilon}$ remains "of order $O(1)$

consider SVD of X:

$$
\begin{aligned}
& X = UDV^T, \\
& U_{n\times n}, V_{p\times n},\ U^TU = V^TV = I, \\
& D = \mathrm{diag}(d_1, \ldots, d_n),\ d_1 \geq d_2 \geq \ldots \geq d_n \geq 0
\end{aligned}
$$

map d_i to $\tilde{d}_i$: spectral transformation is defined as

$$
\begin{aligned}
& F = U\mathrm{diag}(\tilde{d}_1/d_1, \ldots, \tilde{d}_n/d_n)U^T \\
\rightsquigarrow \quad & \tilde{X} = U\tilde{D}V^T
\end{aligned}
$$

Examples of spectral transformations

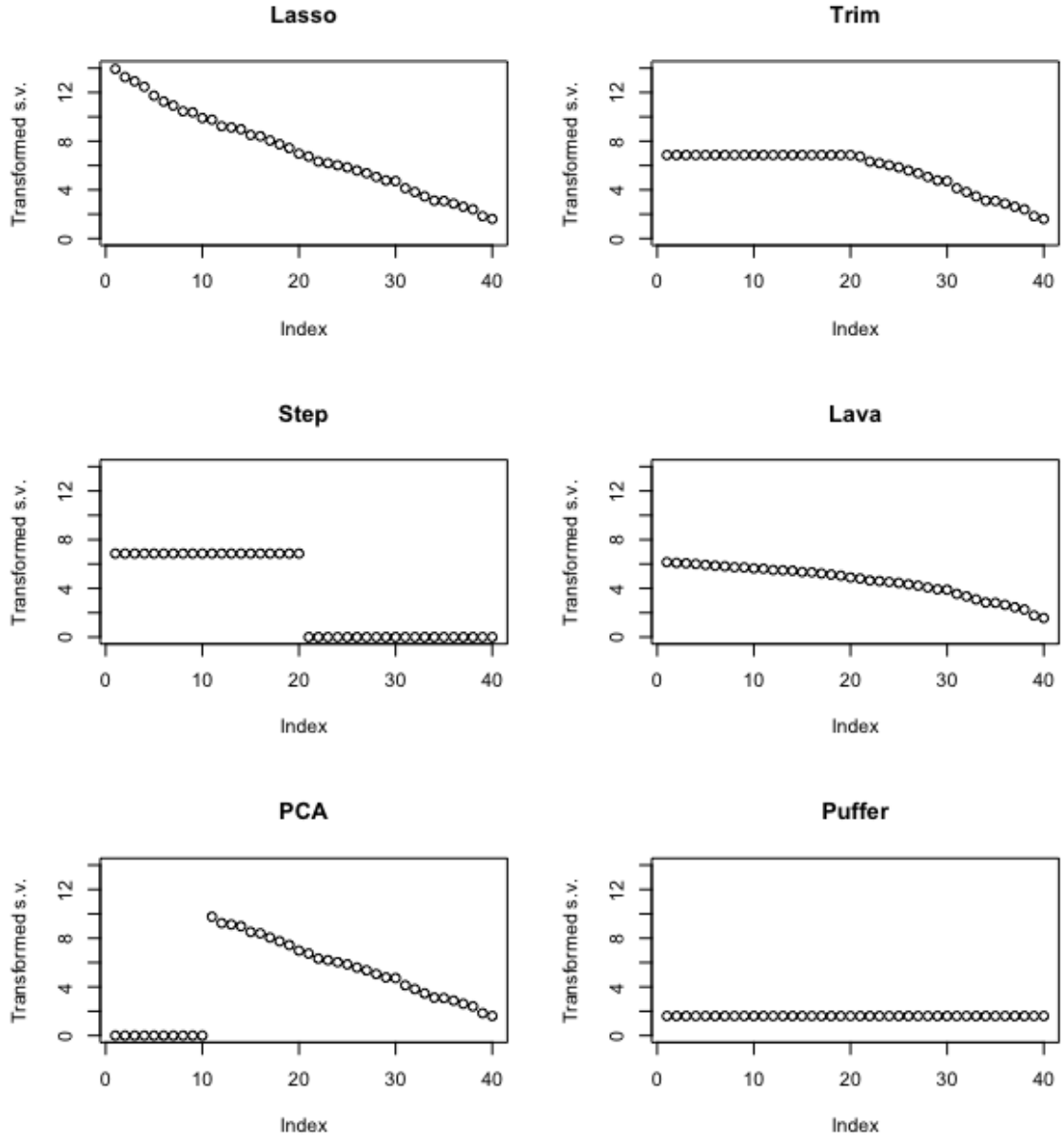
1. adjustment with r largest principal components
   equivalent to $\tilde{d}_1 = \ldots = \tilde{d}_r = 0$
2. Lava (Chernozhukov, Hansen & Liao, 2017)

   $$\text{argmin}_{\beta,b} \| Y - X(\beta + b) \|_2^2 / n + \lambda_1 \|\beta\|_1 + \lambda_2 \|b\|_2^2$$

