Causal Regularization for Distributional Robustness and Replicability

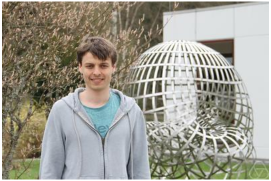
Peter Bühlmann

Seminar for Statistics, ETH Zürich

Supported in part by the European Research Council under the Grant Agreement No. 786461 (CausalStats - ERC-2017-ADG)

Acknowledgments

Dominik Rothenhäusler
Stanford University

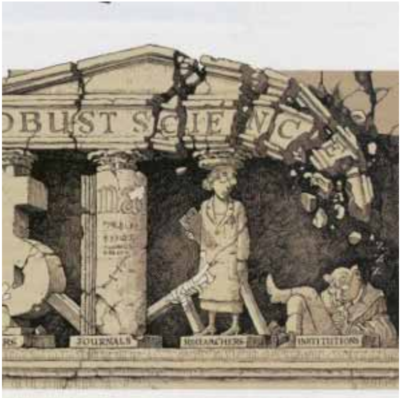


Niklas Pfister
ETH Zürich

Jonas Peters
Univ. Copenhagen

Nicolai Meinshausen
ETH Zürich

The replicability crisis in science



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

John P.A. Ioanidis
(School of Medicine, courtesy appoint. Statistics, Stanford)

Ioanidis (2005): Why Most Published Research Findings Are False (PLOS Medicine)

one among possibly many reasons:
(statistical) methods may not generalize so well...

Single data distribution and accurate inference

say something about generalization to a population from the same distribution as the observed data

Graunt & Petty (1662), Arbuthnot (1710), Bayes (1761), Laplace (1774), Gauss (1795, 1801, 1809), Quetelet (1796-1874),..., Karl Pearson (1857-1936), Fisher (1890-1962), Egon Pearson (1895-1980), Neyman (1894-1981), ...

Bayesian inference, bootstrap, high-dimensional inference, selective inference, ...

Generalization to new data distributions

generalization beyond the population distributions(s) in the data
replicability for new data generating distributions

setting:
observed data from distribution P_0

want to say something about new $P' \neq P_0$

Generalization to new data distributions

generalization beyond the population distributions(s) in the data
replicability for new data generating distributions

setting:
observed heterogeneous data from distributions P_e $(e \in \mathcal{E})$
$\mathcal{E} =$ observed sub-populations

want to say something about new $P_{e'}$ $(e' \notin \mathcal{E})$

$\leadsto$ “some kind of extrapolation”

$\leadsto$ “some kind of causal thinking” can be useful
(as I will try to explain)

see also “transfer learning” from machine learning (cf. Pan and Yang)

GTEx data

Genotype-Tissue Expression (GTEx) project

a (small) aspect of entire GTEx data:

- 13 different tissues, corresponding to $\mathcal{E} = \{1, 2, \ldots, 13\}$
- gene expression measurements for 12’948 genes
  (one of them is the response, the other are covariates)
  sample size between 300 - 700
- we aim for:
  prediction for new tissues $e' \notin \mathcal{E}$
  replication of results on new tissues $e' \notin \mathcal{E}$

it’s very noisy and high-dimensional data!

“Causal thinking”

we want to generalize/transfer to new situations with new unobserved data generating distributions

causality: is giving a prediction (a quantitative answer) to a “what if I do/perturb” question but the perturbation (aka “new situation”) is not observed many modern applications are faced with such prediction tasks:

- genomics: what would be the effect of knocking down (the activity of) a gene on the growth rate of a plant?

  we want to predict this without any data on such a gene knock-out (e.g. no data for this particular perturbation)
- E-commerce: what would be the effect of showing person "*XYZ*" an advertisement on social media?
  no data on such an advertisement campaign for "*XYZ*" or persons being similar to "*XYZ*"
- etc.

