Robust and causal-oriented prediction models from heterogeneous data

Xinwei Shen[†], Peter Bühlmann[†], and Armeen Taeb[‡]

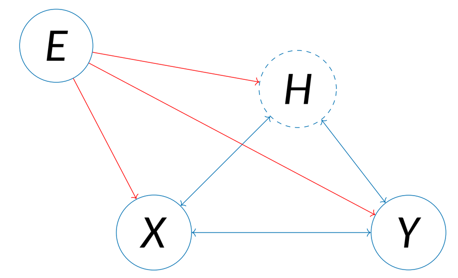
[†] Seminar for Statistics, ETH Zürich, [‡]Department of Statistics, University of Washington

Introduction

- Heterogeneous data provides opportunities for causal inference and for learning prediction models that generalize to unseen environments.
- Perturbations may affect both means and variances of the variables, while previous methods only exploit shifts in the means.
- We propose Distributionally Robust predictions via Invariant Gradients (DRIG), a method that leverages perturbations in the form of both mean and variance shifts for robust predictions.
- Viewing causality as an extreme case of distributional robustness, we investigate the causal identifiability of DRIG under various scenarios of interventions and causal structures.

Linear structural causal model

Covariates $X \in \mathbb{R}^p$ and response variable $Y \in \mathbb{R}$ with latent variables H.



- Training data from multiple environments $e \in \mathcal{E}$:

$$\begin{pmatrix} X^e \\ Y^e \end{pmatrix} = B^\star \begin{pmatrix} X^e \\ Y^e \end{pmatrix} + \varepsilon + \delta^e, \tag{1}$$

where $b^\star := B^\star_{p+1,1:p}$ denotes the causal effects and $\varepsilon \perp\!\!\!\perp \delta^e$.

- *Reference environment*: $0 \in \mathcal{E}$ such that $\sum_{e\in\mathcal{E}} \omega^e \mathbb{E}[\delta^e \delta^{e\top}] \succeq \mathbb{E}[\delta^0\delta^{0\top}]$. e.g., an observational environment with no intervention, i.e., $\delta^0 \equiv 0$.
- Test distribution under new interventions:

$$\begin{pmatrix} X^v \\ Y^v \end{pmatrix} = B^\star \begin{pmatrix} X^v \\ Y^v \end{pmatrix} + \varepsilon + v, \tag{2}$$

Our method DRIG

Given a scalar $\gamma \geq 0$, the population DRIG: $b^{\text{opt}}_\gamma = \operatorname{argmin}_b \mathcal{L}_\gamma(b)$ where

$$\mathcal{L}_\gamma(b) := \mathbb{E}[\ell(X^0, Y^0; b)] + \gamma \sum_{e\in\mathcal{E}} \omega^e \left(\mathbb{E}[\ell(X^e, Y^e; b)] - \mathbb{E}[\ell(X^0, Y^0; b)]\right), \tag{3}$$

where $\ell(x, y; b) := (y - b^\top x)^2$ and γ is a hyperparameter.

Special cases: interpolation between OLS and the causal parameter:

- $\gamma = 0$: observational OLS
- $\gamma = 1$: pooled OLS
- $\gamma \to \infty$: causal parameter (when identifiable)
- δ^e's are deterministic: anchor regression with categorical anchors

In the limit of $\gamma \to \infty$, the DRIG solution b^{opt}_γ satisfies gradient invariance:

Definition (Gradient invariance)

A regression parameter b is said to satisfy the gradient invariance (GI) condition if $\sum_{e\in\mathcal{E}} \omega^e \nabla_b \mathbb{E}[\ell(X^e, Y^e; b)] = \nabla_b \mathbb{E}[\ell(X^0, Y^0; b)]$.

