Causality-oriented robustness:
exploiting data heterogeneity at different levels

Xinwei Shen

Seminar for Statistics, ETH Zurich

Peter Bühlmann, Armeen Taeb, Alexander Henzi, Michael Law

March 5, 2024

Classical statistical learning: identically distributed

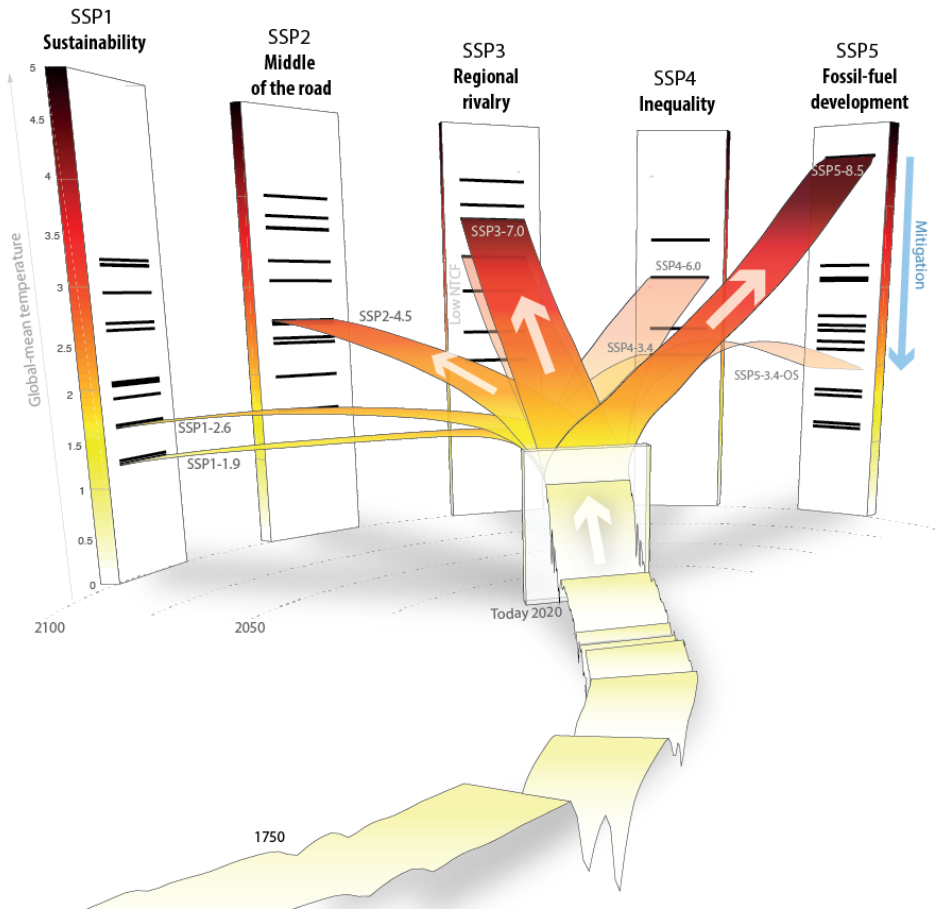
- Predictors $X \in \mathbb{R}^p$, response $Y \in \mathbb{R}$, $(X, Y) \sim P$
- Prediction model $\hat{Y} = f_\theta(x)$, e.g., linear model, neural network
- Optimality of $f_\theta(x)$ for a fixed distribution:

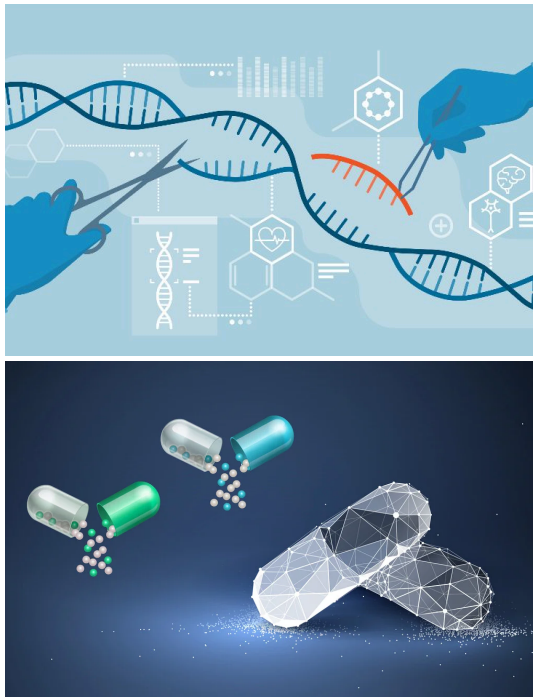
$$\min_\theta \mathbb{E}_{(X,Y)\sim P}[\ell(Y, f_\theta(X))]$$

