Brown Bag Seminar

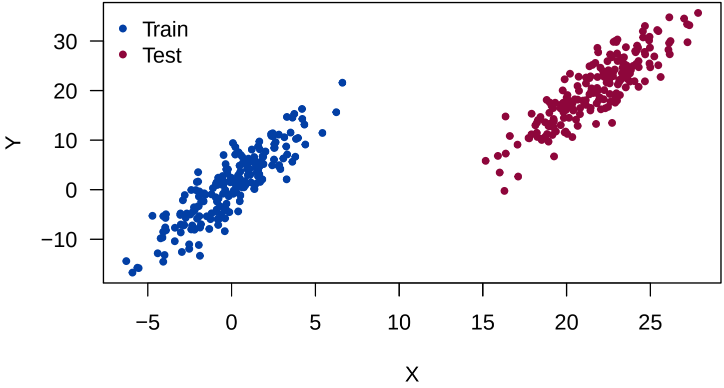
Distributional anchor regression

Lucas Kook
University of Zurich
Zurich University of Applied Sciences

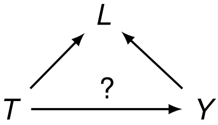
June 18, 2020

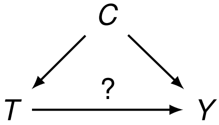
Motivation

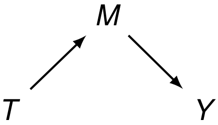
We want to robustly predict an outcome in heterogenous data with potentially unseen perturbations in the test data.



The "Causal Revolution"





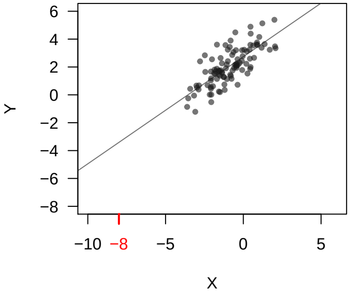


Structural causal models,
Bayesian networks
and causal calculus

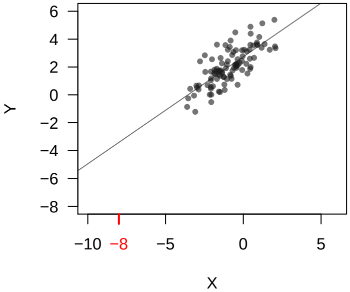
Judea Pearl (Source)

Potential outcomes: What if?

$X \longrightarrow Y$

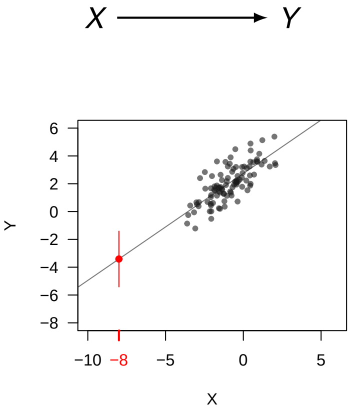


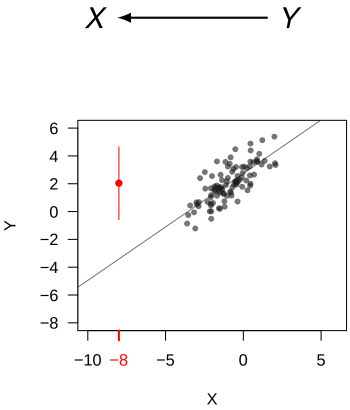
$X \longleftarrow Y$



What is our best prediction for Y if we do$(X = -8)$?

Potential outcomes: What if?





How do we know which one is the right model?

Robustness

"If the answer is highly sensitive to perturbations, you have probably asked the wrong question."

– Lloyd N. Trefethen

Our aim:

Predict the outcome, such that the prediction is robust towards "perturbations" in future data

Robustness

Predict the outcome, such that the prediction is robust towards "perturbations" in future data

These are **future, yet unobserved** perturbations, e.g.,

- data from a different country,
- different point in time,
- different experimental setting,
- different environment,
- ...

