Invariance, Causality and Robustness

2018 Neyman Lecture *

Peter Bühlmann †
Seminar for Statistics, ETH Zürich

December 21, 2018

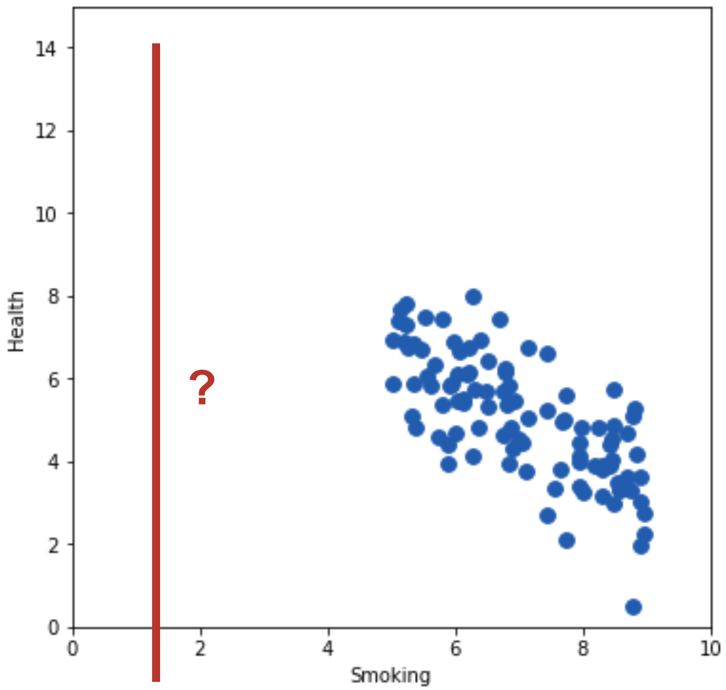
Advanced Topics in Machine Learning and Data Science (2022)
Nora Schneider
27.03.2022

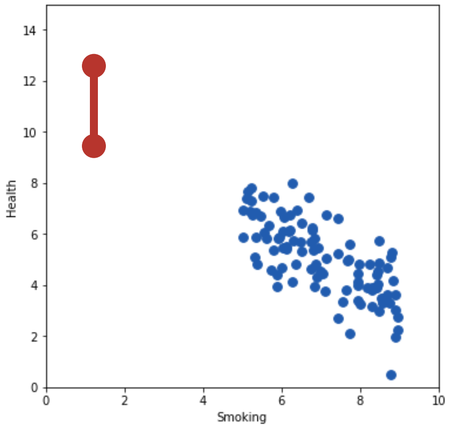
Agenda

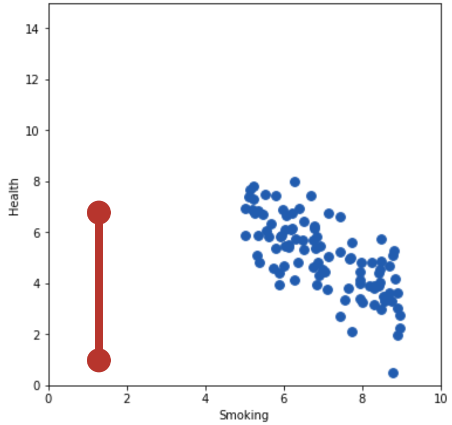
1. Introduction Causality
2. Problem Setting
3. Invariant Causal Prediction
4. Instrumental Variable Regression
5. Anchor Regression
6. Conclusion
7. Q&A and Discussion

Causality: *“What if I do .... (in a heterogenous setting)?"*

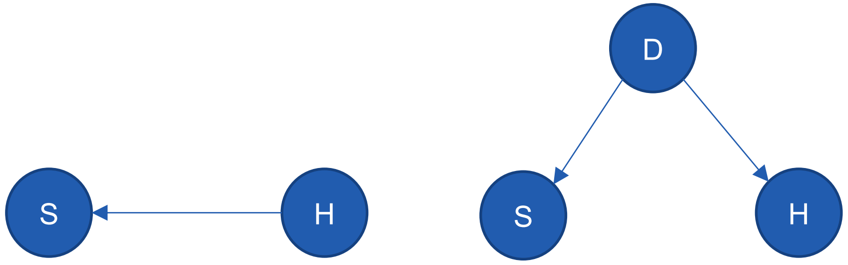
- Gold standard: Randomized Control Trials