  where ℓ is a given loss function.
- Method: empirical risk minimization (ordinary least-squares)

Statistical learning in applications: distribution shifts

- Climate prediction under different climate change scenarios.
- Genomic/proteomic response modeling under genetic perturbations or drug combinations.





Statistical learning in applications: distribution shifts

- Distributional robustness: optimality for a class of distributions $\mathcal{P}$

$$\min_{\theta} \sup_{P \in \mathcal{P}} \mathbb{E}_P[\ell(Y, f_\theta(X))]$$

- Distributionally robust optimization (DRO): $\mathcal{P}_\delta = \{P : D(P, P_0) \leq \delta\}$ with D being the Wasserstein distance or f-divergence.

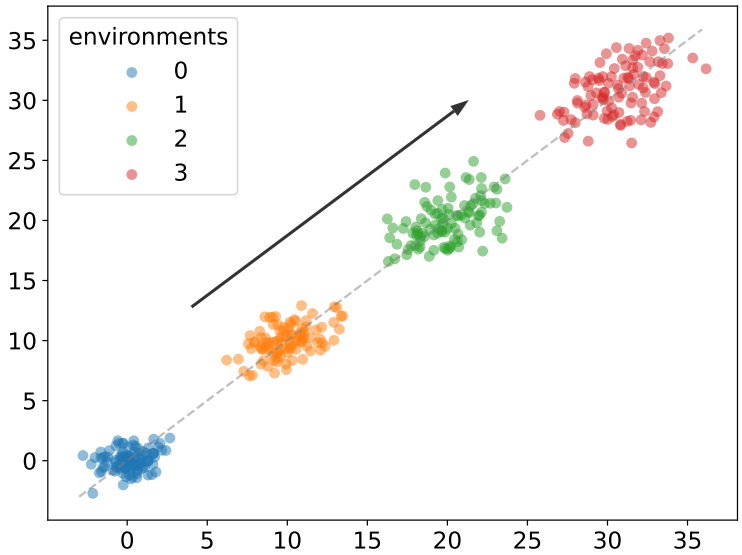
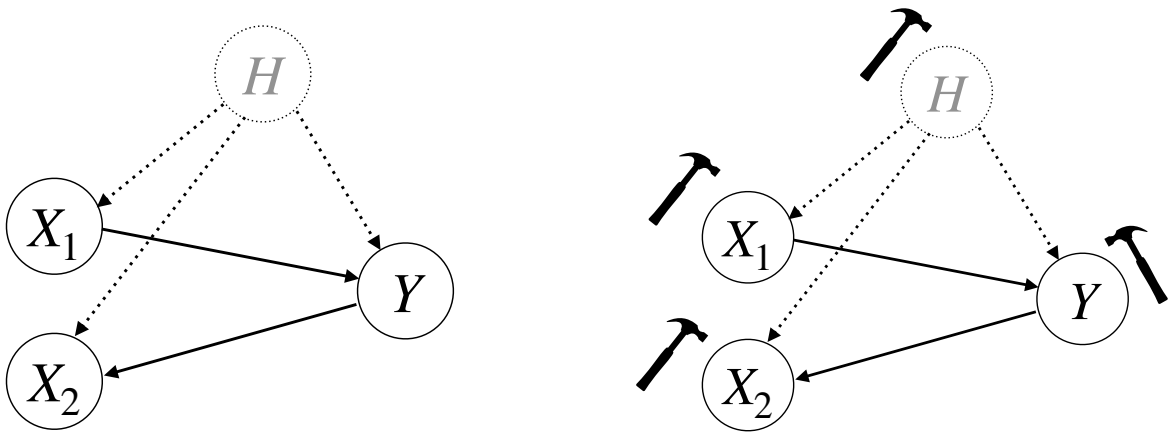


Figure: Shifts in certain directions.