Causality and robustness

Haavelmo (1943)

Causal variables $\Rightarrow$ Robustness

Peters et al. (2016)

Causal structures $\Leftarrow$ Robustness

T. Haavelmo (Source)

Formalize our aim

Data from **observed** environments:

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

Only part of larger class of **unobserved** environments:

$$\mathcal{F} \supset \mathcal{E}$$

Predict Y^e given X^e such that the prediction is robust for all $e \in \mathcal{F}$ based on data from much fewer environments $e \in \mathcal{E}$.

Bühlmann (2018)

Connection to causality

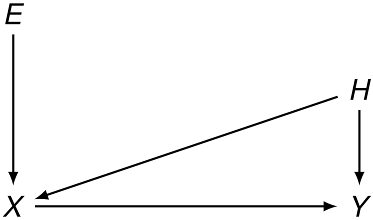
$$\begin{tikzcd} E \arrow[d] & \\ X \arrow[r, "\text{not depending on } E"'] & Y \end{tikzcd}$$

Connection to causality:

$$\arg\min_{\beta} \max_{e \in \mathcal{F}} \mathbb{E}[(Y^e - X^e\beta)^2] = \text{causal parameter}$$

A more realistic problem

Include hidden confounders (H)

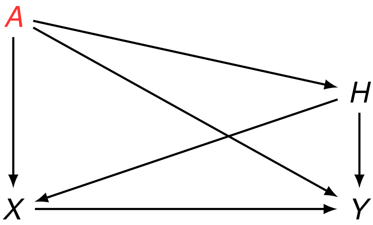


Are these reasonable assumptions?

Equivalent to Instrumental Variable Regression, where E are the IVs

Generalization: Anchor regression

Allow anchors A to influence all variables

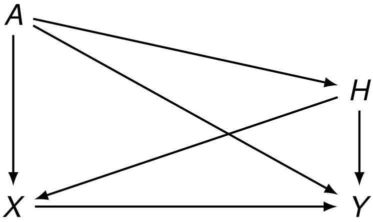


We cannot identify the causal parameter β anymore.

Price to pay for more realistic assumptions than the IV model.

Equivalent to IV Regression with invalid instruments A

Generalization: Anchor regression

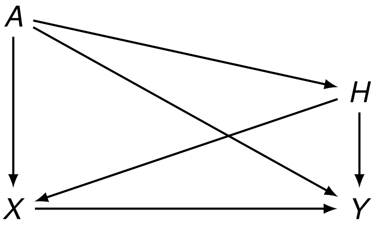


Aim: Induce stability of residuals across environments.

A loss along the lines of

$$L(\beta) = \frac{1}{2n}\left(\|(Y - X\beta)\|_2^2 + \lambda \left\| A^\top (Y - X\beta) \right\|_2^2\right)$$

Generalization: Anchor regression

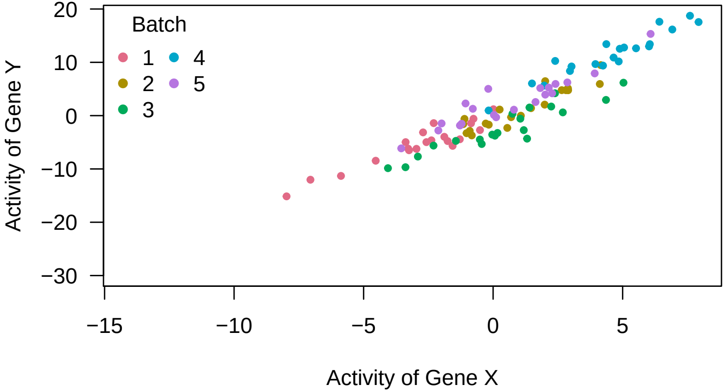


Causal regularization: Decorrelate residuals from anchors

$$L(\beta) = \frac{1}{2n}\left(\|(I - \Pi_A)(Y - X\beta)\|_2^2 + \gamma\,\|\Pi_A(Y - X\beta)\|_2^2\right)$$
$$\Pi_A = A(A^\top A)^{-1}A^\top$$