Heterogeneity, Robustness and a bit of causality

assume heterogeneous data from different known observed environments or experimental conditions or perturbations or sub-populations $e \in \mathcal{E}$:

$$(X^e, Y^e) \sim P_e, \quad e \in \mathcal{E}$$

with response variables Y^e and predictor variables X^e

examples:
- data from 10 different countries
- data from 13 different tissue types in GTEx data

consider "many possible" but mostly non-observed environments/perturbations $\mathcal{F} \supset \underbrace{\mathcal{E}}_{\text{observed}}$

examples for $\mathcal{F}$:
- 10 countries and many other than the 10 countries
- 13 different tissue types and many new ones (GTEx example)

problem:
predict Y given X such that the prediction works well (is "robust"/"replicable") for *"many possible"* new environments $e \in \mathcal{F}$ based on data from much fewer environments from $\mathcal{E}$

trained on designed, known scenarios from $\mathcal{E}$

trained on designed, known scenarios from $\mathcal{E}$

new scenario from $\mathcal{F}$!

a pragmatic prediction problem:
predict Y given X such that the prediction works well (is “robust”/“replicable”) for *“many possible”* environments $e \in \mathcal{F}$ based on data from much fewer environments from $\mathcal{E}$

for example with linear models: find

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

$$\text{it is } \underbrace{\text{“robustness”}}_{\text{distributional robust.}}$$

a pragmatic prediction problem:
predict Y given X such that the prediction works well (is “robust”/“replicable”) for *“many possible”* environments $e \in \mathcal{F}$ based on data from much fewer environments from $\mathcal{E}$

for example with linear models: find

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

$$\text{it is } \underbrace{\text{“robustness”}}_{\text{distributional robust.}}$$

a pragmatic prediction problem:
predict Y given X such that the prediction works well (is “robust”/“replicable”) for *“many possible”* environments $e \in \mathcal{F}$ based on data from much fewer environments from $\mathcal{E}$

for example with linear models: find

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

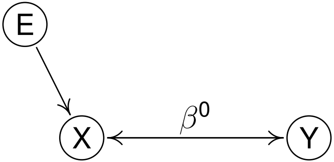
it is $\underbrace{\text{“robustness”}}_{\text{distributional robust.}}$ **and** causality

Causality and worst case risk

for linear models: in a nutshell

for $\mathcal{F} = \{$all perturbations not acting on Y directly$\}$,

$$\mathrm{argmin}_{\beta} \max_{e \in \mathcal{F}} \mathbb{E}|Y^e - X^e \beta|^2 = \text{ causal parameter} = \beta^0$$



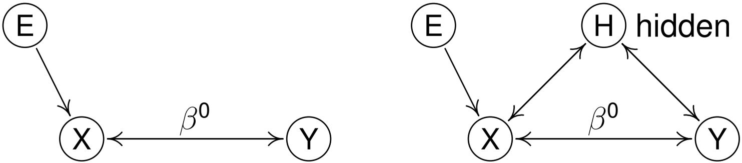
that is:
causal parameter optimizes
worst case loss w.r.t. “very many” unseen (“future”) scenarios

Causality and worst case risk

for linear models: in a nutshell

for $\mathcal{F} = \{$all perturbations not acting on Y directly$\}$,

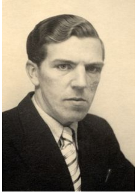
$$\mathrm{argmin}_{\beta} \max_{e \in \mathcal{F}} \mathbb{E}|Y^e - X^e \beta|^2 = \text{ causal parameter } = \beta^0$$



that is:
causal parameter optimizes
worst case loss w.r.t. “very many” unseen (“future”) scenarios

causal parameter optimizes
worst case loss w.r.t. "very many" unseen ("future") scenarios
no causal graphs or potential outcome models (Neyman, Holland, Rubin, ..., Pearl, Spirtes, ...)

causality and distributional robustness are intrinsically related (Haavelmo, 1943)