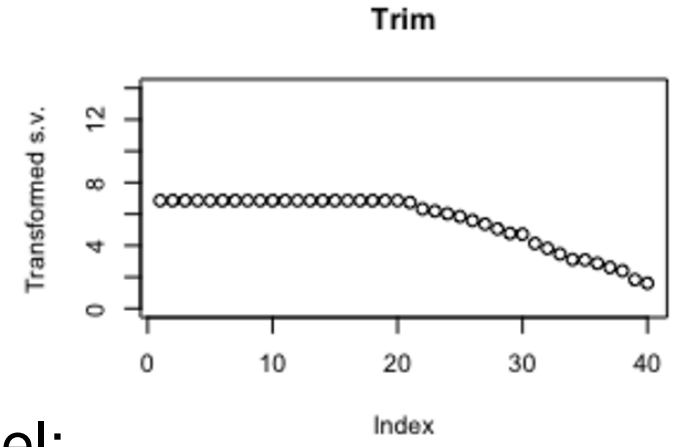
   can be represented as a spectral transform plus Lasso
3. Puffer transform (Jia & Rohe 2015) uses
   $\tilde{d}_i \equiv 1$
   $\rightsquigarrow$ if d_n is small, the errors are inflated...!
4. Trim transform (Ćevid, PB & Meinshausen, 2018)
   $\tilde{d}_i = \min(d_i, \tau)$ with $\tau = d_{\lfloor n/2 \rfloor}$

singular values of $\tilde{X}$

Lasso = no transformation



Heuristics



in hidden confounding model:

- b points towards singular vectors with large singular val.
  $\rightsquigarrow$ it suffices to shrink only large singular values
  to make the "bias" $\|\tilde{X}b\|_2^2/n$ small
- β typically does not point to singular vectors with large singular val.: since β is sparse and V is dense
  (unless there is a tailored dependence between β and the structure of X)
  $\rightsquigarrow$ "signal" $\|\tilde{X}\beta\|_2^2/n$ does not change too much
  when shrinking only large singular values

Some (subtle) theory

consider confounding model

$$Y = X\beta + H\delta + \eta,$$
$$X = H\Gamma + E$$

Theorem (Ćevid, PB & Meinshausen, 2018)
Assume:

- Γ must spread to $O(p)$ components of X
  components of Γ and δ are i.i.d. sub-Gaussian r.v.s (but then thought as fixed)
- condition number of $\Sigma_E = O(1)$
- $\dim(H) = q < s\log(p)$, $s = \mathrm{supp}(\beta)$ (sparsity)

Then, when using Lasso on $\tilde{X}$ and $\tilde{Y}$:

$$\|\hat{\beta} - \beta\|_1 = O_P\left(\frac{\sigma s}{\lambda_{\min}(\Sigma)}\sqrt{\frac{\log(p)}{n}}\right)$$

same optimal rate of Lasso as without confounding variables

limitation: when hidden confounders only spread to/affect m components of X

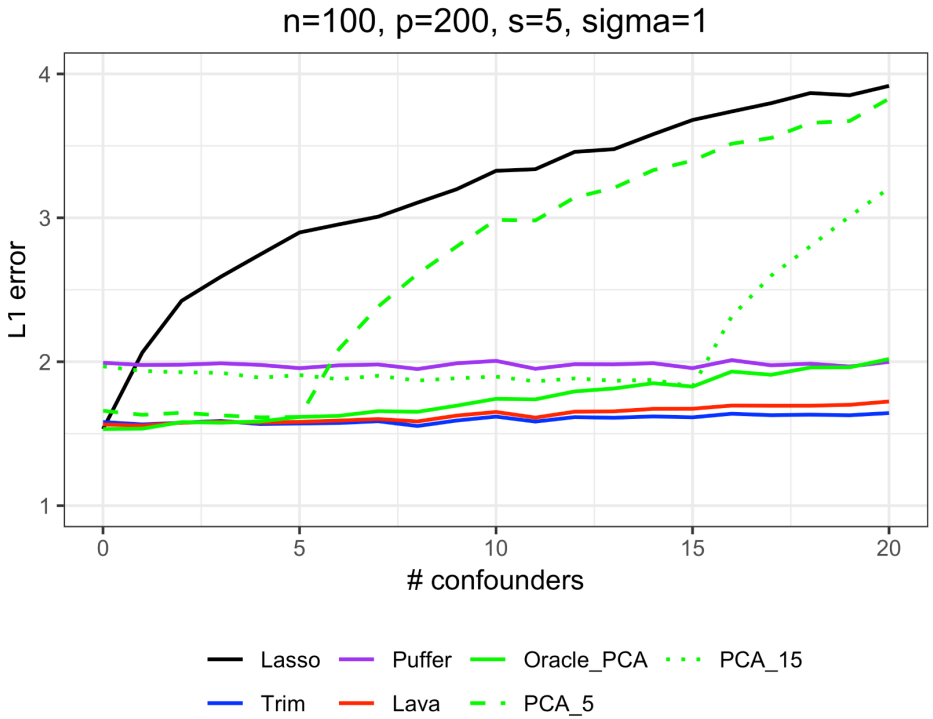
$$\|\hat{\beta} - \beta\|_1 \leq O_P\left(\frac{\sigma s}{\lambda_{\min}(\Sigma)}\sqrt{\frac{\log(p)}{n}} + \frac{\sqrt{s}\|\delta\|_2}{\sqrt{m}}\right)$$