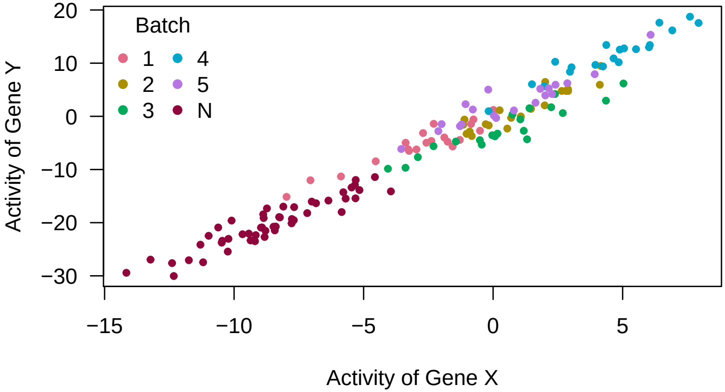
Example: Linear anchor regression

Heterogeneity due to batch-effects in biological experiments



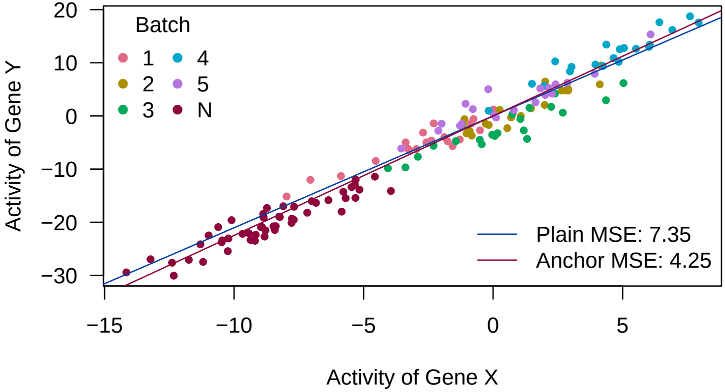
Example: Linear anchor regression

Predict new batch “N”, with (unseen) shift perturbations



Example: Linear anchor regression

Predict and evaluate models on new batch “N”



Linear anchor regression

$$L(\beta) = \frac{1}{2n} \| W_\gamma Y - W_\gamma X\beta \|_2^2, \;\; W_\gamma = \mathrm{I} - (1 - \sqrt{\gamma})\Pi_A$$

Simply compute OLS on $\tilde{Y} = W_\gamma Y$ and $\tilde{X} = W_\gamma X$!

Simulation: A case for anchor regression

Train

$$A \sim \text{Rademacher}$$

$$\varepsilon_H, \varepsilon_X, \varepsilon_Y \overset{\text{iid}}{\sim} \text{N}(0,1)$$

$$H \leftarrow \varepsilon_H$$

$$X \leftarrow A + H + \varepsilon_X$$

$$Y \leftarrow X + 2H + \varepsilon_Y$$

Perturbed

$$\varepsilon_H, \varepsilon_X, \varepsilon_Y \overset{\text{iid}}{\sim} \text{N}(0,1)$$