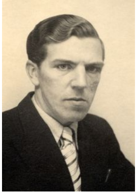


Trygve Haavelmo, Nobel Prize in Economics 1989

$\mathcal{L}(Y^e | X^e_{\text{causal}})$ remains invariant w.r.t. e
causal structure $\Longrightarrow$ invariance/"robustness"

causal parameter optimizes
worst case loss w.r.t. “very many” unseen (“future”) scenarios
no causal graphs or potential outcome models (Neyman, Holland, Rubin, ..., Pearl, Spirtes, ...)

causality and distributional robustness are intrinsically related
(Haavelmo, 1943)



Trygve Haavelmo, Nobel Prize in Economics 1989

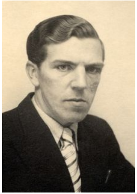
$\mathcal{L}(Y^e|X^e_{\text{causal}})$ remains invariant w.r.t. e

causal structure $\Longleftarrow$ invariance

(Peters, PB & Meinshausen, 2016)

causal parameter optimizes
worst case loss w.r.t. “very many” unseen (“future”) scenarios

causality and distributional robustness are intrinsically related
(Haavelmo, 1943)



Trygve Haavelmo, Nobel Prize in Economics 1989

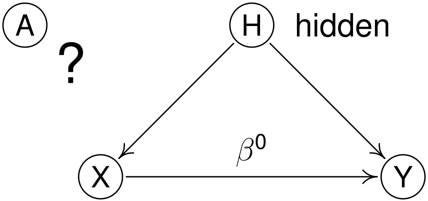
causality $\Longleftrightarrow$ invariance/“robustness”

and novel causal regularization allows to exploit this relation

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

(Rothenhäusler, Meinshausen, PB & Peters, 2018)

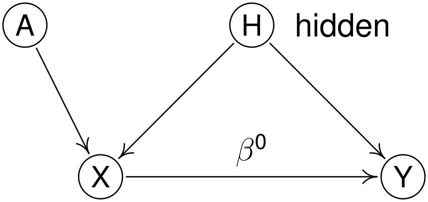
the environments from before, denoted as e:
they are now outcomes of a variable $\underbrace{A}_{\text{anchor}}$



Anchor regression and causal regularization

(Rothenhäusler, Meinshausen, PB & Peters, 2018)

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



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

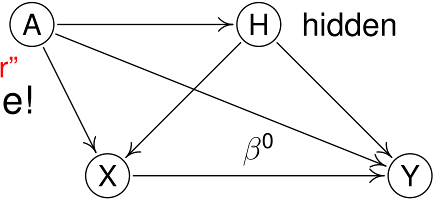
Instrumental variables regression model
(cf. Angrist, Imbens, Lemieux, Newey, Rosenbaum, Rubin,...)

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

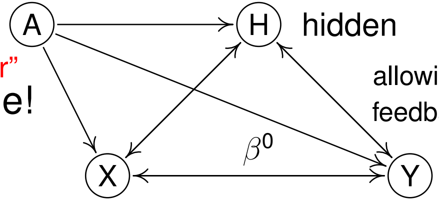
$$\begin{pmatrix} X \\ Y \\ H \end{pmatrix} = B \begin{pmatrix} X \\ Y \\ H \end{pmatrix} + \varepsilon + MA$$

Anchor regression and causal regularization

(Rothenhäusler, Meinshausen, PB & Peters, 2018)

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

A is an “anchor”
source node!



allowing also for
feedback loops

$\rightsquigarrow$ Anchor regression

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

allow that A acts on Y and H

$\leadsto$ 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:

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

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

- for $\gamma = 1$: least squares
- for $0 \leq \gamma < \infty$: general causal regularization

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

causal regularization:

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

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

- for $\gamma = 1$: least squares + ℓ_1-penalty
- for $0 \le \gamma < \infty$: general causal regularization + ℓ_1-penalty

convex optimization problem

... 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 shift perturbations $\mathcal{F}$ (in test data)
shift vectors v

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

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

i.e. $v \in C_\gamma = \{v;\ v = Mu$ for some u 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

$\rightsquigarrow$ just regularize! (instead of l.h.s. which is a difficult object) for any β

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

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

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

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

$\rightsquigarrow$ and “therefore” also finite sample guarantees for predictive stability (i.e. optimizing a worst case risk)