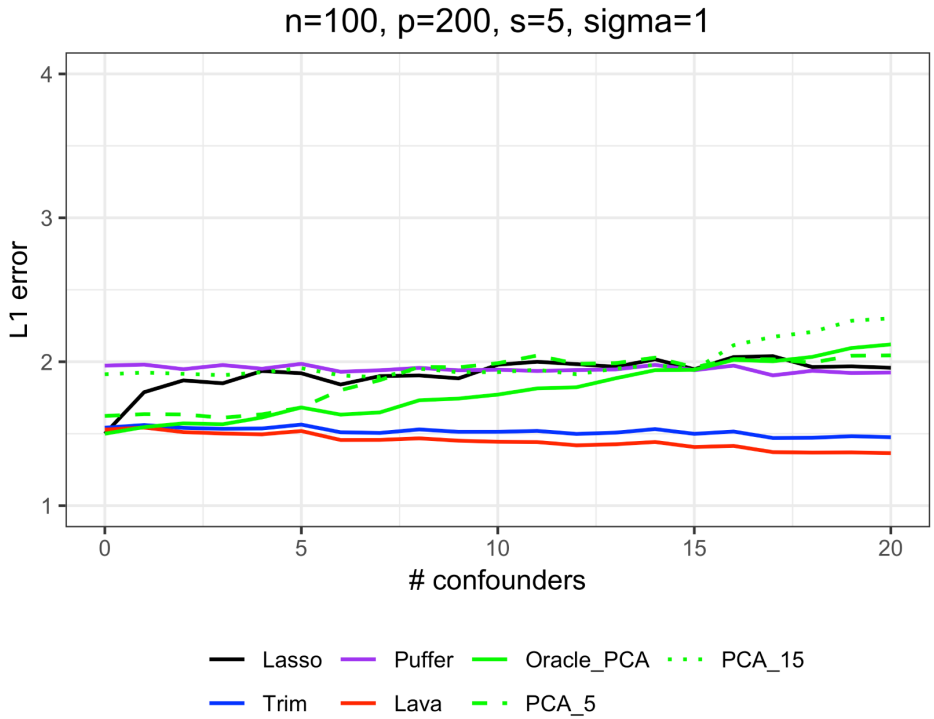
$\rightsquigarrow$ if only few (the number m is small) of the X-components are affected by hidden confounding variables, this and other techniques for adjustment must fail without further information (that is, without going to different settings)