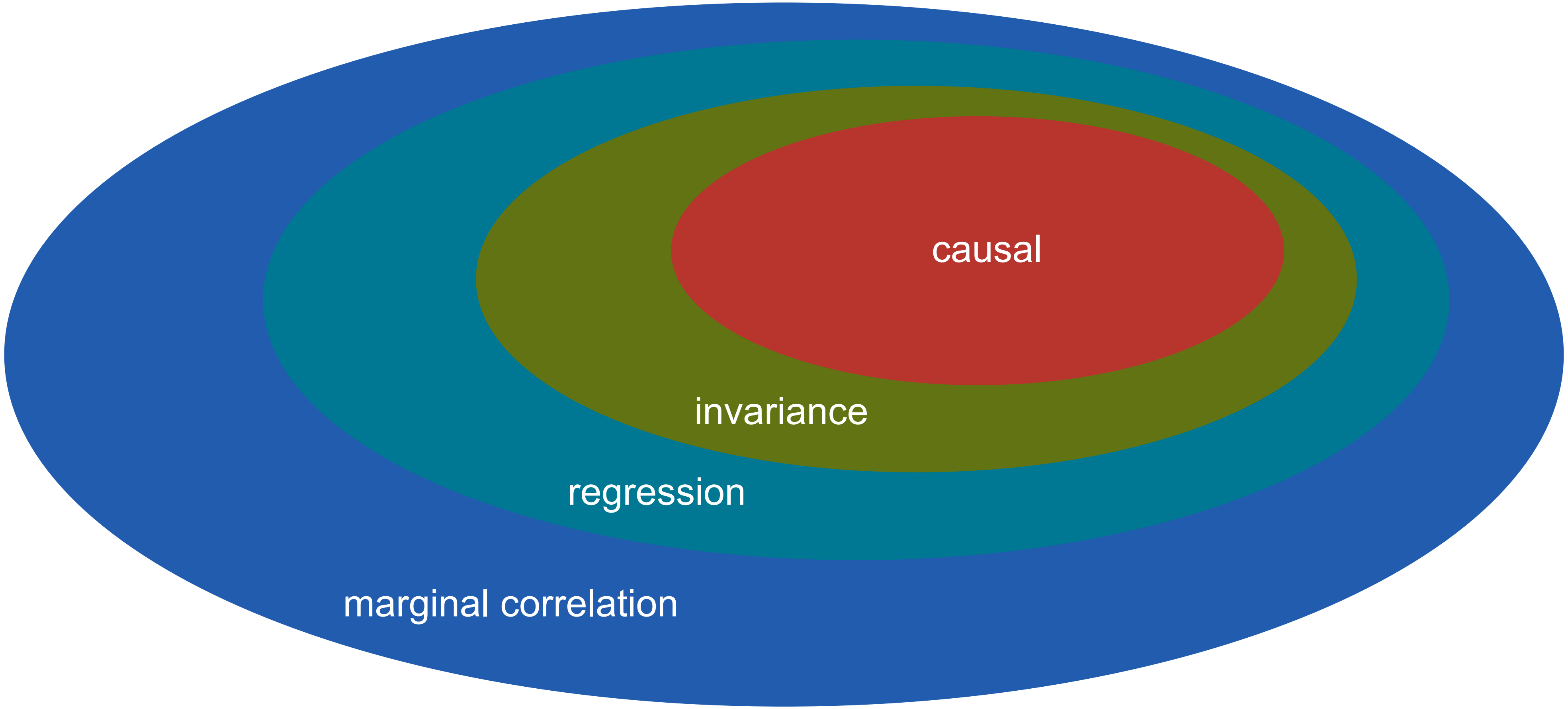




S = Smoking
H = Health
D = Depression level



Associations Between Covariates X and a Response Y



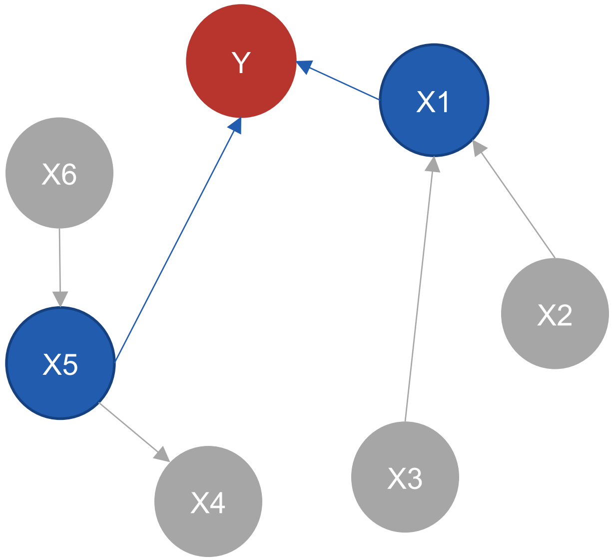
The Problem Setting

Structured Equation Models (SEMs)

$$Y \leftarrow f_Y(X_{\mathrm{pa}(Y)}, \varepsilon_Y),\ \ \varepsilon_Y \text{ independent of } X_{\mathrm{pa}(Y)},$$
$$X \sim F_X,$$

special case:

$$Y \leftarrow f_Y(X_{\mathrm{pa}(Y)}, \varepsilon_Y),$$
$$X_j \leftarrow f_j(X_{\mathrm{pa}(X_j)}, \varepsilon_j),$$



direct causal variables for Y: $S_{causal} = pa(Y) = \{X1, X5\}$

The Problem Setting

Exploit heterogeneities in the data and inspect a certain stability

- Observe data from different *environments* $(\boldsymbol{X^e}, \boldsymbol{Y^e}) \in \mathcal{E}$
- Non-observed environments: $\mathcal{F} \supset \mathcal{E}$

- <u>**ad-hoc conditions $\boldsymbol{B}(\mathcal{E})$**</u>

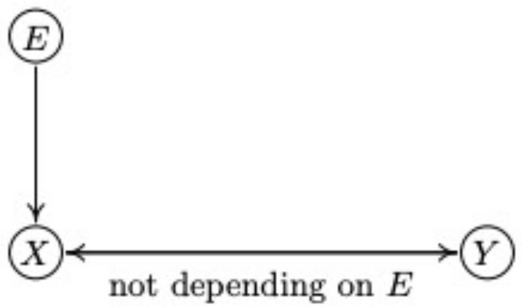
Structural equation model remains the same, that is for all $e \in \mathcal{E}$

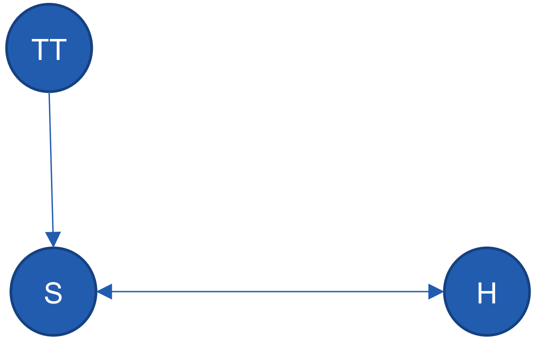
$$Y^e \leftarrow f_Y(X^e_{pa(Y)}, \varepsilon^e_Y)$$

where ε^e_Y is independent of $X^e_{pa(Y)}$ and ε^e_Y has the same distribution as ε_Y.

- <u>**ad-hoc aim:**</u> ideally, e should change the distribution of

- ad-hoc conditions $B(\mathcal{F})$: analogous





TT = Tobacco Taxes, S = Smoking, H = Health

The Problem Setting

Worst case risk optimization and predictive robustness

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

Linear model setting: $\mathrm{argmin}_b \max_{e \in \mathcal{F}} \mathbb{E}[|Y^e - X^e b|^2].$

- Assuming that $B(\mathcal{F})$ holds, then

$$\mathrm{argmin}_b \max_{e \in \mathcal{F}} \mathbb{E}[|Y^e - X^e b|^2] = \text{causal parameter}$$

Causal parameters optimize worst case loss w.r.t. unseen future scenarios/ environments.