(we have worked out all the details)

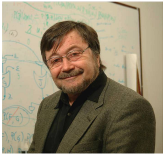
distributional robustness $\longleftrightarrow$ causal regularization

Adversarial Robustness

machine learning, Generative Networks

Causality

e.g. Ian Goodfellow



e.g. Judea Pearl

and indeed, one can improve prediction
with causal-type regularization

- image classification with CNN

  (Heinze-Deml and Meinshausen, 2017)

  for problems with domain shift: gross improvement over non-regularized standard optimization
- causal-robust machine learning

  Leon Bouttou et al. since 2013 (Microsoft and now Facebook)

other examples:

- UCI machine learning and Kaggle datasets
- macro-economics (MSc thesis with KOF Swiss Economic Institute)

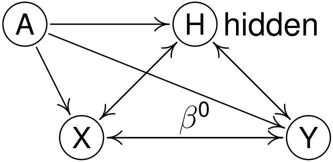
$\rightsquigarrow$ small ($\approx$ 5%) but persistent gains

Science aims for causal understanding

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

causal inference necessarily requires (often untestable) additional assumptions

e.g. in anchor regression model: we cannot find/identify the causal (“systems”) parameter β^0



Invariance and "diluted causality"

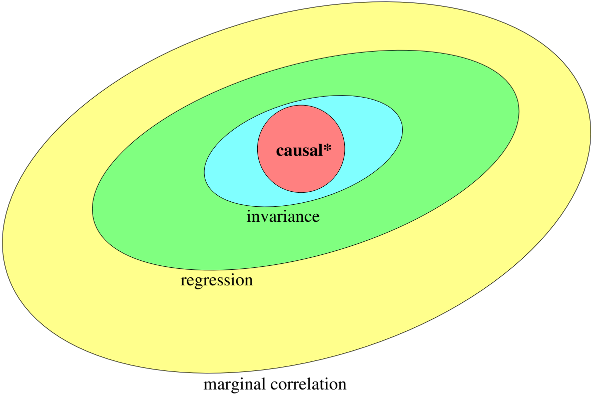
by the fundamental duality in anchor regression:

$\gamma \to \infty$ leads to shift invariance of residuals

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

$b^{\to\infty}$ 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



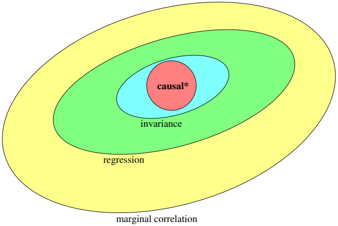
under faithfulness conditions, the figure is valid (causal* are the causal variables as in e.g. large parts of Dawid, Pearl, Robins, Rubin, ...)

Stabilizing

John W. Tukey (1915 – 2000)

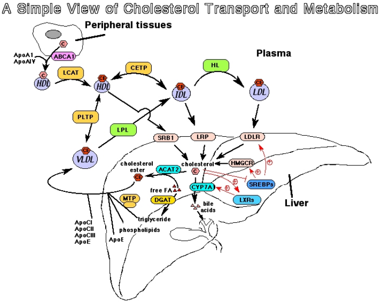
Tukey (1954)

> *“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”: important proteins for cholesterol

Ruedi Aebersold, ETH Zürich



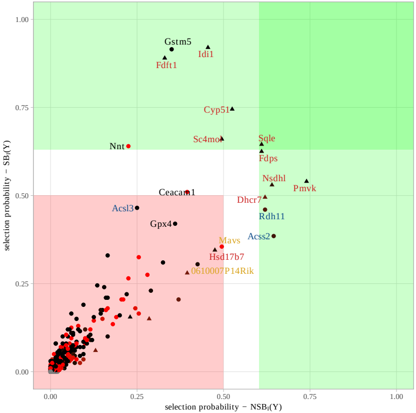
3934 other proteins
which of those are
“diluted causal”
for cholesterol