Some numerical examples

$\|\hat{\beta} - \beta\|_1$ versus no. of confounders

left: the confounding model

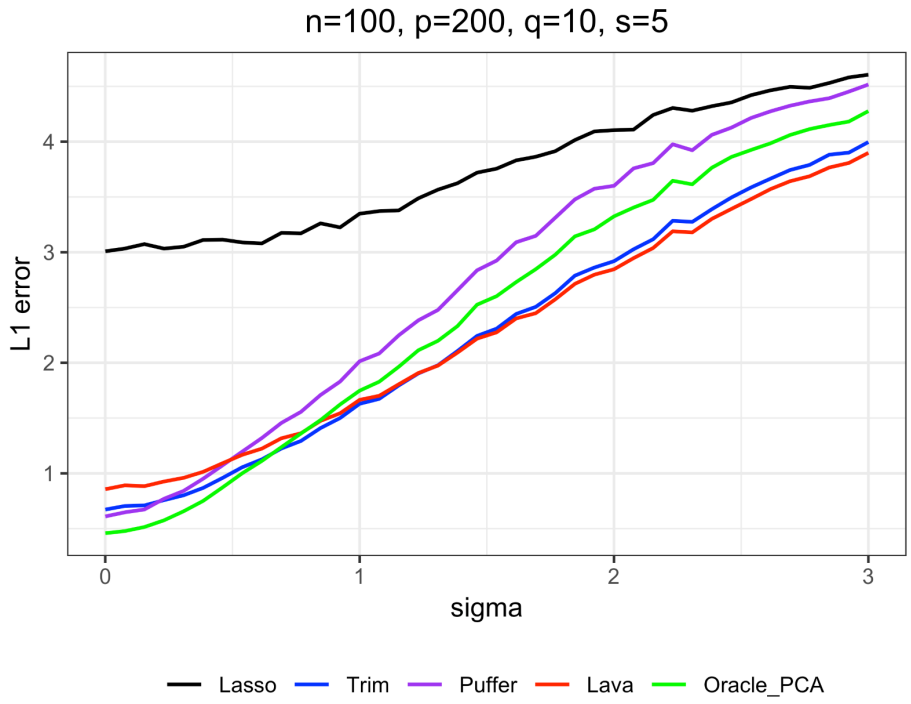


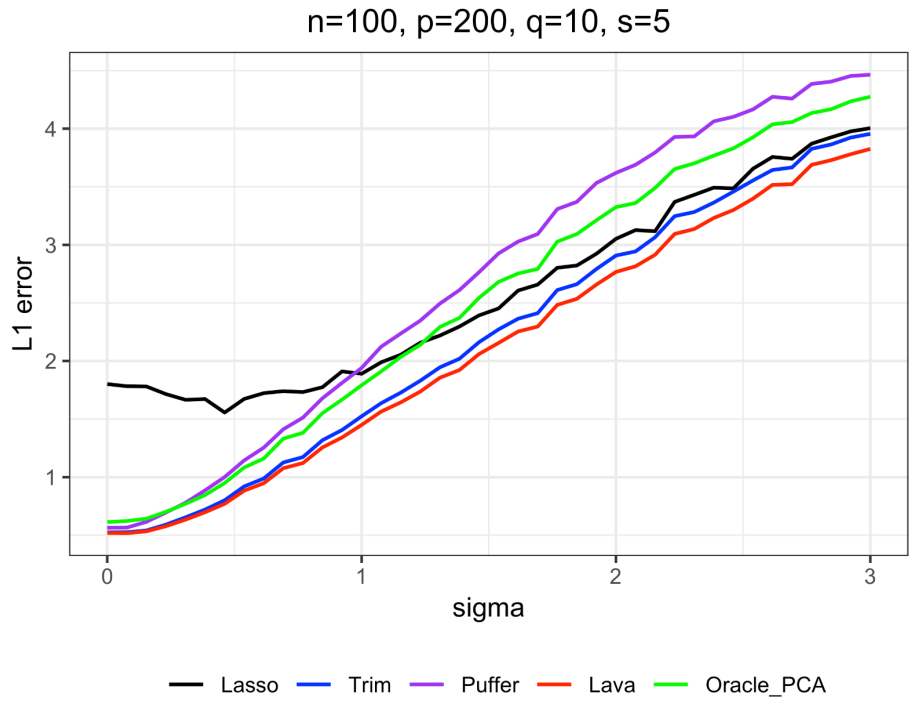


black: Lasso, blue: Trim transform, red: Lava, PCA adjustment

$\|\hat{\beta} - \beta\|_1$ versus σ

left: the confounding model

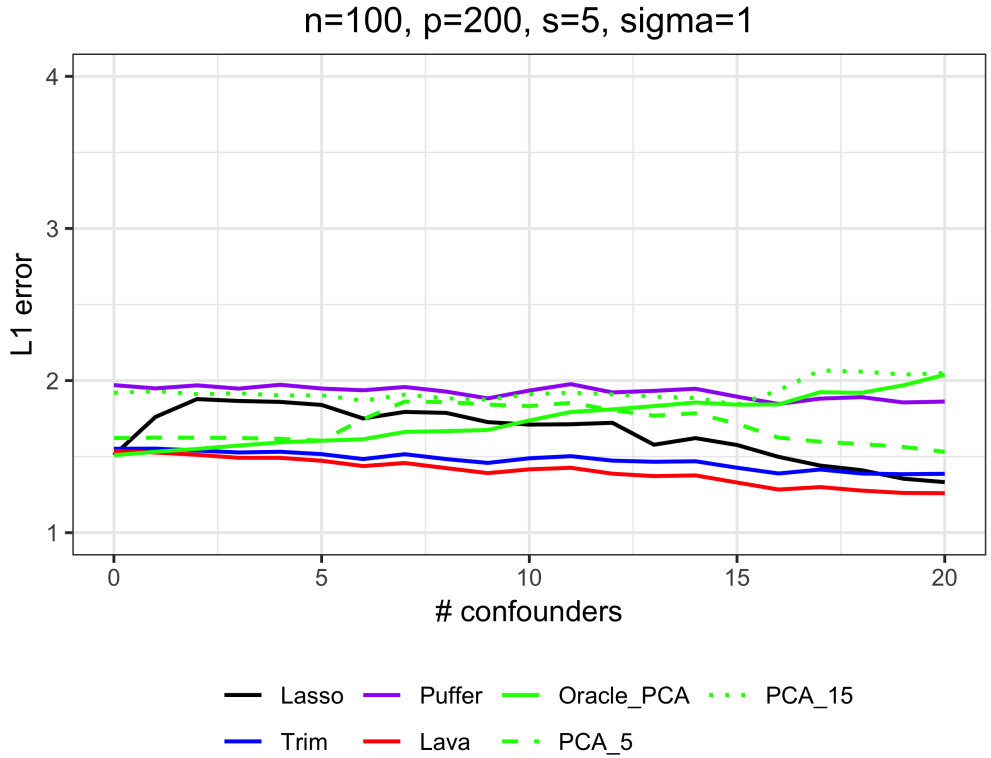




black: Lasso, blue: Trim transform, red: Lava, PCA adjustment

$\|\hat{\beta} - \beta\|_1$ versus no. of factors ("confounders")

but with $b = 0$ (no confounding)



black: Lasso, blue: Trim transform, red: Lava, PCA adjustment

using Trim transform does not hurt: plain Lasso is not better

using Trim transform does not hurt: plain Lasso is not better

spectral deconfounding leads to robustness against hidden confounders

- much improvement in presence of confounders
- (essentially) no loss in cases with no confounding!