$$H \leftarrow \varepsilon_H$$

$$X \leftarrow {\color{red}1.8} + H + \varepsilon_X$$

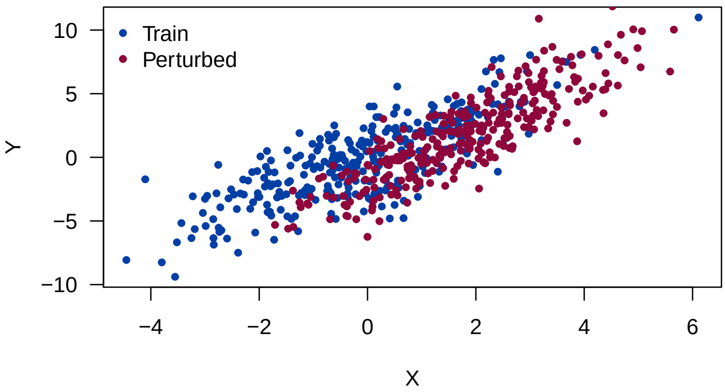
$$Y \leftarrow X + 2H + \varepsilon_Y$$

$$\begin{tikzcd}
A \arrow[d, "\text{shift}"] & & H \arrow[dll, "1"'] \arrow[d, "2"] \\
X \arrow[rr, "1"] & & Y
\end{tikzcd}$$

Example from Rotenhäusler (2018)

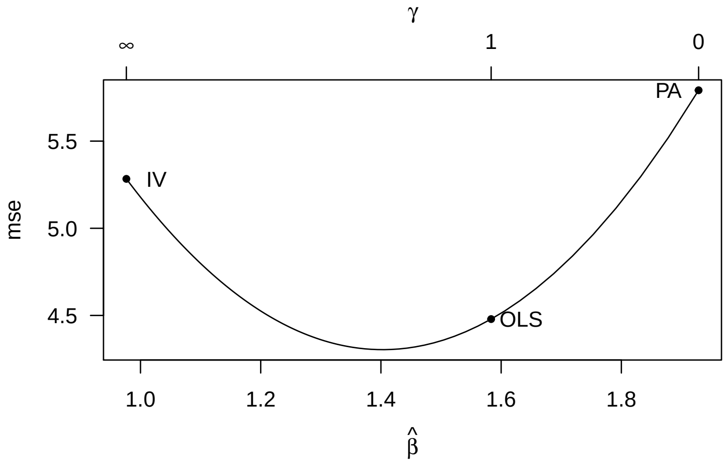
Simulation: A case for anchor regression

The IV assumptions hold . . .



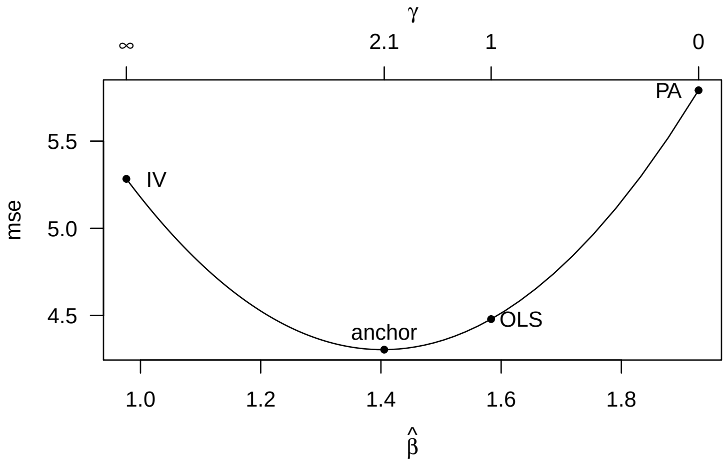
Simulation: A case for anchor regression

But OLS outperforms IV



Simulation: A case for anchor regression

OLS is still not optimal, but $\gamma = 2.1$ anchor regression is



Non-linear anchor regression

Anchor boosting or anchor neural networks

$$L(\beta) = \frac{1}{2n} \|W_\gamma(Y - f)\|_2^2, \; W_\gamma = \mathrm{I} - (1 - \sqrt{\gamma})\Pi_A$$

with complex conditional expectation function

$$f(x) = \mathbb{E}(Y|X = x)$$

Non-linear anchor regression

Anchor boosting or anchor neural networks

$$L(\beta) = \frac{1}{2n} \| W_\gamma (Y - f) \|_2^2, \; W_\gamma = \mathrm{I} - (1 - \sqrt{\gamma}) \Pi_A$$