The Problem Setting
Invariance Assumption

- $A_S(\mathcal{E})$: The subset S of covariates fulfills invariance saying that

$$\mathcal{L}(Y^e \mid X_S^e) \text{ is the same (= invariant) across all } e \in \mathcal{E}$$

- $A_S(\mathcal{F})$: analogous
- Linear model setting:

Subset S* and regression coefficients β^* with $supp(\beta^*) = \{j; \beta_j^* \neq 0\} = S^*$ such that

$$\text{For all } e \in \mathcal{E}\text{:} \quad Y^e = X^e\beta^* + \varepsilon^e, \quad \text{and } \varepsilon^e \text{ independent of } X_{S^*}^e,\ \varepsilon^e \sim F_\varepsilon$$

The Problem Setting
Invariance Assumption

- $A_S(\mathcal{E})$: The subset S of covariates fulfills invariance saying that

$$\mathcal{L}(Y^e |\, X_S^e) \text{ is the same (= invariant) across all } e \in \mathcal{E}$$

- $A_S(\mathcal{F})$: analogous
- Linear model setting:

Subset S* and regression coefficients β^* with $supp(\beta^*) = \{j; \beta_j^* \neq 0\} = S^*$ such that

$$\text{For all } e \in \mathcal{E}: \quad Y^e = X^e\beta^* + \varepsilon^e, \quad \text{and } \varepsilon^e \text{ independent of } X_{S^*}^e,\ \varepsilon^e \sim F_\varepsilon$$

- **Proposition 1:** Assume a partial structural equation model. Consider the set of environments $\mathcal{F}$ such that $B(\mathcal{F})$ holds. Then, the set of causal variables $S_{causal} = pa(Y)$ satisfies the invariance assumption with respect to $\mathcal{F}$, that is $A_{S^*}(\mathcal{F})$ holds:

$$\text{causal variables} \implies \text{Invariance.}$$

$$\text{causal structures} \overset{?}{\impliedby} \text{Invariance}$$

Invariant Causal Prediction – ICP (Peters et al., 2016)

Procedure

for all subsets
$S_k \in S$

test if S_k fulfills
invariance

$H_{0,S_k}(\mathcal{E})$: assumption $A_{S_k}(\mathcal{E})$ holds

$$\hat{S}(\mathcal{E}) = \bigcap_{S_k} \{S_k; H_0 \text{ not rejected at } \alpha\}$$

Theorem 1: Assume a structural equation model for response Y and that the environments/ perturbations in $\mathcal{E}$ *satisfy (B)*. Furthermore, assume that the tests are valid, controlling the type 1 error. Then, for alpha in 0,1 we have that

$$\mathbb{P}[\hat{S}(\mathcal{E}) \subseteq \mathrm{pa}(Y)] \geq 1 - \alpha.$$

- No information about the power/ completeness of estimates:
  - Roughly: power increases as $\mathcal{E}$ becomes larger

Invariant Causal Prediction – ICP (Peters et al., 2016)

Application: Single gene knock-out experiments in yeast

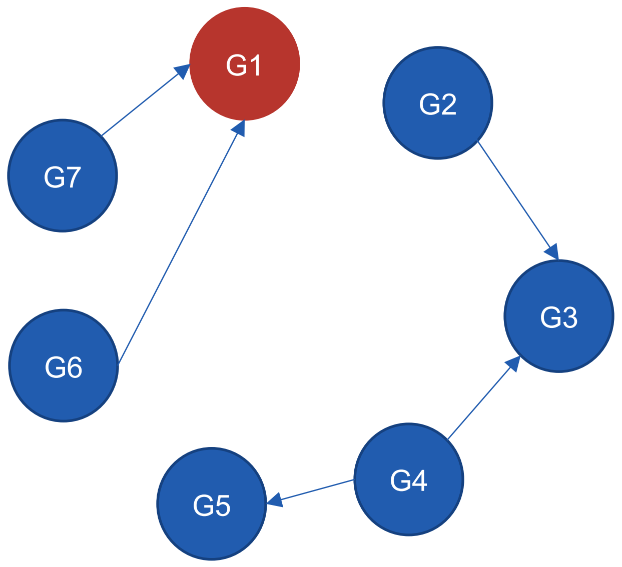
- mRNA expression levels for 6,170 genes

e = 1 — 160 wild-type observations (“normal”)

e = 2 — 1,479 interventional observations (“single gene deletion”)

Goal: predict the expression levels of all (except the deleted) genes of a new and unseen single gene deletion intervention

no intervention



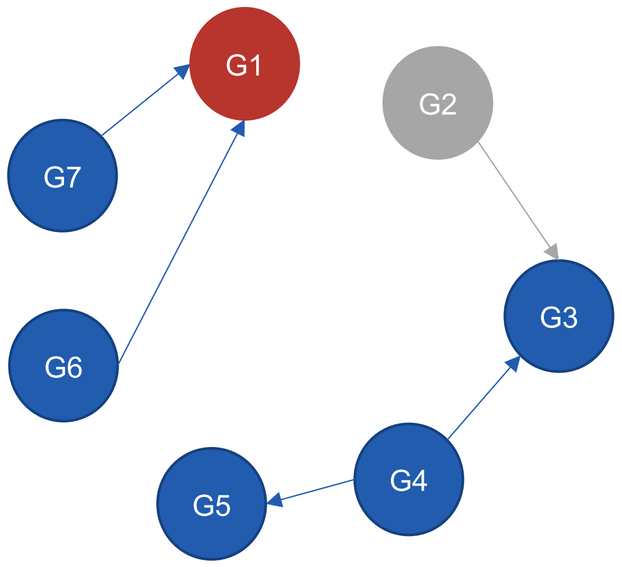
Invariant Causal Prediction – ICP (Peters et al., 2016)