Ben-Tal et al. '98; Duchi and Namkoong '17; Sagawa et al. '20

A causal perspective for distribution shifts

- Model (X, Y) using a causal framework — structural causal model
- Model distribution shifts as "interventions"



Meinshausen '18; Bühlmann '20; Christiansen et al. '21

Exploiting heterogeneity in multi-environment data

- Data are collected from multiple environments/sources:

$$Z^e := (X^e, Y^e) \sim P^e, \ e \in \mathcal{E}$$

- E.g., climate change scenarios, genetic perturbations, hospitals, etc.
- Heterogeneity occurs at different levels:
  - Mean shift: $\mathbb{E}[Z^e] = \mathbb{E}[Z^{e'}]$ for $e \neq e'$
  - Variance shift: $\mathrm{Var}(Z^e) = \mathrm{Var}(Z^{e'})$ for $e \neq e'$; could be $\mathbb{E}[Z^e] \equiv \mu$
  - Distribution shift: P^e's can differ in general ways, e.g., quantiles or higher-order moments; could be $\mathbb{E}[Z^e] \equiv \mu$, $\mathrm{Var}(Z^e) \equiv \Sigma$
- Goal: a prediction model that is robust against future perturbations along the "directions" of observed heterogeneity.

Overview

Under a causal framework, we aim at robust prediction by **exploiting data heterogeneity** at different levels.

- Mean shift: anchor regression (Rothenhäusler et al. '21)
- **Variance shift**: *Causality-oriented robustness: exploiting general additive interventions*. arXiv:2307.10299
- **Distribution shift**: *Invariant Probabilistic Prediction*. arXiv:2309.10083

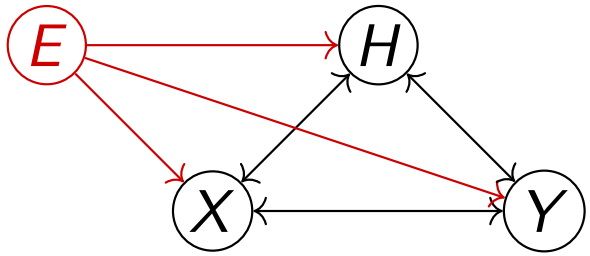
Causality-oriented robustness: exploiting general additive interventions

with Peter Bühlmann and Armeen Taeb

Training setup: multi-environment data

Training data $Z^e := (X^e, Y^e)$, $e \in \mathcal{E}$ are generated via a linear SCM:

$$\begin{pmatrix} X^e \\ Y^e \end{pmatrix} = B^\star \begin{pmatrix} X^e \\ Y^e \end{pmatrix} + \varepsilon + \delta^e$$



- $B^\star$: adjacency matrix; ε: exogenous variables (correlated components)
- δ^e: additive interventions that generate variance shifts $\mathrm{Var}(Z^e) = \Sigma^e$
- Structural equation for Y (where $b^\star$ represents the causal effects):

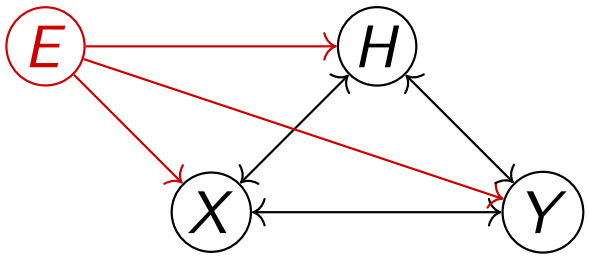
$$Y^e = {b^\star}^\top X^e + \varepsilon_Y + \delta_Y^e.$$

- Observational environment $e = 0$: $\delta^0 = 0$, no intervention
- Interventional environments $e \neq 0$: $\delta^e \neq 0$