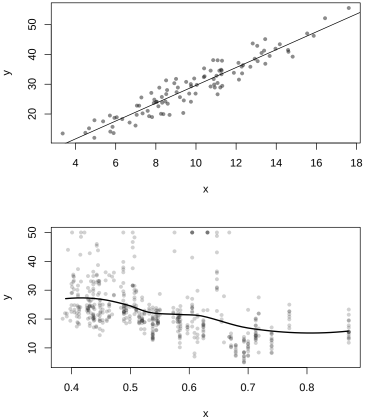
with complex conditional expectation function

$$f(x) = \mathbb{E}(Y | X = x)$$

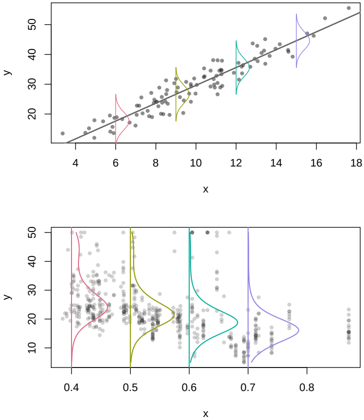
Up until now this was all "anchor curve fitting" ...

Distributional regression

Curve Fitting



Distributional regression



Distributional anchor regression

Aim: Derive a probabilistic anchor loss function

$$L(\beta) = -\text{log-likelihood} + \xi \cdot \text{causal regularizer}$$

Changing perspective

- MSE $\rightarrow$ (negative) log-likelihood
- Least squares residuals $\rightarrow$ score-based residuals
- Any kind of response (continuous, ordinal, survival)
- Allows for uninformative censoring

Score-based residuals

Score contribution for a newly introduced intercept $\alpha \equiv 0$

$$r_i = \partial_\alpha \ell(h, \alpha; y_i, \boldsymbol{x}_i)\big|_{\hat{h},\, \alpha=0}$$

for a model of the form

$$F_Y(Y|\boldsymbol{x}) = F_Z(h(y|\boldsymbol{x}) - \alpha)$$

Equivalent to a score test for testing H_0: $\alpha = 0$ for a covariate, that is not yet included in the model (Lagakos, 1980)

Distributional anchor regression

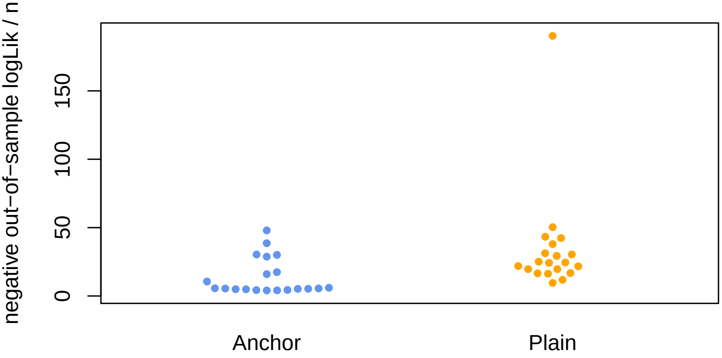
Probabilistic anchor loss function

$$L(h) = \underbrace{-\ell(h; y, \boldsymbol{x})}_{-\text{log-likelihood}} \quad + \quad \underbrace{\xi \, \|\Pi_A r\|_2^2}_{\text{causal regularizer}}$$

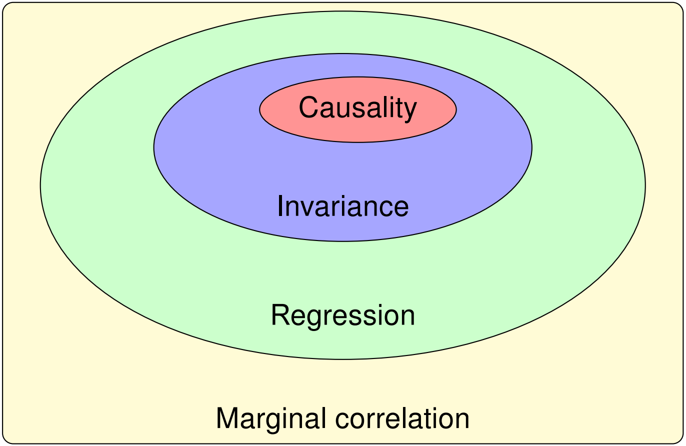
$$r = \partial_\alpha \ell(h, \alpha; y, \boldsymbol{x})\big|_{\hat{h},\, \alpha=0}$$