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

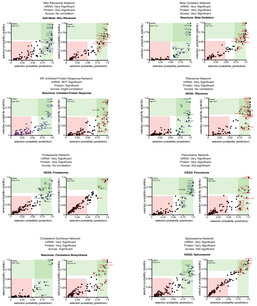
high-dimensional regression, total sample size $n = 270$

Y = cholesterol pathway activity, X = 3934 protein expressions x-axis: importance w.r.t regression but non-invariant

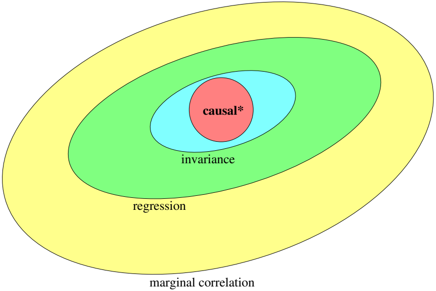
y-axis: importance w.r.t. invariance



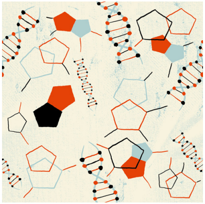
beyond cholesterol: with transcriptomics and proteomics



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

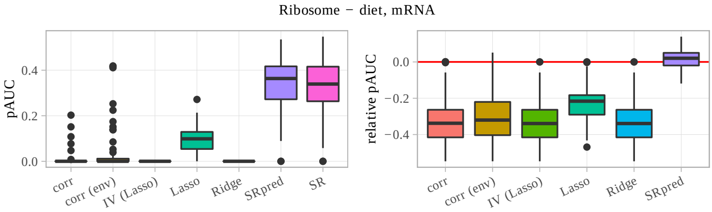


and we actually find promising candidates



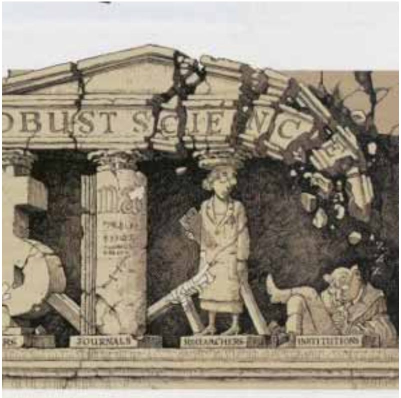
we “checked” in independent datasets the top hits
⇝ has worked “quite nicely”

further “validation” with respect to finding known pathways (here for Ribosome pathway)



Distributional Replicability

The replicability crisis



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

more severe issue than just “accurate confidence”, “selective inference”, ...

The “diluted causal” parameter $b^{\rightarrow\infty}$ is replicable

assume

- new dataset for replication arises from shift perturbations (as before)
- a practically checkable so-called projectability condition

  $\inf_b \mathbb{E}[Y - Xb|A] = 0$

consider

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

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

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

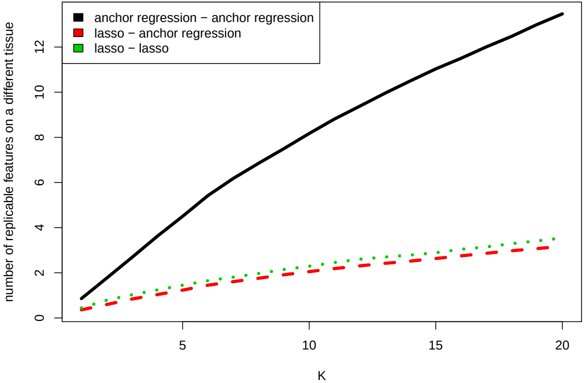
Replicability for $b^{\rightarrow\infty}$ in GTEx data across tissues

- 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

$\leadsto$ are they replicable on other tissues?

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



x-axis: number K for the top K features
y-axis: overlap of the top K ranked variables/features
(found by a method on tissue t and on tissue $t' \neq t$)

averaged over all 13 t and averaged over 1000 random choices 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!