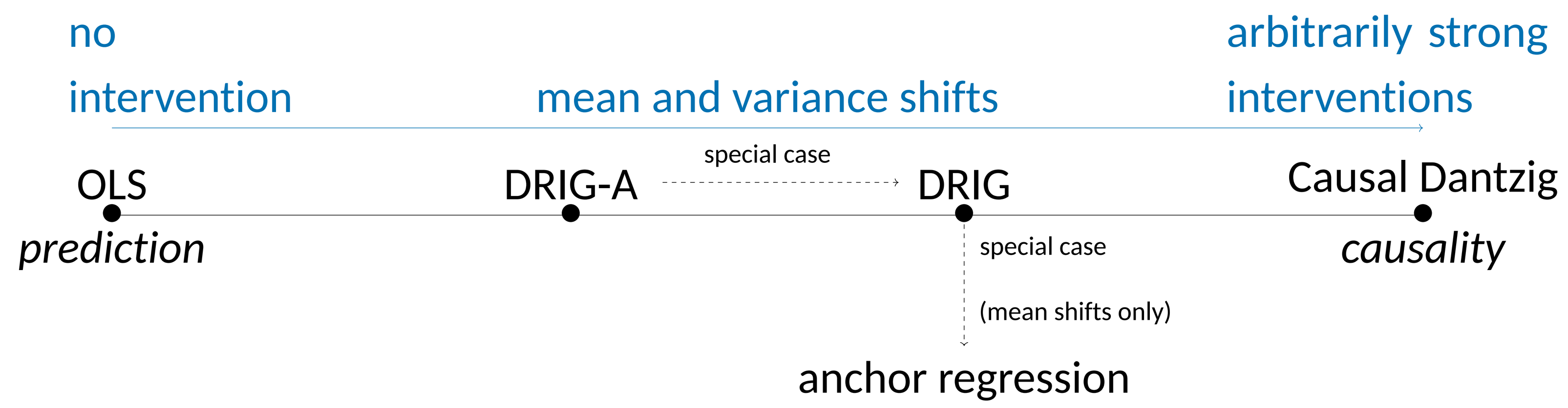


Figure: Trade-off between prediction and causality

Distributional robustness

Minimizing the worst-case risk over test perturbations $v \in \mathcal{C} \subseteq \mathbb{R}^{p+1}$, i.e.

$$\operatorname*{argmin}_{b\in\mathbb{R}^p} \sup_{v\in\mathcal{C}} \mathbb{E}_v[\ell(X^v, Y^v; b)]. \tag{4}$$

Theorem (Robustness guarantee)

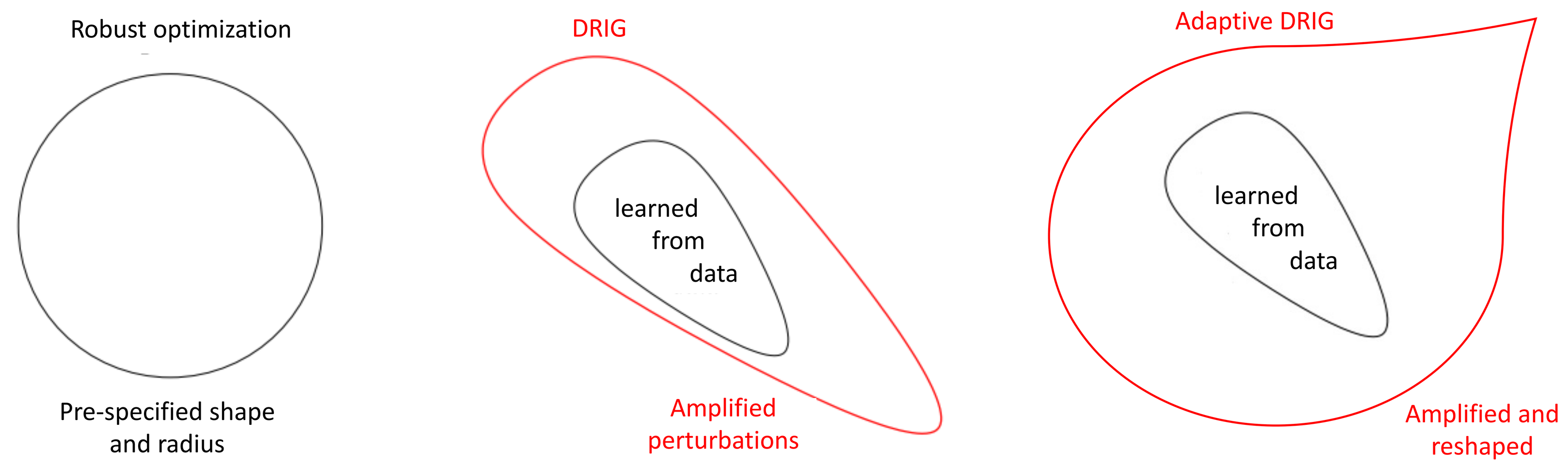
The prediction model by DRIG is the solution to the worst-case risk (4) with

$$\mathcal{C}^\gamma_{\mathrm{DRIG}} := \left\{ v \in \mathbb{R}^{p+1} : \mathbb{E}[vv^\top] \preceq \gamma \sum_{e\in\mathcal{E}} \omega^e \left(S^e - S^0 + \mu^e\mu^{e\top}\right) \right\},$$

where $S^e := cov(\delta^e)$ and $\mu^e := \mathbb{E}[\delta^e]$

- Strength of perturbations: controlled by γ
- Directions of perturbations: row and column spaces of $\sum_{e\in\mathcal{E}} \omega^e(S^e - S^0 + \mu^e\mu^{e\top})$