Simulation: Distributional anchor regression

$$F_Y(y|\boldsymbol{x}) = \Phi\left(\vartheta_1 + \vartheta_2 y - \boldsymbol{x}^\top \boldsymbol{\beta}\right)$$



Taking a step back



Future work

- Implement distributional anchor regression in {`anchor`}
- Combine distributional anchor regression with DNNs
- Apply distributional anchor regression to real-world data
- Theoretical properties of the probabilistic anchor loss
- Estimate anchor variables from data

Acknowledgements

Beate Sick

Torsten Hothorn

Susanne Wegener

Helmut Grabner

Lisa Herzog

References

Bühlmann, Peter. "Invariance, causality and robustness." *arXiv preprint arXiv:1812.08233* (2018).

Haavelmo, Trygve. "The statistical implications of a system of simultaneous equations." *Econometrica, Journal of the Econometric Society* (1943): 1-12.

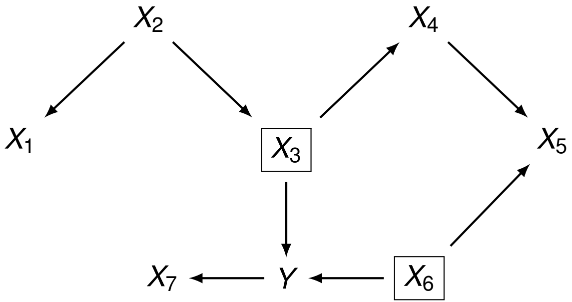
Lagakos, S. W. "The graphical evaluation of explanatory variables in proportional hazard regression models." *Biometrika* 68.1 (1981): 93-98.

Peters, Jonas, Peter Bühlmann, and Nicolai Meinshausen. "Causal inference by using invariant prediction: identification and confidence intervals." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78.5 (2016): 947-1012.

Rothenhäusler, Dominik, et al. "Anchor regression: heterogeneous data meets causality." *arXiv preprint arXiv:1801.06229* (2018).

Appendix

(Another) connection to causality



$$\underset{\beta}{\arg\min} \max_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[(Y - X\beta)^2] = \text{causal parameter}$$

if $\mathcal{P}$ contains *all* possible interventional distributions $\mathbb{P}$ on components of X.

In other words, conditioning on pa(Y) shields against arbitrarily strong interventions on X.

Motivation for the anchor estimator

Prerequisites

Start from the linear SEM

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

corresponding to the anchor regression problem.

The anchor estimator is given by

$$\hat{\beta}(\gamma) = \arg\min_{\beta} \frac{1}{2n} \left\{ \|(I - \Pi_A)(Y - X\beta)\|_2^2 + \gamma \, \|\Pi_A(Y - X\beta)\|_2^2 \right\}.$$

Worst case risk optimization

$\hat{\beta}(\gamma)$ solves a worst case optimization problem over a class of shift perturbations C_γ. The linear SEM for the perturbed set is

$$\begin{pmatrix} X^v \\ Y^v \\ H^v \end{pmatrix} = B \begin{pmatrix} X^v \\ Y^v \\ H^v \end{pmatrix} + \varepsilon + v = (\mathrm{I} - B)^{-1}(\varepsilon + v),$$

where $v \in \mathrm{span}(M)$.