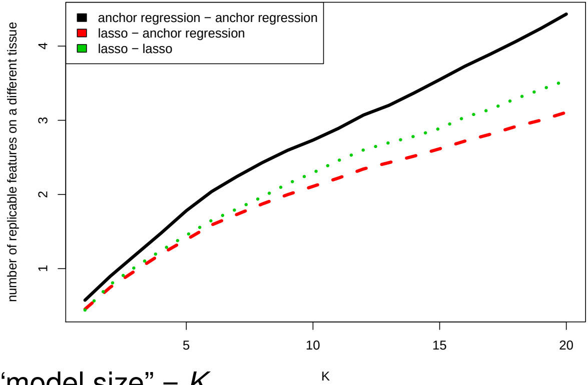
assume:

- anchor stability
- projectability condition

$\leadsto$ 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



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 (not shown today): based on transcriptomics
  (Meinshausen, Hauser, Mooij, Peters, Versteeg, and PB, 2016)
- large-scale kinetic systems (not shown today): based on metabolomics (Pfister, Bauer and Peters, 2019)

Conclusions

- causal regularization is for the population case
  (not because of “complexity” in relation to sample size)
  $\rightsquigarrow$ distributional robustness and replicability
  (not claiming to find “truly causal” structure)
- the key is to exploit certain invariances

- anchor regression (with γ large) justifies instrumental variables regression when IV assumptions are violated
  $\rightsquigarrow$ “diluted causality” and invariance of residuals

make heterogeneity or non-stationarity your friend

(rather than your enemy)!

make heterogeneity or non-stationarity your friend

(rather than your enemy)!

Theorem (Rothenhäusler, Meinshausen, PB & Peters, 2018)

assume:

- a "causal" compatibility condition on X (weaker than the standard compatibility condition);
- (sub-) Gaussian error;
- $\dim(A) \le C < \infty$ for some C;

Then, for $R_\gamma(u) = \max_{v \in C_\gamma} \mathbb{E}|Y^v - X^v u|^2$ and any $\gamma \ge 0$:

$$R_\gamma(\hat{\beta}_\gamma) = \underbrace{\min_u R_\gamma(u)}_{\text{optimal}} + O_P(s_\gamma \sqrt{\log(d)/n}),$$

$$s_\gamma = \mathrm{supp}(\beta_\gamma),\ \beta_\gamma = \mathrm{argmin}_\beta R_\gamma(u)$$

if dim(A) is large: use ℓ_∞-norm causal regularization

- good for identifiability (lots of heterogeneity) regularization
- a statistical price of $\log(|A|)$

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

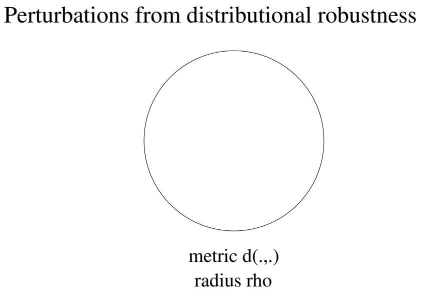
$$\text{armin}_{\beta} \max_{P \in \mathcal{P}} \mathbb{E}_P[(Y - X\beta)^2]$$

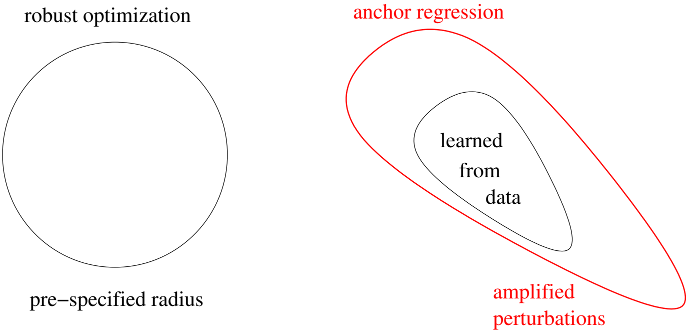
perturbations are within a class of distributions

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

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

often as Wasserstein distance





robust optimization

pre–specified radius

anchor regression

learned from data

amplified perturbations

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