Test setup: new additive intervention

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

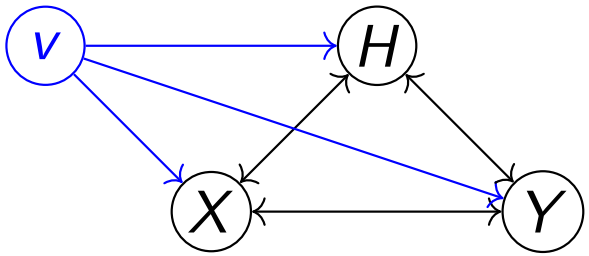
$$\begin{pmatrix} X^e \\ Y^e \end{pmatrix} = B^\star \begin{pmatrix} X^e \\ Y^e \end{pmatrix} + \varepsilon + \delta^e$$



Test data (X^v, Y^v):

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

where random variable v is a new intervention.



Goal: robust against test distributions of (X^v, Y^v) for v in a certain (data dependent) distribution class.

Our method DRIG

DRIG (Distributional Robustness via Invariant Gradients)

The population DRIG estimator for $\gamma \geq 0$ is defines as

$$b_\gamma := \operatorname*{argmin}_b \Big\{ \underbrace{\mathbb{E}[\ell(X^0, Y^0; b)]}_{\text{observational MSE}} + \gamma \sum_{e \in \mathcal{E}} \omega^e \underbrace{\big(\mathbb{E}[\ell(X^e, Y^e; b)] - \mathbb{E}[\ell(X^0, Y^0; b)]\big)}_{\text{difference of interventional and observational MSE}} \Big\}$$

where $\ell(x, y; b) := (y - b^\top x)^2$, $\sum_{e \in \mathcal{E}} \omega^e = 1$, and $\omega^e \geq 0$.

When $\gamma \to \infty$, optimal b satisfies the **gradient invariance condition**:

$$\sum_{e \in \mathcal{E}} \omega^e \nabla_b \mathbb{E}[\ell(X^e, Y^e; b)] = \nabla_b \mathbb{E}[\ell(X^0, Y^0; b)]$$

DRIG regularizes towards gradient invariance.

Special cases of DRIG

The population DRIG estimator:

$$b_\gamma := \operatorname*{argmin}_b \Big\{ \underbrace{\mathbb{E}[\ell(X^0, Y^0; b)]}_{\text{observational MSE}} + \gamma \sum_{e \in \mathcal{E}} \omega^e \underbrace{\big(\mathbb{E}[\ell(X^e, Y^e; b)] - \mathbb{E}[\ell(X^0, Y^0; b)]\big)}_{\text{difference of interventional and observational MSE}} \Big\}$$

Special cases:

- $\gamma = 0$: OLS with the observational data
- $\gamma = 1$: OLS with the pooled data
- $\gamma \to \infty$: causal effects $b^\star$ (when identifiable)

"Causality-oriented": DRIG interpolates between OLS and causal effects.

Special cases of DRIG

Special cases when there are only **mean shifts**, i.e., δ^e's are deterministic:

$$\begin{pmatrix} X^e \\ Y^e \end{pmatrix} = B^\star \begin{pmatrix} X^e \\ Y^e \end{pmatrix} + \varepsilon + \delta^e$$

- $\gamma \geq 0$: anchor regression (Rothenhäusler et al. '21) with categorical anchor E

$$\min_b \mathbb{E}[((Id - P_E)(Y - b^\top X))^2] + \gamma \mathbb{E}[(P_E(Y - b^\top X))^2]$$

  where $P_E(\cdot) = \mathbb{E}[\cdot|E]$ and $Id(\cdot) = \cdot$.

- $\gamma \to \infty$: two-stage least squares (IV regression)

Robustness guarantee for general γ

Test distribution

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

Theorem

The population DRIG estimator for $\gamma \geq 0$ satisfies

$$b_\gamma = \underset{b}{\operatorname{argmin}} \sup_{v \in \mathcal{C}^\gamma} \mathbb{E}[(Y^v - b^\top X^v)^2]$$

where $\mathcal{C}^\gamma := \left\{ v \in \mathbb{R}^{p+1} : \mathbb{E}[vv^\top] \preceq \gamma \sum_{e \in \mathcal{E}} \omega^e \mathbb{E}[\delta^e \delta^{e\top}] \right\}$.

