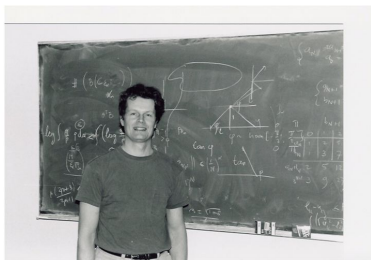


Replicability

(when new distributions arise)



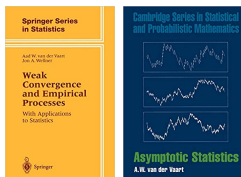
Amsterdam, 2000

Happy Birthday Aad!

Peter Bühlmann

Seminar for Statistics, ETH Zürich

- ▶ $\approx$ 1992: Aad gave a seminar talk at ETH Zurich (and I think I've met him again at the EMS in Bath, 1992)



more as a user of the techniques and machinery...

fascinated by the clarity of Aad's writing/presentation style
admired his mathematical/statistical insights



dinner in Singapore 2012



festive dinner in Zürich, 2015

Distributional Replicability

main collaborators:



Dominik Rothenhäusler
now UC Berkeley

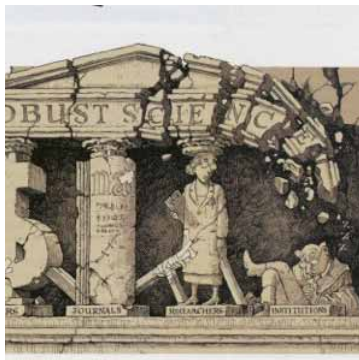


Jonas Peters
Univ. Copenhagen



Nicolai Meinshausen
ETH Zürich

The replicability crisis



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

I am sure you have all heard about this...
but I want to present a somewhat different viewpoint

we cannot replicate Aad!

we cannot replicate Aad!

- ▶ it is impossible to replicate his books

- A. van der Vaart (1988).

Statistical Estimation in Large Parameter Spaces. CWI tract 44.



his scientific and mathematical insights

e.g. Ghosal, S. and van der Vaart, A. (2000). Convergence rates of posterior distributions. Annals of Statistics

...

- ▶ *he is truly unique!*

we cannot replicate Aad!

- ▶ it is impossible to replicate his books

- A. van der Vaart (1988).

Statistical Estimation in Large Parameter Spaces. CWI tract 44.



his scientific and mathematical insights

e.g. Ghosal, S. and van der Vaart, A. (2000). Convergence rates of posterior distributions. Annals of Statistics

...

- ▶ *he is truly unique!*

► it is impossible to replicate his books

- A. van der Vaart (1988).

Statistical Estimation in Large Parameter Spaces. CWI tract 44.



his scientific and mathematical insights

...

► *he is truly unique!*

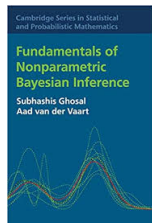
and thus... we realize that there is a replicability crisis and why it is impossible to address it completely

but nevertheless...

we should at least try to partially address the issue!

Single data distribution and accurate inference

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



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**)



Genotype-Tissue Expression (GTEx) project

a (small) aspect of entire GTEx data:

- ▶ 13 different tissues, corresponding to $\mathcal{E} = \{1, 2, \dots, 13\}$
- ▶ 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

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

it is “robustness”

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

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

it is “robustness”

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

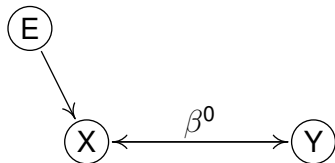
$$\operatorname{argmin}_{\beta} \max_{e \in \mathcal{F}} \mathbb{E} |Y^e - X^e \beta|^2$$

it is “robustness” **and** causality

Causality and worst case risk

for linear models: in a nutshell

for $\mathcal{F} = \{\text{all perturbations not acting on } Y \text{ directly}\}$,
$$\operatorname{argmin}_{\beta} \max_{e \in \mathcal{F}} \mathbb{E} |Y^e - X^e \beta|^2 = \text{causal parameter} = \beta^0$$



that is:

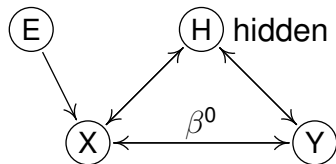
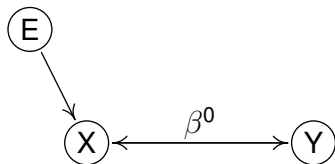
causal parameter optimizes