Example from genomics (GTEx data)

a (small) aspect of GTEx data

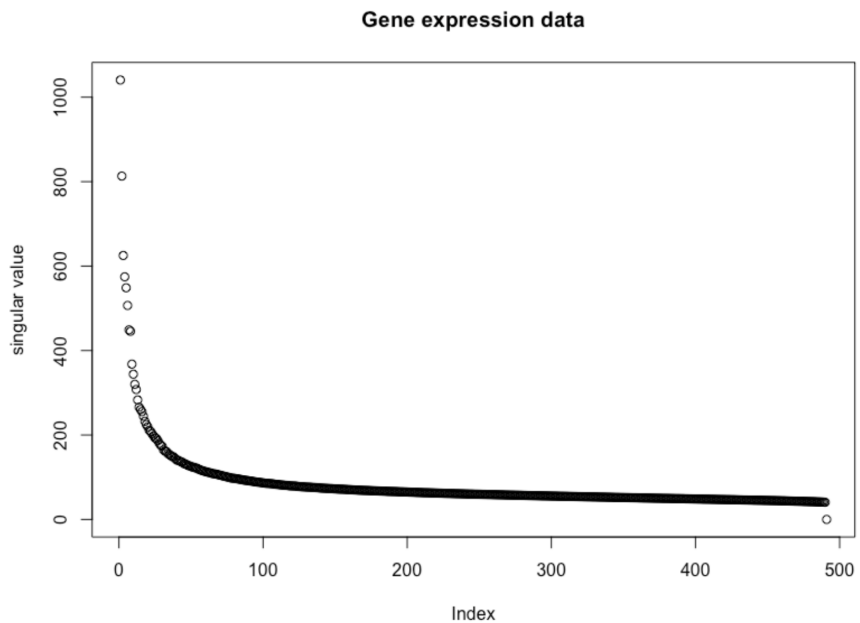
$p = 14713$ protein-coding gene expressions
$n = 491$ human tissue samples (same tissue)

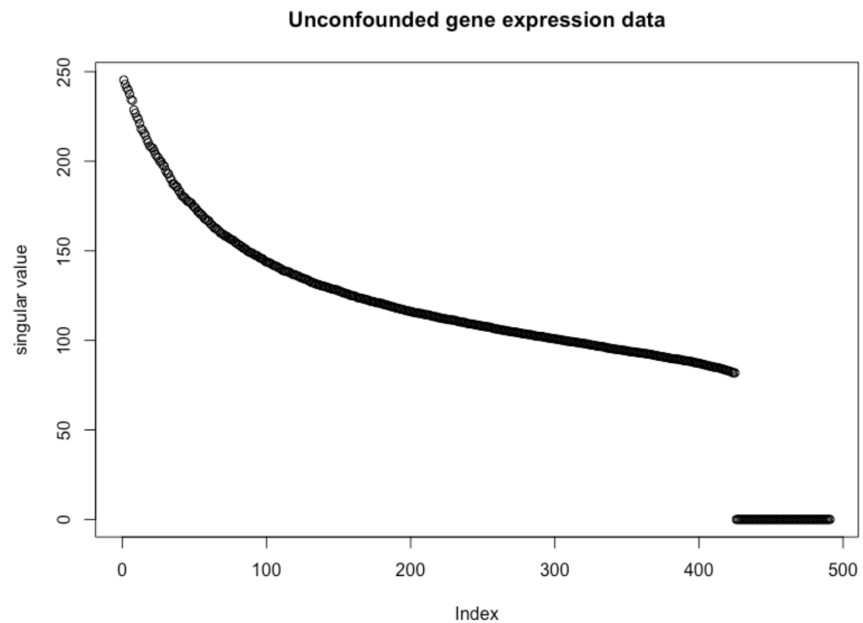
$q = 65$ different covariates which are proxys for hidden confounding variables

$\rightsquigarrow$ we can check robustness/stability of Trim transform in comparison to adjusting for proxys of hidden confounders

singular values of X

adjusted for 65 proxys of confounders





$\rightsquigarrow$ some evidence for factors, potentially being confounders

robustness/stability of selected variables

do we see similar selected variables for the original and the proxy-adjusted dataset?