DRIG prediction model is robust against perturbations in the class $\mathcal{C}^\gamma$.

- γ: perturbation **strength**
- $\sum_{e \in \mathcal{E}} \omega^e \mathbb{E}[\delta^e \delta^{e\top}]$: perturbation **directions**

More robustness by exploiting variance shifts

Larger rank$(\sum_{e\in\mathcal{E}} \omega^e \mathbb{E}[\delta^e \delta^{e\top}]) \rightarrow$ more "directions" we are robust against

- Anchor regression exploits deterministic δ^e (mean shifts):
  #robust directions $\leq$ #observed environments
- DRIG exploits general δ^e (variance shifts):
  can be robust in all directions (with 2 environments)

DRIG can exploit heterogeneity in variance to protect against perturbations in more directions.

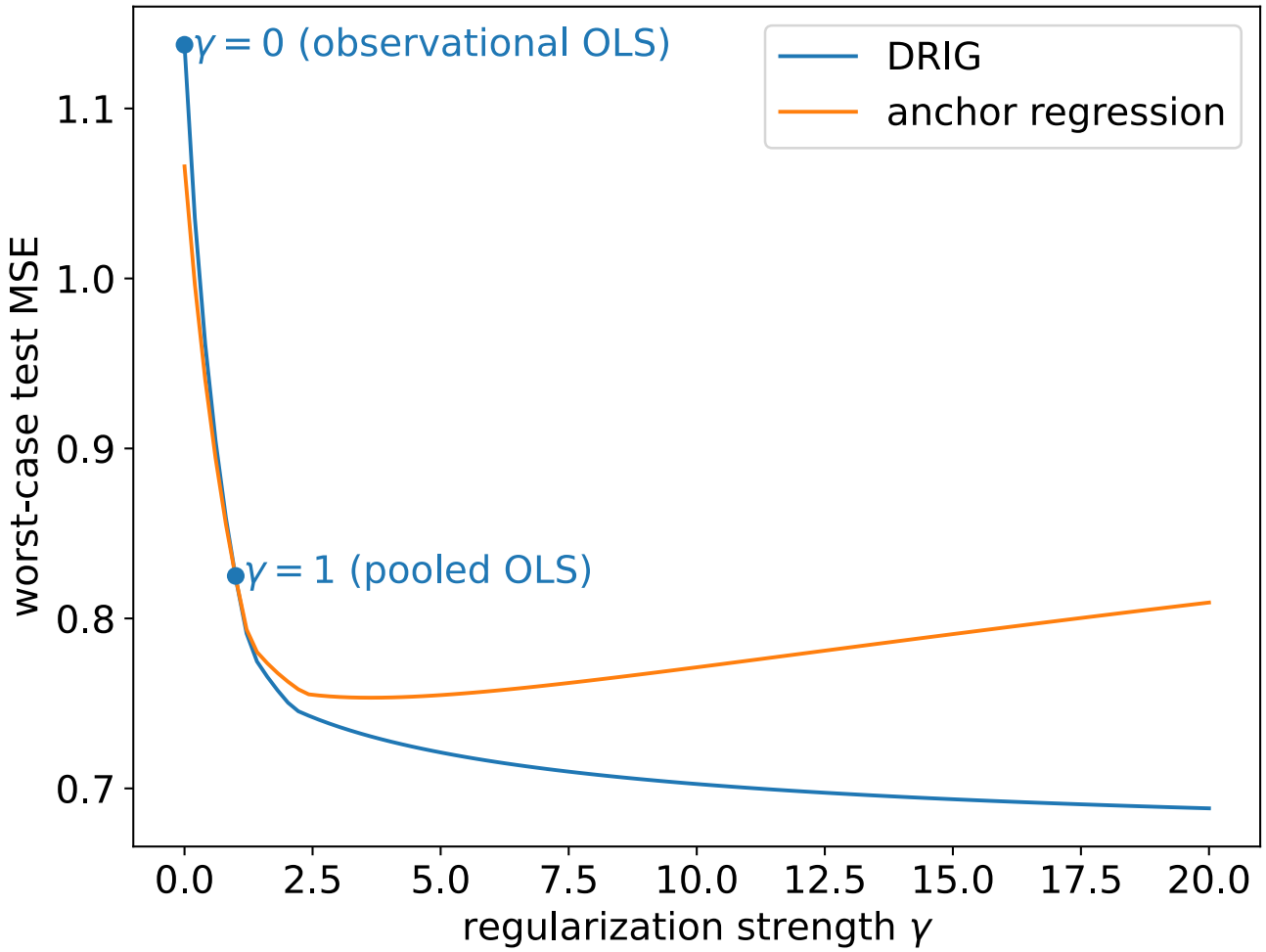
Single-cell data application

Single-cell RNA-sequencing data (Replogle et al. '22)

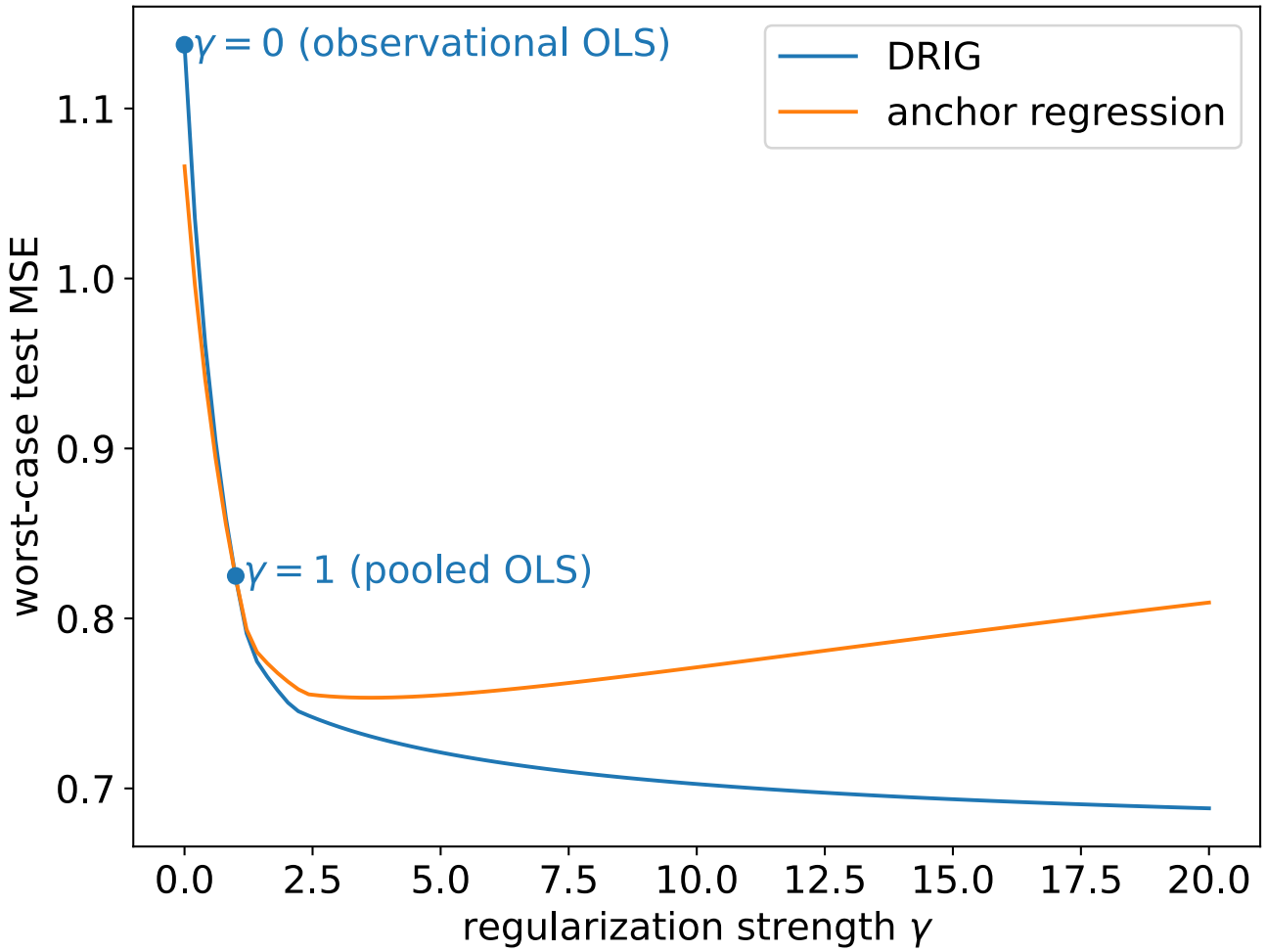
- 10 genes: a response and 9 predictors
- 10 training environments: 1 observational + 9 interventional
- 50 test environments that can be very different from training environments, due to interventions on unobserved genes.