Comparing to traditional distributionally robust optimization (DRO):



Comparing to other methods: for $\gamma \geq 1$, we have $\mathcal{C}_{\mathrm{oOLS}} \subseteq \mathcal{C}_{\mathrm{pOLS}} \subseteq \mathcal{C}^\gamma_{\mathrm{anchor}} \subseteq \mathcal{C}^\gamma_{\mathrm{DRIG}}$.

- observational OLS: $\mathcal{C}_{\mathrm{oOLS}} = \left\{v \in \mathbb{R}^{p+1} : \mathbb{E}[vv^\top] \preceq \sum_{e\in\mathcal{E}} \omega^e \left(S^e - S^0 + \mu^e\mu^{e\top}\right)\right\}$
- pooled OLS: $\mathcal{C}_{\mathrm{pOLS}} = \left\{v \in \mathbb{R}^{p+1} : \mathbb{E}[vv^\top] \preceq \sum_{e\in\mathcal{E}} \omega^e \left(S^e - S^0 + \mu^e\mu^{e\top}\right)\right\}$
- anchor regression: $\mathcal{C}^\gamma_{\mathrm{anchor}} = \left\{v \in \mathbb{R}^{p+1} : \mathbb{E}[vv^\top] \preceq \sum_{e\in\mathcal{E}} \omega^e \left(S^e - S^0 + \gamma\mu^e\mu^{e\top}\right)\right\}$
- causal parameter: $\mathcal{C}_{\mathrm{causal}} = \left\{v \in \mathbb{R}^{p+1} : v_{p+1} \equiv 0\right\}$

Causal identification

DRIG solution with $\gamma \to \infty$:

$$b^{\text{opt}}_\infty := \lim_{\gamma\to\infty} b^{\text{opt}}_\gamma = \operatorname*{argmin}_{b \text{ satisfies GI}} \mathbb{E}[(Y^0 - b^\top X^0)^2].$$

Identifiable cases: $b^{\text{opt}}_\infty = b^\star$

- Sufficient interventions on X & no interventions on Y or H
- Sufficient interventions on X & independent interventions on Y & Y is childless.

Unidentifiable cases: approximate identifiability $\|b^{\text{opt}}_\infty - b^\star\| \leq c$

- Interventions on the latent variables with dense latent effects
- Insufficient interventions on X

DRIG-A: adaptive DRIG in semi-supervised domain adaptation settings

- DRIG with matrix Γ for more flexible robustness: b^{opt}_Γ is minimizing

$$\mathcal{L}_\Gamma(b) := \mathbb{E}[(Y^0 - b^\top X^0)^2] + \sum_{e\in\mathcal{E}} \omega^e \left(\mathbb{E}[\gamma_y Y^e - b^\top \Gamma_x X^e]^2 - \mathbb{E}[\gamma_y Y^0 - b^\top \Gamma_x X^0]^2\right).$$

with a closed-form solution

$$b^{\text{opt}}_\Gamma = [\mathbb{E}X^0 X^{0\top} + \Gamma_x \Delta_x \Gamma_x]^{-1}[\mathbb{E}X^0Y^0 + \gamma_y \Gamma_x \Delta_{xy}].$$

where $\Delta_x = \sum_{e\in\mathcal{E}} \omega^e[\mathbb{E}X^eX^{e\top} - \mathbb{E}X^0X^{0\top}]$ and $\Delta_{xy} = \sum_{e\in\mathcal{E}} \omega^e[\mathbb{E}X^eY^e - \mathbb{E}X^0Y^0]$.