Application: Single gene knock-out experiments in yeast

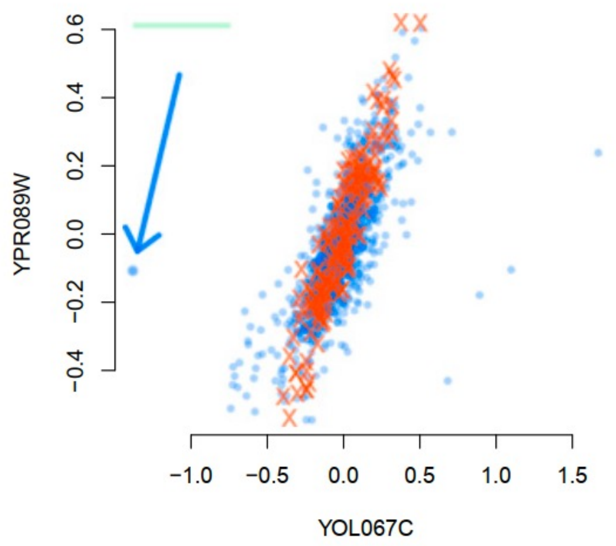
- mRNA expression levels for 6,170 genes
  - e = 1: 160 wild-type observations ("normal")
  - e = 2: 1,479 interventional observations ("single gene deletion")

Goal: predict the expression levels of all (except the deleted) genes of a new and unseen single gene deletion intervention

knock-out of G2



No Strong Intervention Effect



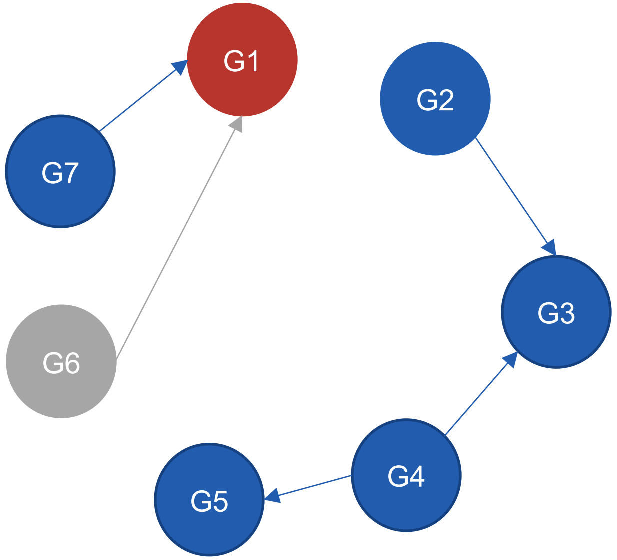
Invariant Causal Prediction – ICP (Peters et al., 2016)

Application: Single gene knock-out experiments in yeast

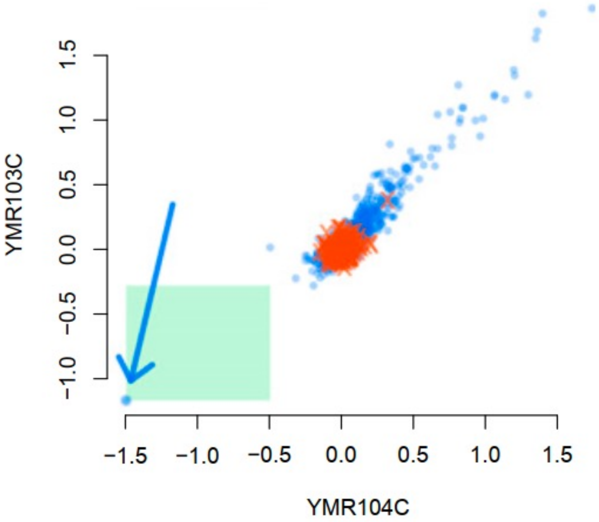
- mRNA expression levels for 6,170 genes
  - e = 1: 160 wild-type observations (“normal”)
  - e = 2: 1,479 interventional observations (“single gene deletion”)

<u>**Goal**:</u> predict the expression levels of all (except the deleted) genes of a new and unseen single gene deletion intervention

knock-out of G6

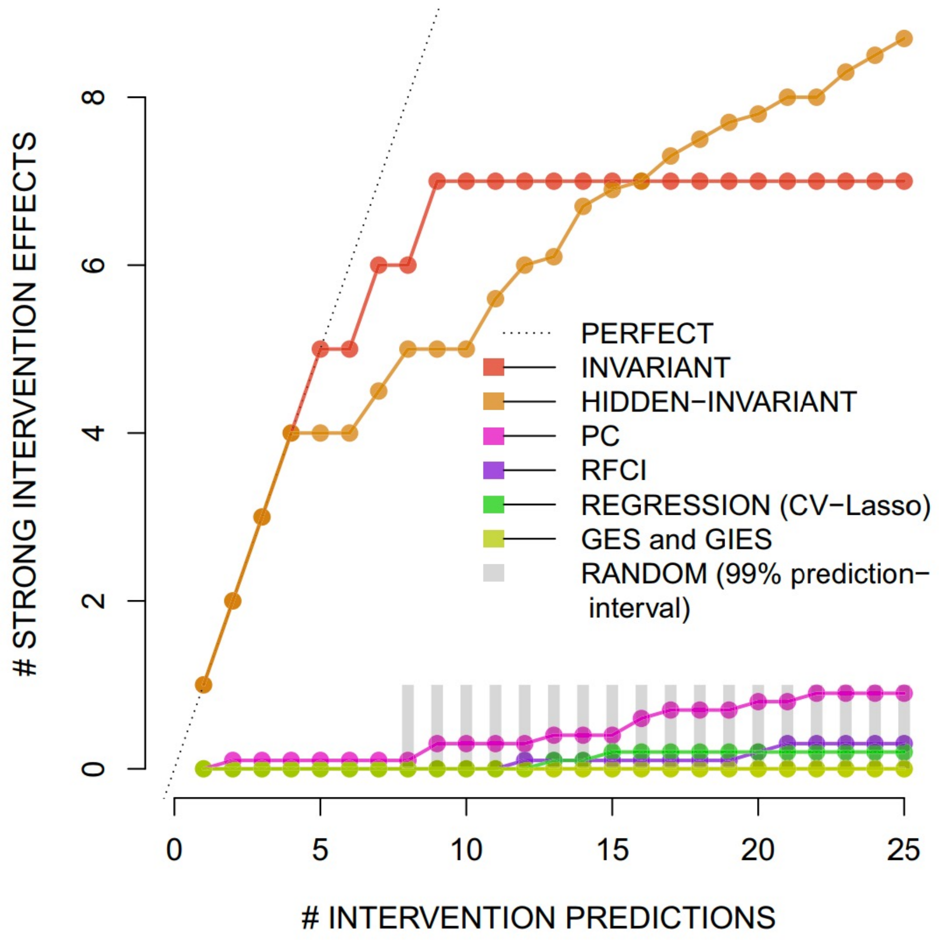


Strong Intervention Effect



Invariant Causal Prediction – ICP (Peters et al., 2016)