How robust is our prediction model on test environments?

Results on single-cell data



Results on single-cell data



Want to exploit shifts beyond the mean and variance → distribution!

Invariant Probabilistic Prediction

with Alexander Henzi, Michael Law and Peter Bühlmann

Invariant Probabilistic Prediction (IPP)

- Training data: for $e = 1, \ldots, m$,

$$X^e = h^e(\varepsilon_X)$$
$$Y^e = g^\star(X^e, \varepsilon_Y)$$

- Given a proper scoring rule S and a model $P_\theta(y|x)$, risk per environment:[1]

$$\mathcal{R}_S^e(\theta) = \mathbb{E}[-S(P_\theta(y|X^e), Y^e)]$$

- Population **IPP**:

$$\min_\theta \ \frac{1}{m}\sum_{e=1}^m \mathcal{R}_S^e(\theta) + \lambda D(\mathcal{R}_S^1(\theta), \ldots, \mathcal{R}_S^m(\theta))$$

where $D(v) = \frac{1}{m^2}\sum_{i,j=1}^m (v_i - v_j)^2$, $\lambda \geq 0$ tuning parameter.

[1] Engression: Extrapolation for Nonlinear Regression. **S.** and Meinshausen '23

Identification for distributional causal effects

- Do-intervention: $P^\star(y|x)$ denotes the distribution of Y under $\mathrm{do}(X = x)$

$$X^{\mathrm{int}} = x \text{ for any } x$$
$$Y^{\mathrm{int}} = g^\star(x, \varepsilon_Y)$$