- expression of one randomly chosen gene is response Y; all other gene expressions are the covariates X
- use a variable selection method $\hat{S} = \mathrm{supp}(\hat{\beta})$:

  $\hat{S}^{(1)}$ based on original dataset

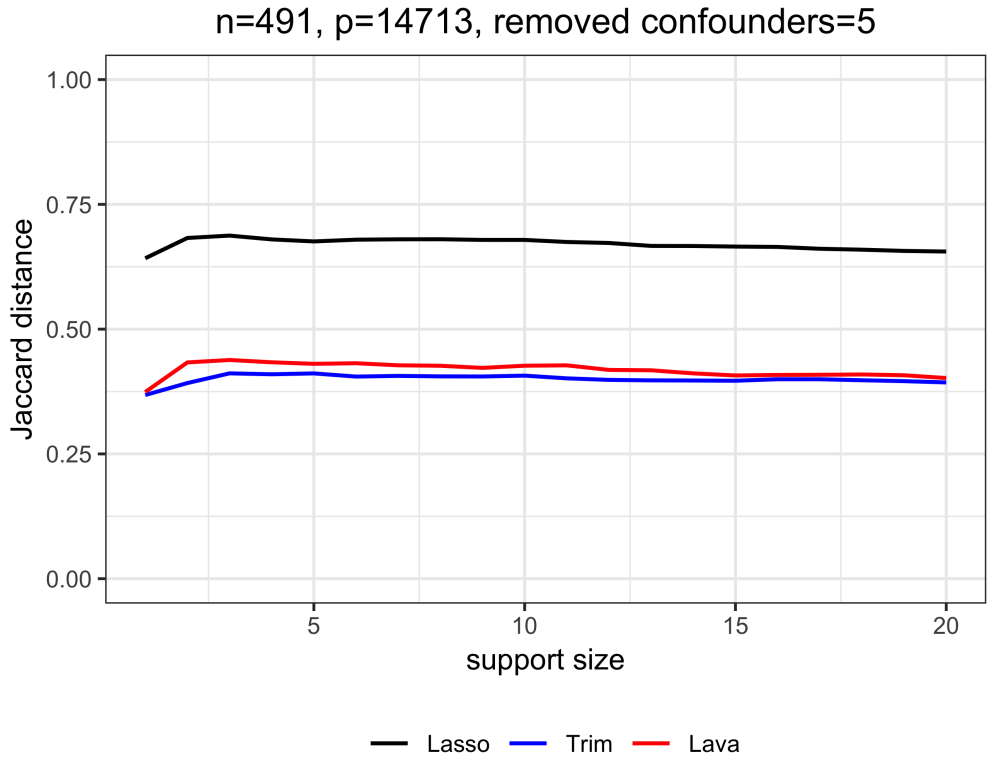
  $\hat{S}^{(2)}$ based on dataset adjusted with proxies

- compute Jaccard distance $d(\hat{S}^{(1)}, \hat{S}^{(2)}) = 1 - \frac{|\hat{S}^{(1)} \cap \hat{S}^{(2)}|}{|\hat{S}^{(1)} \cup \hat{S}^{(2)}|}$
- repeat over 500 randomly chosen genes

Jaccard distance $d(\text{supp}(\hat{\beta}_{\text{original}}, \text{supp}(\hat{\beta}_{\text{adjusted}})$ (vs. size) between original and adjusted data

averaged over 500 randomly chosen responses

adjusted for 5 proxy-confounders



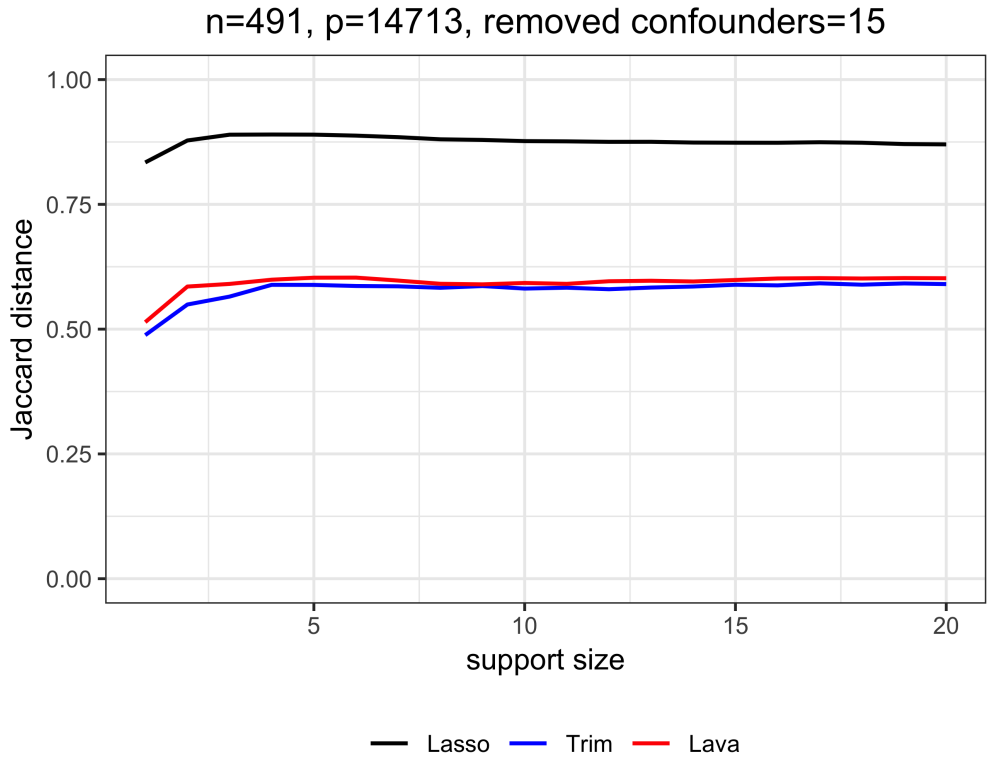
black: Lasso, blue: Trim transform, red: Lava