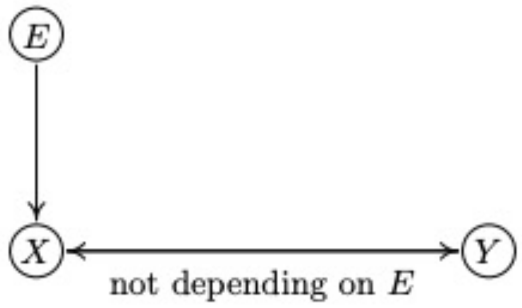
Application: Single gene knock-out experiments in yeast



k most often selected edges (x-axis), how many of them correspond to a true SIE based on test data (y-axis)?

More realistic setting – Relaxing conditions

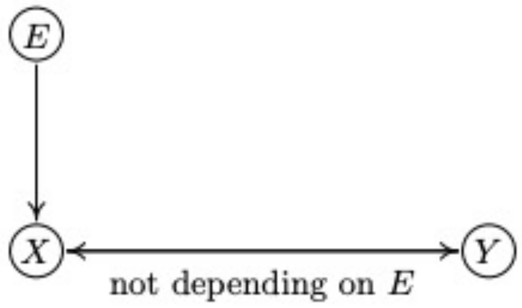
ICP Model



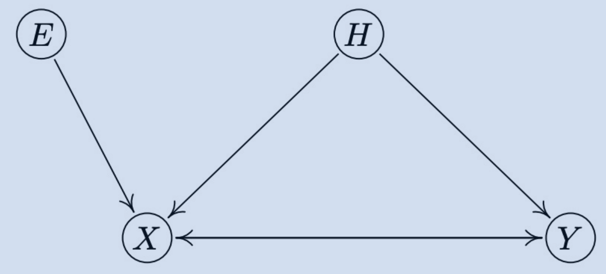
- Approximate instead of exact invariance holds
- Residuals not invariant for all environments
- Different regression parameters for varying environments
- Hidden confounding factors
- ...

Instrumental Variable Regression

ICP Model

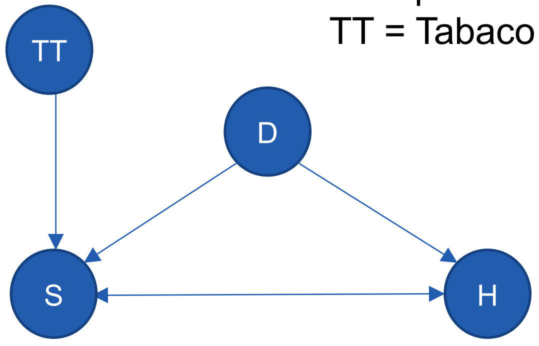


IV Regression Model



$$Y \leftarrow f_Y(X_{\mathrm{pa}_X(Y)}, H, \varepsilon_Y),$$
$$X_j \leftarrow f_j(X_{\mathrm{pa}_X(X_j)}, H, E, \varepsilon_j),$$

S = Smoking
H = Health
D = Depression level
TT = Tabaco Taxes



(Assume a linear setting)

- Two Stage Least Squares (2SLS) estimation:

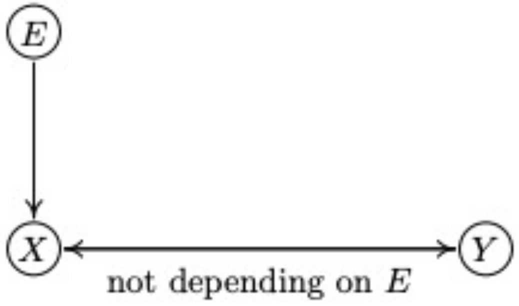
1. Regress each column of X on instruments (E) to obtain $\hat{X}$ by OLS
2. Regress Y on the predicted values from stage 1 $\hat{X}$

➢ **Can identify causal mechanism between X and Y.**

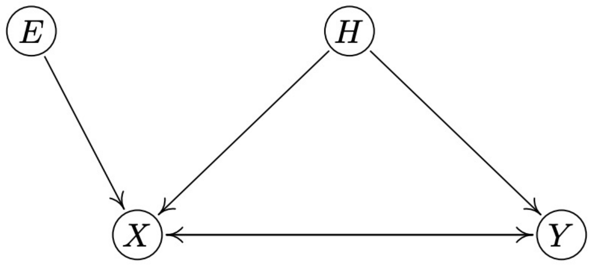
Anchor Regression (Rothenhäusler et al. 2018)

Model ("invalid instruments")

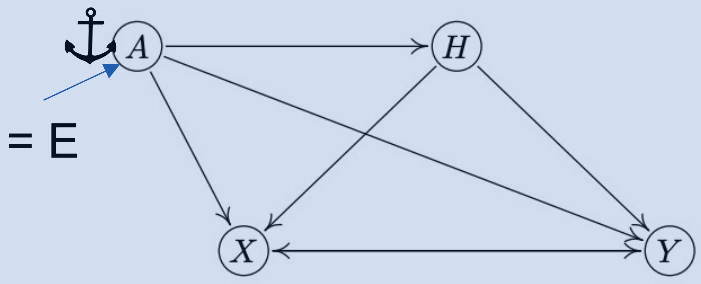
ICP Model



IV Regression Model



Anchor Regression Model



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

with $(I - B)$ being invertible and allowing for feedback loops

$$Y = X^T \beta + H^T \alpha + A^T \xi + \varepsilon_Y,$$

- **Fundamental identifiability problem** (cannot identify causal mechanism between X and Y)

Anchor Regression (Rothenhäusler et al. 2018)

... but with causal regularization we can still infer interesting properties

Motivation: invariance for residuals

Proposition 2: We can show that in the Anchor model

A uncorrelated with $(Y - X^T b)$ $\Leftrightarrow$ $(Y - X^T b)$ is “shift-invariant"

! Remember: causal parameters would lead to general invariance!