The class of shift perturbations C_γ is now defined as

$$C_\gamma := \left\{ v : v = M\delta \text{ for some } \delta \text{ s.t. } \mathrm{Corr}(\delta, \varepsilon) = 0 \text{ and } \mathbb{E}(\delta^\top \delta) \preccurlyeq \gamma \mathbb{E}(A^\top A) \right\},$$

which allows to formulate the population version of the worst case risk

$$\sup_{v \in C_\gamma} \mathbb{E}[(Y^v - X^v b)^2] = \mathbb{E}[((I - P_A)(Y - Xb))^2] + \gamma \mathbb{E}[(P_A(Y - Xb))^2].$$

Theorem 1 in Rothenhäusler (2018).

Distributional robustness

Assume X and Y have mean zero, then $\mathbb{E}[A(Y - Xb)] = \text{Cov}(A, Y - Xb)$. Let

$$I := \{ b \in \mathbb{R}^p : \mathbb{E}[A(Y - Xb)] = 0 \},$$

then

$$\beta \in I \Leftrightarrow Y^v - X^v \beta \text{ has the same distribution } \forall\ v \in \text{span}(M),$$

which shows the duality between the worst case optimization over a class of shift perturbations and a distributional robustness over the same class.

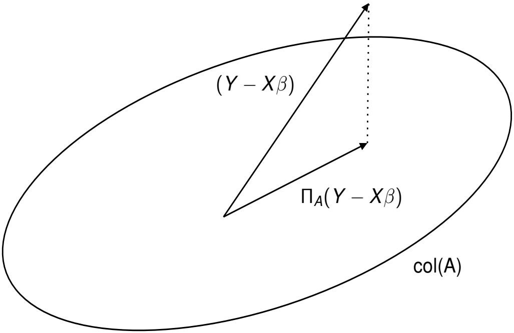
Thus we established

$$\beta(\gamma) = \arg\min_b \sup_{v \in C_\gamma} \mathbb{E}[(Y^v - X^v b)^2].$$

Theorem 3 in Rothenhäusler (2018).

Geometric interpretation

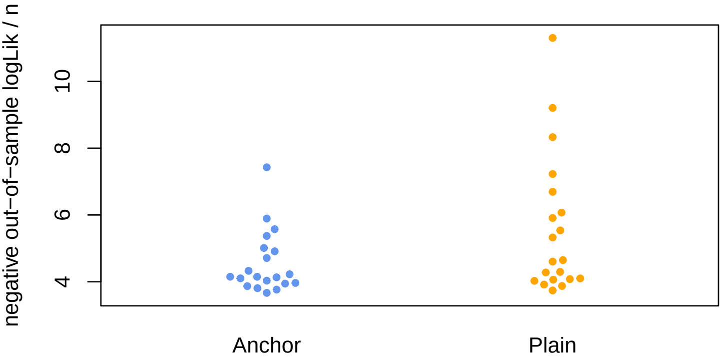
The causal regularization term $\gamma \, \|\Pi_A(Y - X\beta)\|_2^2$ encourages orthogonality (uncorrelatedness) between the anchor variables A and the residuals $Y - X\beta$ for larger γ.



More empirical results

Simulation: Box-Cox anchor regression

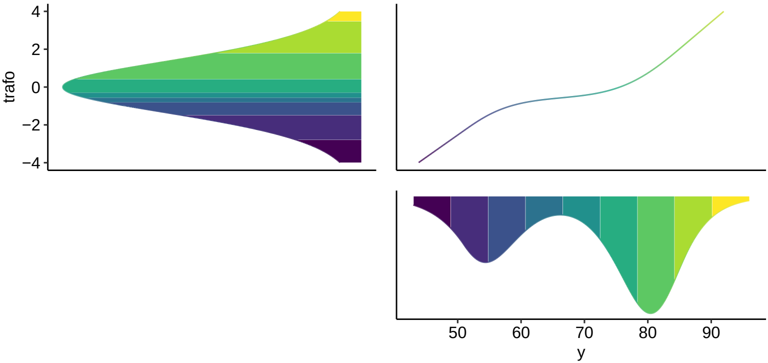
$$F_Y(y|\boldsymbol{x}) = \Phi\left(\boldsymbol{a}_{\text{Bs},6}(y)^\top \boldsymbol{\vartheta} - \boldsymbol{x}^\top \beta\right)$$