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

Causality and worst case risk

for linear models: in a nutshell

for $\mathcal{F} = \{\text{all perturbations not acting on } Y \text{ directly}\}$,
 $\operatorname{argmin}_{\beta} \max_{e \in \mathcal{F}} \mathbb{E} |Y^e - X^e \beta|^2 = \text{causal parameter} = \beta^0$



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

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



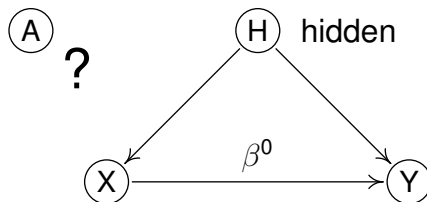
Trygve Haavelmo, Nobel Prize in Economics 1989

and 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 A
anchor

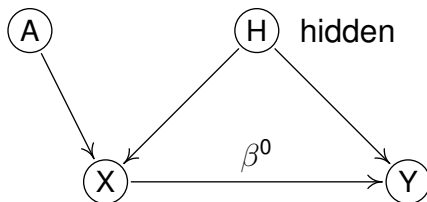


Anchor regression and causal regularization

(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

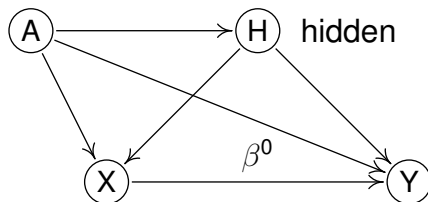
(cf. Angrist, Imbens, Lemieux, Newey, Rosenbaum, Rubin,...)

Anchor regression and causal regularization

(Rothenhäusler, Meinshausen, PB & Peters, 2018)

the environments from before, denoted as e :

they are now outcomes of a variable A
anchor



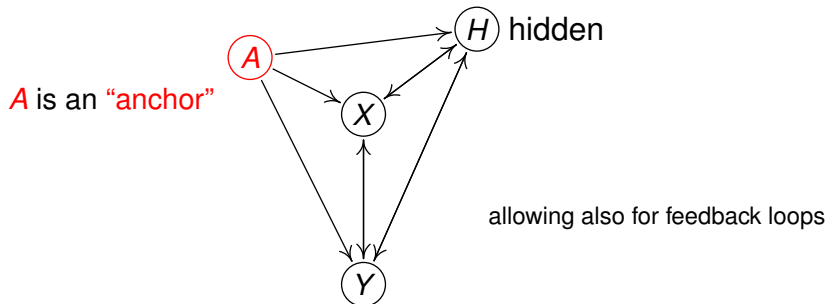
$\leadsto$ Anchor regression

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

Anchor regression with hidden confounders

allow that A acts on Y and H

more realistic but has been believed in the past as “ill-posed”



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

allow that A acts on Y and H

~> 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

namely: to find a parameter vector β such that the residuals $(Y - X\beta)$ **stabilize**, have the same distribution across different perturbations of A = environments/sub-populations

we want to encourage orthogonality of A with the residuals something like

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

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

causal regularization:

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

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

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

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

causal regularization:

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

$\Pi_A = A(A^T A)^{-1} A^T$ (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

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

but causal regularization solves for

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

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

recap: causal parameter solves for

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

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

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

worst case shift interventions $\longleftrightarrow$ regularization!
in the population case

for any β

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

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

and “therefore”

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

leads to

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

and indeed, we can improve prediction

- ▶ UCI machine learning and Kaggle datasets
- ▶ macro-economics
(MSc thesis with KOF Swiss Economic Institute)
→ small ($\approx 5\%$) but persistent gains

The parameter $b^{\rightarrow\infty}$

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

by the fundamental duality: $b^{\rightarrow\infty}$ is the parameter which optimizes worst case prediction risk over shift interventions of arbitrary strength

it is generally not the causal parameter

but leads to **shift invariance** of residuals

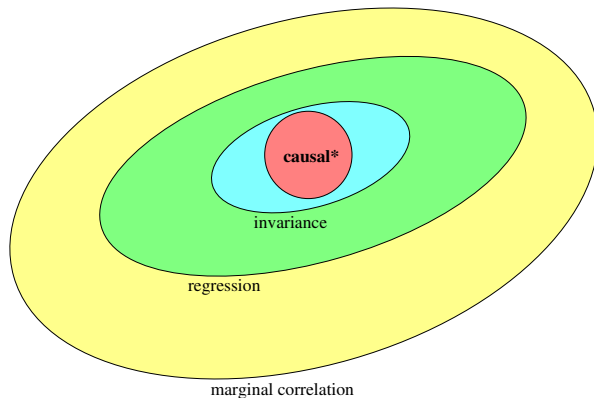
$\leadsto$ name it **“diluted causal”**

– thanks to Ed George!