Anchor Regression (Rothenhäusler et al. 2018)

... but with causal regularization we can still infer interesting properties

Motivation: invariance for residuals

Proposition 2: We can show that in the Anchor model

A uncorrelated with $(Y - X^T b)$ $\Leftrightarrow$ $(Y - X^T b)$ is “shift-invariant"

! Remember: causal parameters would lead to general invariance!

Estimator: $\hat{\beta}(\gamma) = \text{argmin}_b \left(\|(I - \Pi_{\mathbf{A}})(\mathbf{Y} - \mathbf{X}b)\|_2^2/n + \gamma \|\Pi_{\mathbf{A}}(\mathbf{Y} - \mathbf{X}b)\|_2^2/n \right)$

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

- For $\gamma = 1$: OLS
- For $\gamma = 0$: Adjusting for heterogenity due to A
- For $\gamma = \infty$: Two-stage least square in IV model
- **For $0 \leq \gamma < \infty$: causal regularization**

Anchor Regression (Rothenhäusler et al. 2018)

... but with causal regularization we can still infer interesting properties

Motivation: invariance for residuals

Proposition 2: We can show that in the Anchor model

A uncorrelated with $(Y - X^T b)$ $\Leftrightarrow$ $(Y - X^T b)$ is "shift-invariant"

! Remember: causal parameters would lead to general invariance!

Estimator: $$\hat{\beta}(\gamma) = \text{argmin}_b \left(\|(I - \Pi_{\mathbf{A}})(\mathbf{Y} - \mathbf{X}b)\|_2^2/n + \gamma \|\Pi_{\mathbf{A}}(\mathbf{Y} - \mathbf{X}b)\|_2^2/n \right)$$

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

- For $\gamma = 1$: OLS
- For $\gamma = 0$: Adjusting for heterogenity due to A
- For $\gamma = \infty$: Two-stage least square in IV model
- **For $0 \leq \gamma < \infty$: causal regularization**

- Trivial computation by *linear transformation* of the data + OLS estimation

$$\tilde{Y} = W_\gamma Y, \;\; \tilde{X} = W_\gamma X,$$
$$W_\gamma = I - (1 - \sqrt{\gamma})\Pi_{\mathbf{A}}.$$

Anchor Regression (Rothenhäusler et al. 2018)

... but with causal regularization we can still infer interesting properties

With causal regularization we can minimize the worst case risk over a certain class of shift perturbations, meaning

$$\arg\min_{b} \max_{e \in \mathcal{F}} \mathbb{E}|Y^e - (X^e)^T b|^2$$

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

! Remember: causal parameters minimize worst case risk for “essentially all” perturbations!

Anchor Regression (Rothenhäusler et al. 2018)

Class of Shift Perturbations (Environments) $\mathcal{F}$

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

$$\begin{aligned} C_\gamma = \{v;\ \ & v = M\delta \text{ for random or deterministic } \delta, \text{ uncorrelated with } \varepsilon \\ & \text{and } \mathbb{E}[\delta\delta^T] \preceq \gamma\mathbb{E}[AA^T]\}. \end{aligned}$$

- Shift vector $v \in span(M)$ with "strength" $\|v\|^2 = O(\gamma)$
  - $\gamma = 1$: v is up to the order MA = heterogeneity in the (observed) data
  - $\gamma \gg 1$: v can be a stronger perturbation being an amplification of the observed heterogeneity MA

Anchor Regression (Rothenhäusler et al. 2018)

Worst case risk minimization

With causal regularization we can minimize the worst case risk over a certain class of shift perturbations, meaning

$$\arg\min_{b} \max_{e \in \mathcal{F}} \mathbb{E}|Y^e - (X^e)^T b|^2$$

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

<u>**Theorem 2:**</u> For any $b \in \mathbb{R}^p$

$$\underbrace{\mathbb{E}_{\text{train}}[((\mathrm{Id} - \mathrm{P}_A)(Y - X^\intercal b))^2] + \gamma \mathbb{E}_{\text{train}}[(\mathrm{P}_A(Y - X^\intercal b))^2]}_{\text{causal regularized risk}} = \underbrace{\sup_{v \in C^\gamma} \mathbb{E}_v[(Y - X^\intercal b)^2]}_{\substack{\text{worst case risk} \\ \text{(shift perturbations)}}},$$

Anchor Regression (Rothenhäusler et al. 2018)

Worst case risk minimization & diluted form of causality

Therefore

$$\hat{\beta}(\gamma) = \text{argmin}_b \left(\|(I - \Pi_{\mathbf{A}})(\mathbf{Y} - \mathbf{X}b)\|_2^2/n + \gamma \|\Pi_{\mathbf{A}}(\mathbf{Y} - \mathbf{X}b)\|_2^2/n \right)$$

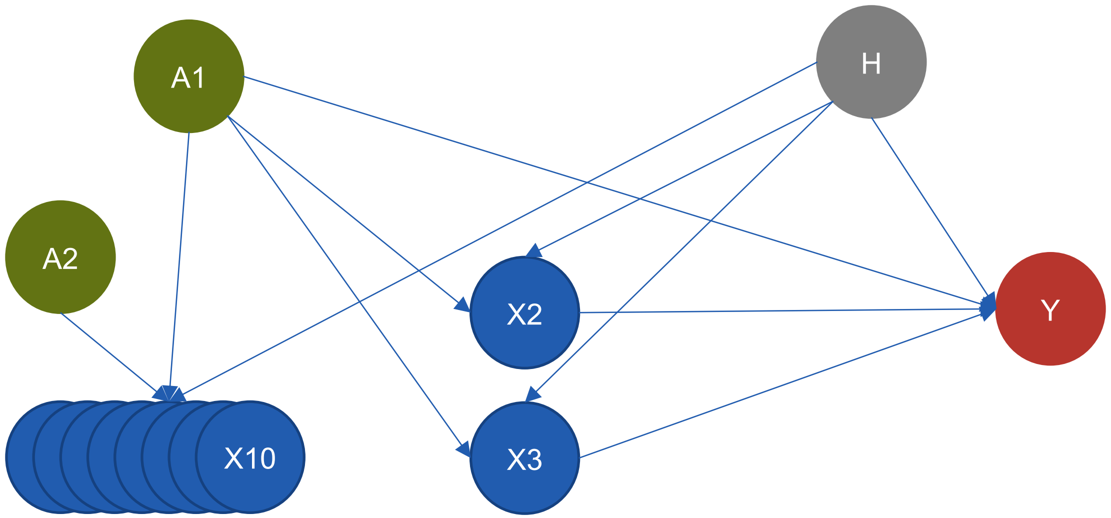
protects against worst case shift perturbations scenarios and leads to **prediction robustness.**

- Variables corresponding to large entries in $\hat{\beta}(\gamma)$ are “key drivers” for explaining Y (in a stable way).
- For $\gamma \to \infty$, define $\text{supp}(\beta(\gamma \to \infty))$ as the **variables which are diluted causal** for Y.