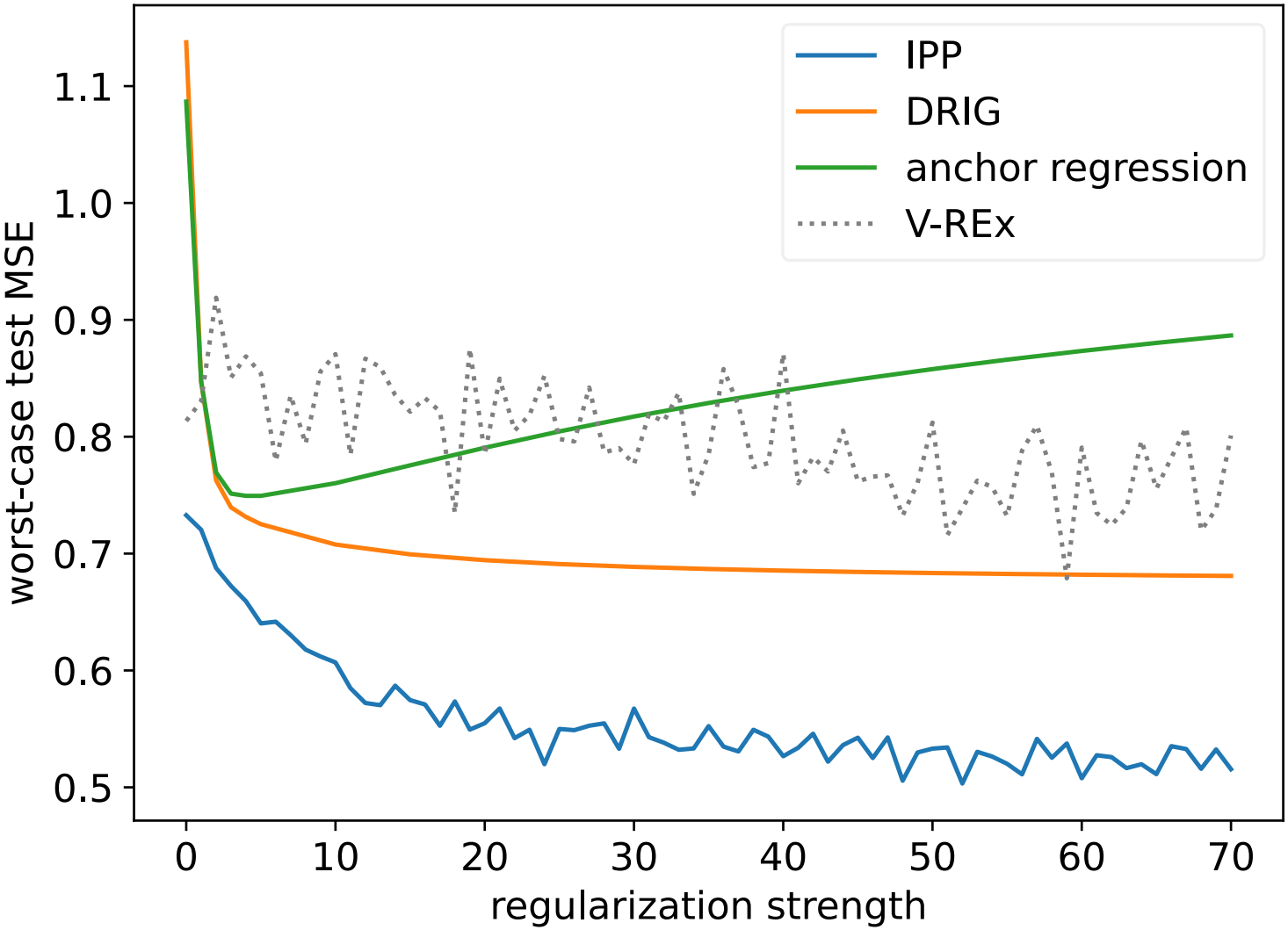
  When there is hidden confounding, we may have $P^\star(y|x) \neq P_{\mathrm{obs}}(y|x)$.

- In model

$$Y = \beta^\top X + \exp(\gamma^\top X)\varepsilon_Y,$$

  where $\varepsilon_Y \sim \mathcal{N}(0, 1)$, we provide sufficient conditions for *IPP (as* $\gamma \to \infty$*) to identify* $P^\star$ with the logarithmic score and SCRPS (Bolin & Wallin '23).

Results on single-cell data



Krueger et al. '21

Summary

Causality-oriented robust prediction by **exploiting data heterogeneity** at different levels.

- Mean shift: anchor regression (Rothenhäusler et al. '21)
- **Variance shift**: *Causality-oriented robustness: exploiting general additive interventions*. arXiv:2307.10299
- **Distribution shift**: *Invariant Probabilistic Prediction*. arXiv:2309.10083

Outlook:

- Robustness guarantee for finite perturbation strengths in distributional or nonlinear settings
  - *Engression: Extrapolation for Nonlinear Regression*. arXiv:2307.00835
- Causal effects identification by exploiting heterogeneity in distributions
  - *Distributional Instrumental Variable Regression* (coming soon)

Summary

Thank you!

Causality-oriented robust prediction by **exploiting data heterogeneity** at different levels.

- Mean shift: anchor regression (Rothenhäusler et al. '21)
- **Variance shift**: *Causality-oriented robustness: exploiting general additive interventions*. arXiv:2307.10299
- **Distribution shift**: *Invariant Probabilistic Prediction*. arXiv:2309.10083

Outlook:

- Robustness guarantee for finite perturbation strengths in distributional or nonlinear settings
  - *Engression: Extrapolation for Nonlinear Regression*. arXiv:2307.00835
- Causal effects identification by exploiting heterogeneity in distributions
  - *Distributional Instrumental Variable Regression* (coming soon)