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 **Robins, Rubin, Pearl, Dawid, ...**)



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.”

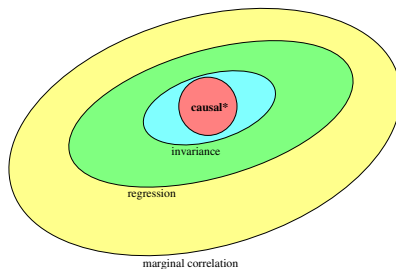


Tukey (1954)

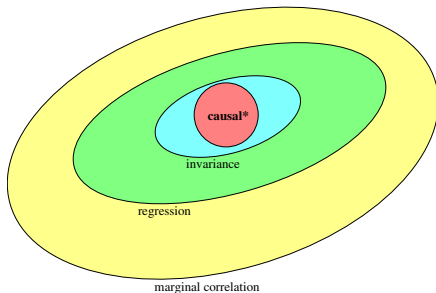
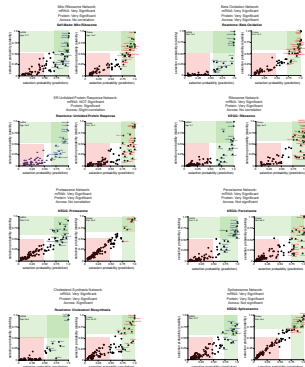
John W. Tukey (1915 – 2000)

*“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.”



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



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 $b^{\rightarrow\infty}$ is replicable

assume

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

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

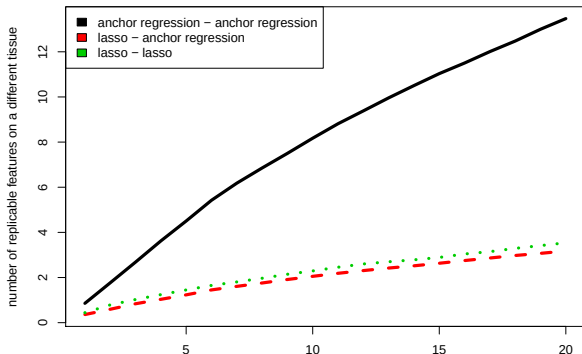


- ▶ 13 tissues
- ▶ gene expression measurements for 12'948 genes, sample size between 300 - 700
- ▶ Y = expression of a target gene
 X = expressions of all other genes
 A = 65 PEER factors (potential confounders)

estimation and findings on one tissue

~> are they replicable on other tissues?

Replicability for $b \rightarrow \infty$ 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**:

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

checkable!

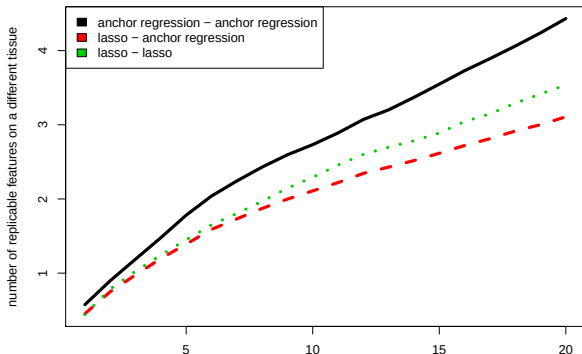
assume:

- ▶ 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

Conclusions

- ▶ causal regularization – including the population case!
leads to distributional robustness and replicability
(not claiming to find “truly causal” structure)
- ▶ the key is to enforce certain invariances

Conclusions

- ▶ causal regularization – including the population case!
leads to distributional robustness and replicability
(not claiming to find “truly causal” structure)
- ▶ the key is to enforce certain invariances

Replicability is challenging –
but it's worth (trying) to improve on it



Happy Birthday Aad!

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

assume:

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

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

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

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

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

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