- Note, if IV assumptions hold, we can identify ”normal” causal variables using Anchor Regression too.

Anchor Regression (Rothenhäusler et al. 2018)

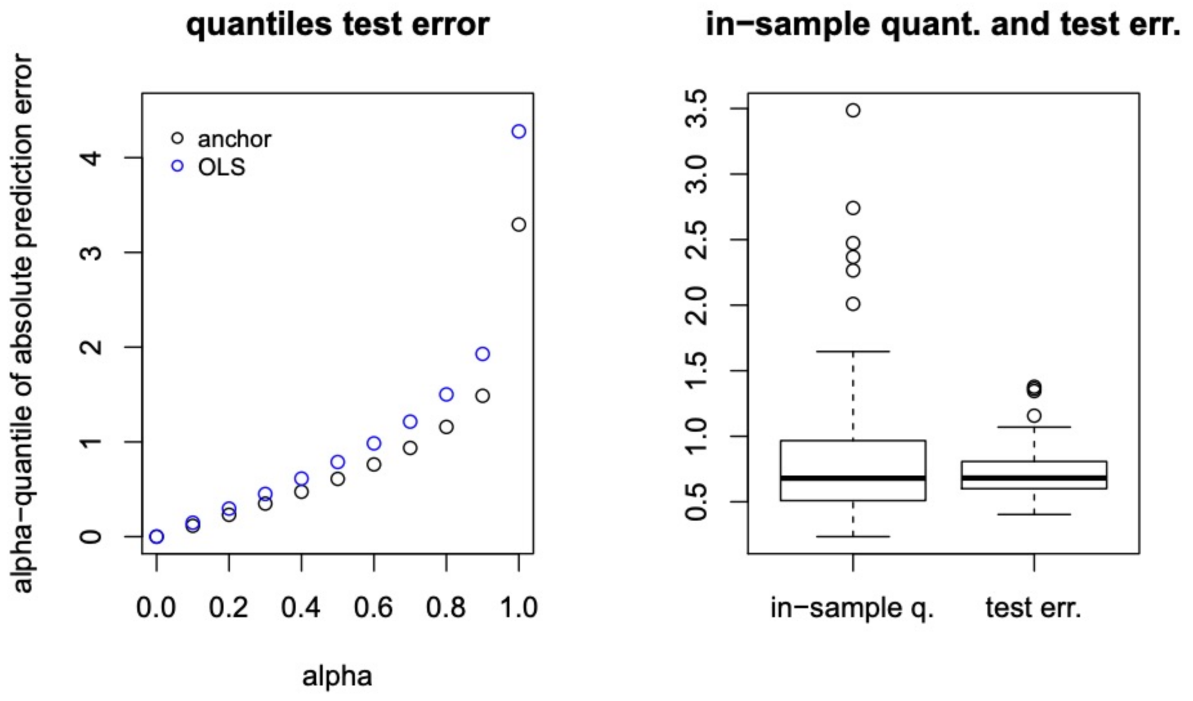
Application – simulation study



- Training data: n = 200
- Test data: n = 2,000 and perturbation by multiplying A1 & A1 with factor $\sqrt{10}$

Anchor Regression (Rothenhäusler et al. 2018)

Application – simulation study

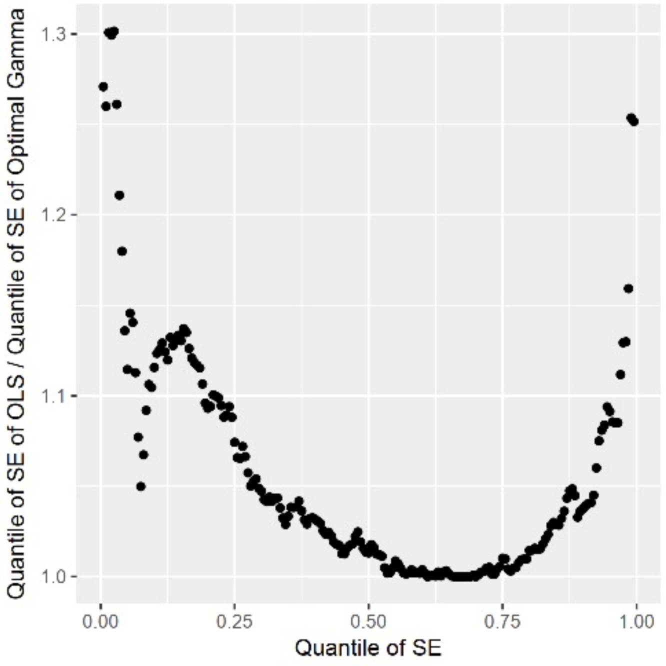


- Anchor regression exhibits **better prediction performance** under (out-sample) perturbation than OLS.
- If out-sample data similar to train data (**no new perturbations**), then there would be **no gain** (even a slight loss) compared to OLS.