Trim transform (and Lava): more stable w.r.t. confounding

Jaccard distance $d(\text{supp}(\hat{\beta}_{\text{original}}, \text{supp}(\hat{\beta}_{\text{adjusted}})$ (vs. size) between original and adjusted data

averaged over 500 randomly chosen responses

adjusted for 15 proxy-confounders



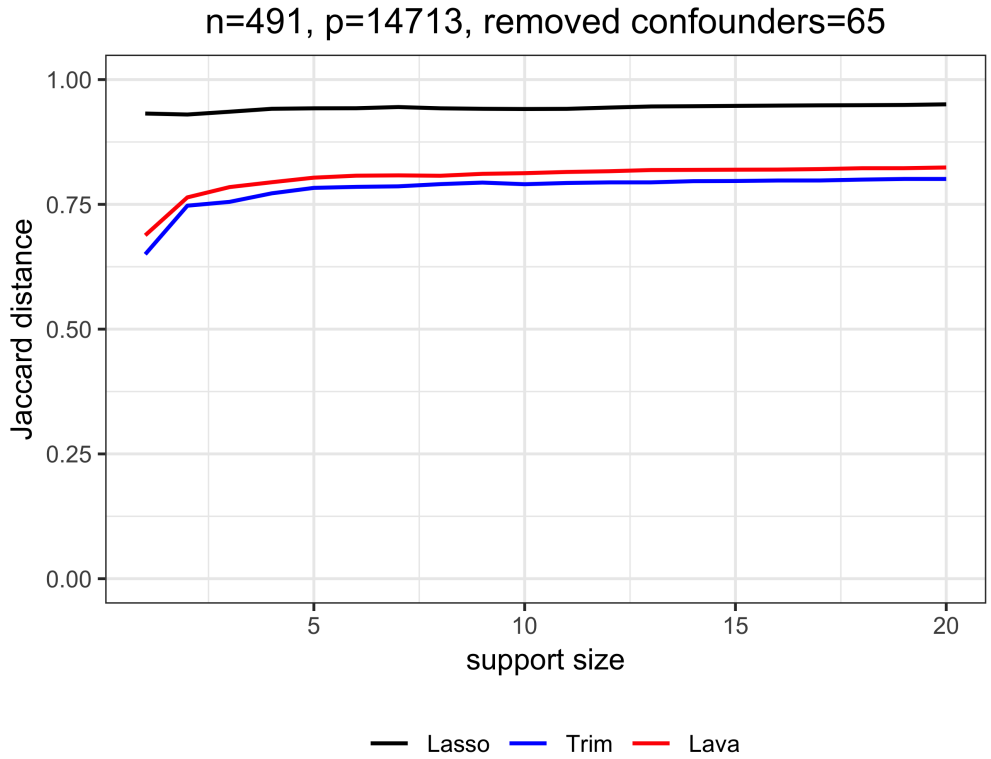
black: Lasso, blue: Trim transform, red: Lava

Trim transform (and Lava): more stable w.r.t. confounding

Jaccard distance $d(\text{supp}(\hat{\beta}_{\text{original}}, \text{supp}(\hat{\beta}_{\text{adjusted}})$ (vs. size) between original and adjusted data