Transformation models

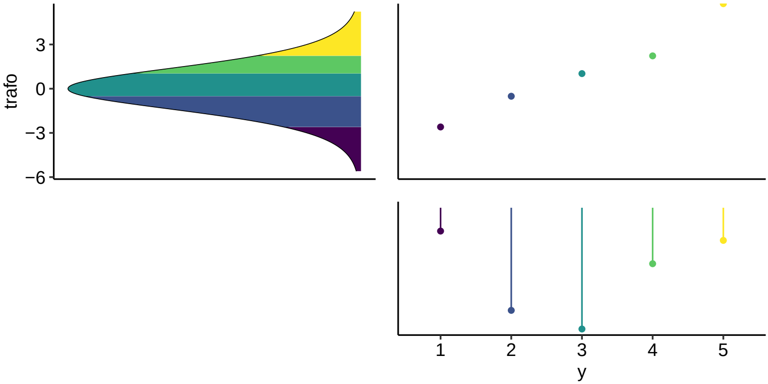
Transformation models

$$F_Y(y|\boldsymbol{x}) = F_Z(h(y|\boldsymbol{x})), \; y \in \mathbb{R}$$



Transformation models

$$F_Y(y|\boldsymbol{x}) = F_Z(h(y|\boldsymbol{x})),\ y \in \{y_1 < \cdots < y_K\}$$



Simulation schemes

Simulation: Linear anchor regression

$X \in \mathbb{R}^{10},\ A \in \mathbb{R}^2,\ H \in \mathbb{R}$

$A \sim \mathrm{N}_2(0, \mathrm{I}),\ H \sim \mathrm{N}(0, 1)$

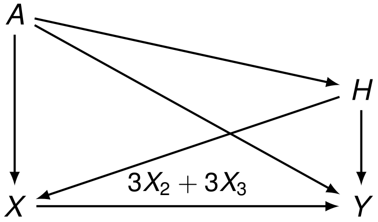
$Y \leftarrow 3X_2 + 3X_3 + H - 2A_1 + \varepsilon_Y$

$X \leftarrow A_1\eta_1 + A_2\eta_2 + H + \varepsilon_{X_j}$

$\gamma_1, \gamma_2, \varepsilon_{X_j}, \varepsilon_Y \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, 1)$

$n_{\text{train}} = 300,\ n_{\text{test}} = 2000$

Shift perturbation: $\sqrt{10}A_{\text{test}}$



Simulation: Non-linear anchor regression

$X \in \mathbb{R}^{10},\ A \in \mathbb{R}^2,\ H \in \mathbb{R}$

$A \sim \mathrm{N}_2(0, \mathrm{I}),\ H \sim \mathrm{N}(0, 1)$

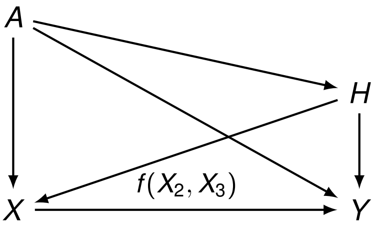
$Y \leftarrow f(X_2, X_3) + 3H - 2A_1 + \varepsilon_Y$

$X \leftarrow A_1 + A_2 + 2H + \varepsilon_{X_j}$

$\varepsilon_{X_j} \sim \mathrm{N}(0, 0.5^2)$

$\varepsilon_Y \sim \mathrm{N}(0, 0.25^2)$

$n_{\text{train}} = 300,\ n_{\text{test}} = 2000$



Shift perturbation: $A_{\text{test}} \sim \mathrm{N}_{n_{\text{test}}}(\mu, \mathrm{I}),\ \mu \sim \mathrm{N}_{n_{\text{test}}}(1, 2^2\mathrm{I})$

$f(X_2, X_3) = X_2 + X_3 + I(X_2 \le 0) + I(X_2 \le -0.5)I(X_3 \le 1)$