Anchor Regression (Rothenhäusler et al. 2018)

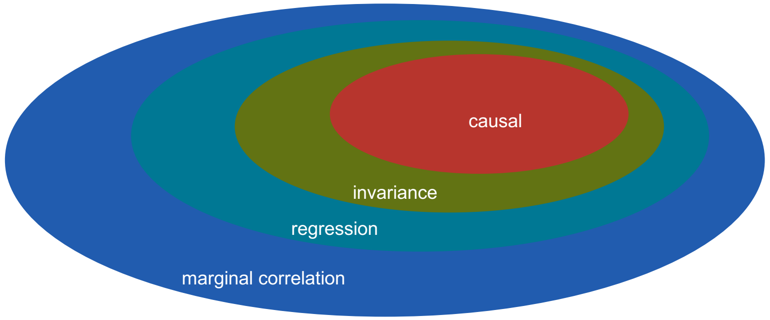
Application – Bike sharing data set (real dataset) with strong heterogeneities

- Predict bike rental count based on d = 4 covariates (weather data) and a sample with n = 17,379
- Discrete anchor variable = “time” (one level per day)

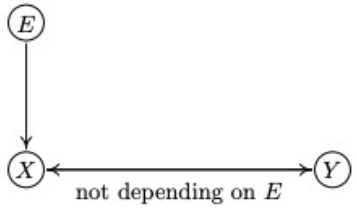


$\approx 15-20\%$ performance gain compared to OLS

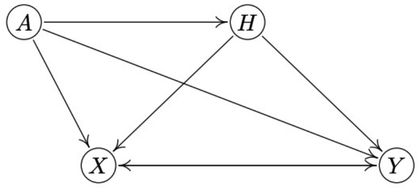
Conclusion



- **Causality**, as predictive **robustness** in **heterogeneous** setting
  by **exploiting invariance** from heterogeneous data (important to infer invariance)



- Invariant Causal Prediction
- Invariance corresponds to causality (= worst case risk optimization)



- IV model assumptions: identify causal relationship (2SLS, Anchor regression)
- Relaxing assumptions: limiting to shift perturbations we can infer invariance of residuals and "diluted causality" (= worst case risk optimization)

Even when inferring causal effects are non-identifiable, identifying variables that fulfill invariance can provide more meaningful insights than methods like regression.

Backup

Referenced Papers

Invariance, Causality and Robustness

2018 Neyman Lecture *

Peter Bühlmann †
Seminar for Statistics, ETH Zürich

December 21, 2018

Methods for causal inference from gene perturbation experiments and validation

Nicolai Meinshausen[a], Alain Hauser[b], Joris M. Mooij[c], Jonas Peters[d], Philip Versteeg[c], and Peter Bühlmann[a,1]

[a]Seminar for Statistics, Eidgenössische Technische Hochschule (ETH) Zurich, CH-8092 Zurich, Switzerland; [b]Department of Engineering and Information Technology, Bern University of Applied Sciences, CH-3400 Burgdorf, Switzerland; [c]Informatics Institute, University of Amsterdam, 1090 GH Amsterdam, The Netherlands; and [d]Max Planck Institute for Intelligent Systems, D-72076 Tuebingen, Germany

Edited by Richard M. Shiffrin, Indiana University, Bloomington, IN, and approved April 5, 2016 (received for review June 5, 2015)

Received: 1 June 2018 | Accepted: 11 September 2020

DOI: 10.1111/rssb.12398

ORIGINAL ARTICLE

Anchor regression: Heterogeneous data meet causality

Dominik Rothenhäusler[1] | Nicolai Meinshausen[2] | Peter Bühlmann[2] | Jonas Peters[3]

Invariant Causal Prediction – ICP (Peters et al., 2016)

Remarks

- Computation of ICP can be expensive
  - Existence of algorithm which computes ICP without necessarily going through all subsets (in worst case this cannot be avoided)
  - In high dimensional setting: variable screening
- Presence of hidden confounding factors, ICP leads to

$$\mathbb{P}[\hat{S}(\mathcal{E}) \subseteq \mathrm{an}(Y)] \geq 1 - \alpha,$$

  with an(Y) = ancestors of Y

- Direct effects of environments on Y (Violation of $\boldsymbol{B}(\mathcal{E})$):
  - Infer that no set S_k fulfills invariance condition
  - As sample size gets sufficiently large, rejecting $H_{0,S_k}(\mathcal{E})$ for all S_k
  - $\hat{S}(\mathcal{E}) = \emptyset$

Invariant Causal Prediction – ICP (Peters et al., 2016)

Concrete test for invariance

$$Y = \sum_{j \in \mathrm{pa}(Y)} \beta_j X_j + \varepsilon_Y, \ \varepsilon_Y \sim \mathcal{N}(0, \sigma_Y^2)$$

and ε_Y is independent of $X_{\mathrm{pa}(Y)}$. The invariance hypotheses in $H_{0,S}(\mathcal{E})$ then becomes:

$H_{0,S}(\mathcal{E})^{\mathrm{lin-Gauss}}$**:** for all $e \in \mathcal{E}$ its holds that,

$$\begin{aligned} &Y^e = X_S^e \beta_S + \varepsilon_S^e, \ \varepsilon_S^e \text{ independent of } X_S^e \text{ (the same } \beta_S \text{ for all } e \in \mathcal{E}), \\ &\varepsilon_S^e \sim F_{\varepsilon_S} \text{ (the same for all } e \in \mathcal{E}). \end{aligned}$$

- Exact tests exist, e.g. Chow test (tests if true coefficients in two linear regressions on different data sets are equal)
- Variable screening using e.g. LASSO

Invariant Causal Prediction – ICP (Peters et al., 2016)

Unknown environments

- Estimate from data
- Type 1 error control holds as long as estimated partition does not involve descendant variables of the response Y
- Use clustering algorithm based on non-descendants of Y

Anchor Regression (Rothenhäusler et al. 2018)

Choosing amount of regularization

- γ relates to the class of shift perturbations over which we achieve protection (in worst case)
  - Decide via cross validation
  - Decide a-priori based on expected perturbation in data (domain knowledge)

- If anchor variables are continuous: Interpretation as a quantile
  - Assume joint Gaussian distribution over A, X, Y

$$
\begin{aligned}
& \alpha - \text{quantile of } \mathbb{E}[(Y - X^T b)^2 | A] \\
= \; & \mathbb{E}[((I - P_A)(Y - X^T\beta))^2] + \gamma \mathbb{E}[(P_A(Y - X^T\beta))^2], \\
& \text{for } \gamma = \alpha - \text{quantile of } \chi_1^2.
\end{aligned}
$$

  - Thus, choose α and then calculate the γ which optimizes this quantile