averaged over 500 randomly chosen responses

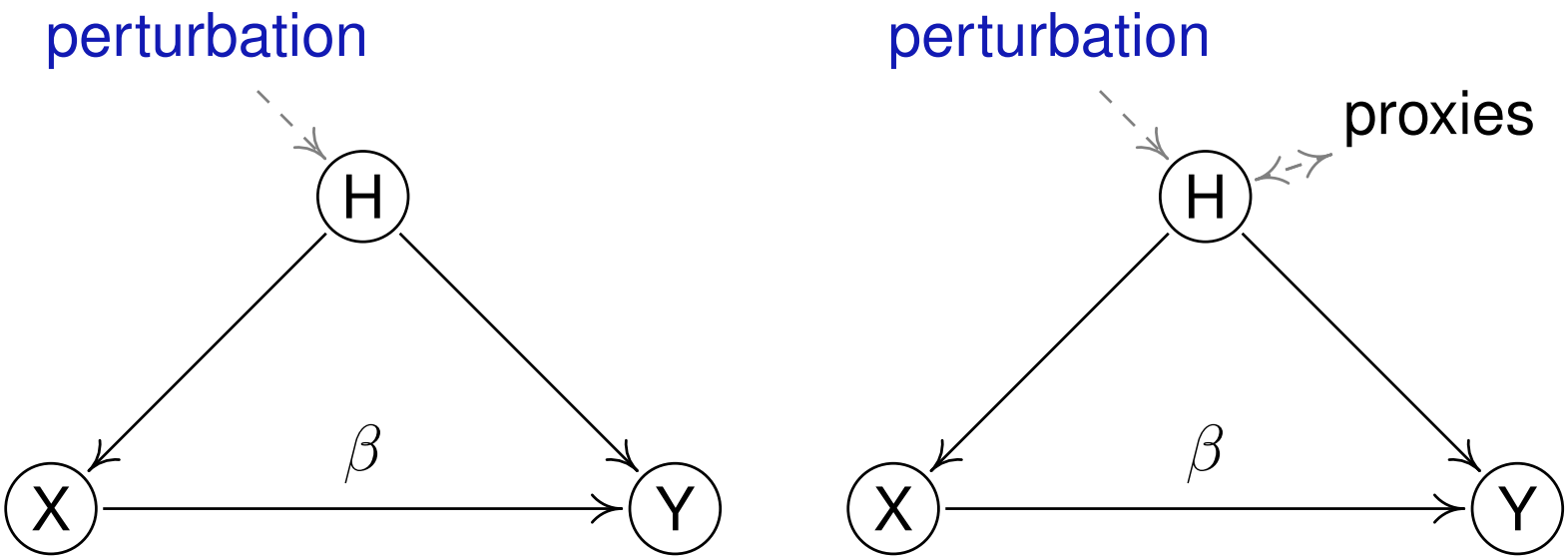
adjusted for 65 proxy-confounders



black: Lasso, blue: Trim transform, red: Lava

Trim transform (and Lava): more stable w.r.t. confounding

when “being able to do approximate deconfounding” $\rightsquigarrow$ more stability under perturbations of the hidden confounders



for replicability (reproducibility): want to be robust against heterogeneities or perturbations (of the hidden confounders)

$\rightsquigarrow$ see the results for the GTEx data

Spectral deconfounding: some conclusions

spectral deconfounding, especially the Trim transform:

- is extremely easy to use: linear transformation of X and Y
  (no tuning parameter with the default choice)
- leads to robustness of Lasso against hidden confounding and increases the "degree of replicability"
  with (essentially) no harm if there is no confounding and a standard linear model is correct

perhaps always to be used when aiming to interpret high-dimensional regression coefficients

Spectral deconfounding: some conclusions

spectral deconfounding, especially the Trim transform:

- is extremely easy to use: linear transformation of X and Y
  (no tuning parameter with the default choice)
- leads to robustness of Lasso against hidden confounding and increases the "degree of replicability"
  with (essentially) no harm if there is no confounding and a standard linear model is correct

perhaps always to be used when aiming to interpret high-dimensional regression coefficients

Conclusions

- causality can be framed as worst case risk optimization!
- causality can be inferred from invariance and a “stability” argument
- ICP (Invariant Causal Prediction) is a conceptual approach and method
  Causal Dantzig is more powerful and “makes more statistical sense”, at the price of restricting the interventions

- causality and distributional robustness are related to each other!

  causal regularization is a technique which enables a spectrum between invariance and “diluted causality”, and least squares (adjusted for anchor variables)
- there is much open space for improving distributional robustness (and hence performance) and interpretability beyond regression/classification association

  (invariance/“diluted causality” being one first example)

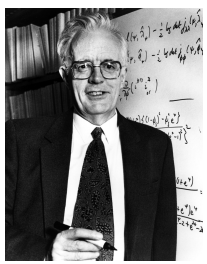
Conclusions

large on-going “dynamics” in data science, machine learn., “AI”, ...

in the topic area of this course but also in other fields:

“Statistical Thinking”






Tukey Fienberg Cox Wahba Efron Donoho

... ...

will remain to be important

Thank you!

Thank you!

I really enjoy(ed) being here!

References

- Bühlmann, P. (2018). Invariance, Causality and Robustness. To appear in Statistical Science. Preprint arXiv:1812.08233
- Ćevid, D., Bühlmann, P. and Meinshausen, N. (2018). Spectral deconfounding and perturbed sparse linear models. Preprint arXiv:1811.05352
- Meinshausen, N., Hauser, A., Mooij, J.M., Peters, J., Versteeg, P. and Bühlmann, P. (2016). Methods for causal inference from gene perturbation experiments and validation. Proceedings of the National Academy of Sciences USA 113, 7361-7368.
- Peters, J., Bühlmann, P. and Meinshausen, N. (2016). Causal inference using invariant prediction: identification and confidence intervals (with discussion). Journal of the Royal Statistical Society, Series B 78, 947-1012.
- Pfister, N., Bühlmann, P. and Peters, J. (2018). Invariant causal prediction for sequential data. Journal of the American Statistical Association, published online DOI 10.1080/01621459.2018.1491403.
- Rothenhäusler, D., Bühlmann, P. and Meinshausen, N. (2019). Causal Dantzig: fast inference in linear structural equation models with hidden variables under additive interventions. Annals of Statistics 47, 1688-1722.
- Rothenhüsler, D., Meinshausen, N., Bühlmann, P. and Peters, J. (2018). Anchor regression: heterogeneous data meets causality. Preprint arXiv:1801.06229.