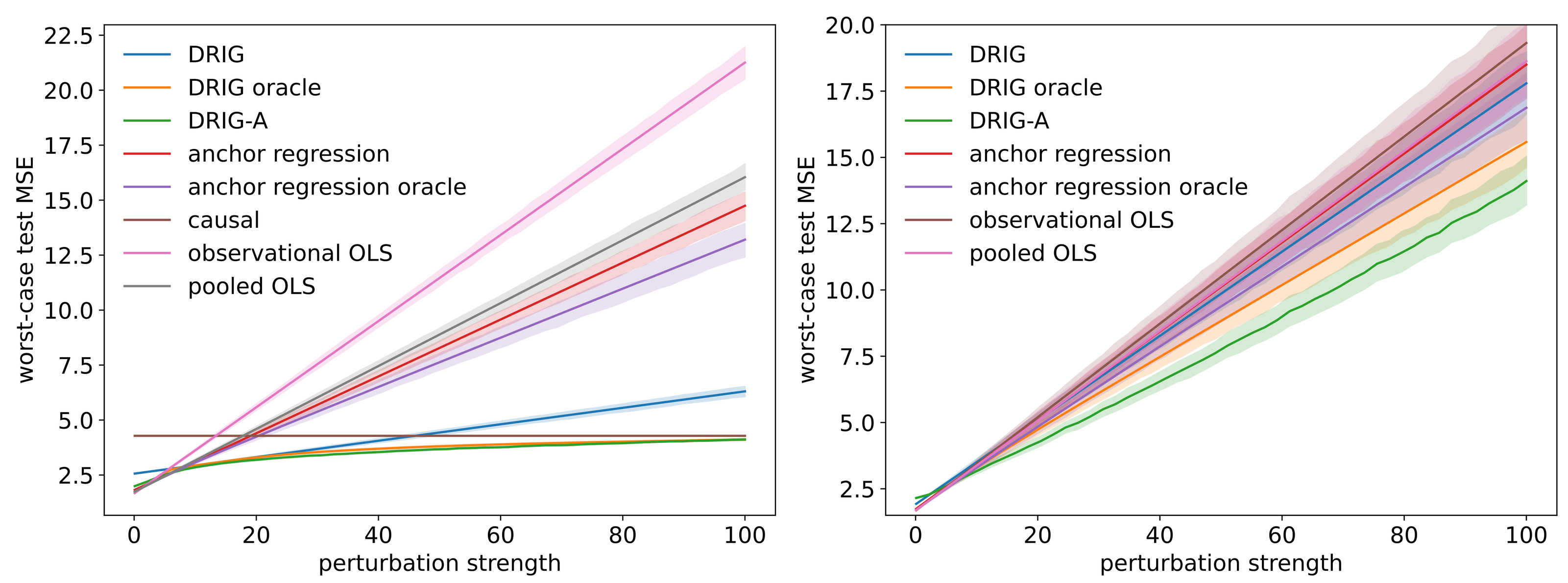
- Test distribution P_{test} according to SCM (2)
  - a small labeled sample $\{(X^v_i, Y^v_i) \sim P_{\mathrm{test}}, i = 1\ldots, n_l\}$
  - a large unlabeled test samples $\{X^v_i \sim P^x_{\mathrm{test}}, i = 1\ldots, n_u\}$.
- Test OLS $\hat{b}_{\mathrm{tOLS}} := (\frac{1}{n_u}\sum_{i=1}^{n_u} X^v_i X^{v\top}_i)^{-1}(\frac{1}{n_l}\sum_{i=1}^{n_l} X^v_i Y^v_i)$
- Choosing Γ based on the semi-supervised test sample: in population

$$\begin{aligned} \min_{\Gamma_x, \gamma_y} \quad & \mathbb{E}[(Y^v - b^{\text{opt}\top}_\Gamma X^v)^2] \\ \text{s.t.} \quad & \mathbb{E}X^0X^{0\top} + \Gamma_x\Delta_x\Gamma_x = \mathbb{E}X^vX^{v\top} \end{aligned}$$

Theorem. *Assume $p > 1$; $\mathrm{Var}(X^uY^u) \succ \left(\mathbb{E}[X^uY^u] - \mathbb{E}[X^0Y^0]\right)^{\otimes 2}$. Then $\exists N_u, N_l > 0$ such that when $n_u \geq N_u$ and $n_l \leq N_0$, we have $\mathbb{E}[\mathcal{L}_{test}(b^{opt}_{\hat{\Gamma}})] < \mathbb{E}[\mathcal{L}_{test}(\hat{b}_{\mathrm{tOLS}})]$, where the expectation is taken over all test samples.*

Simulations

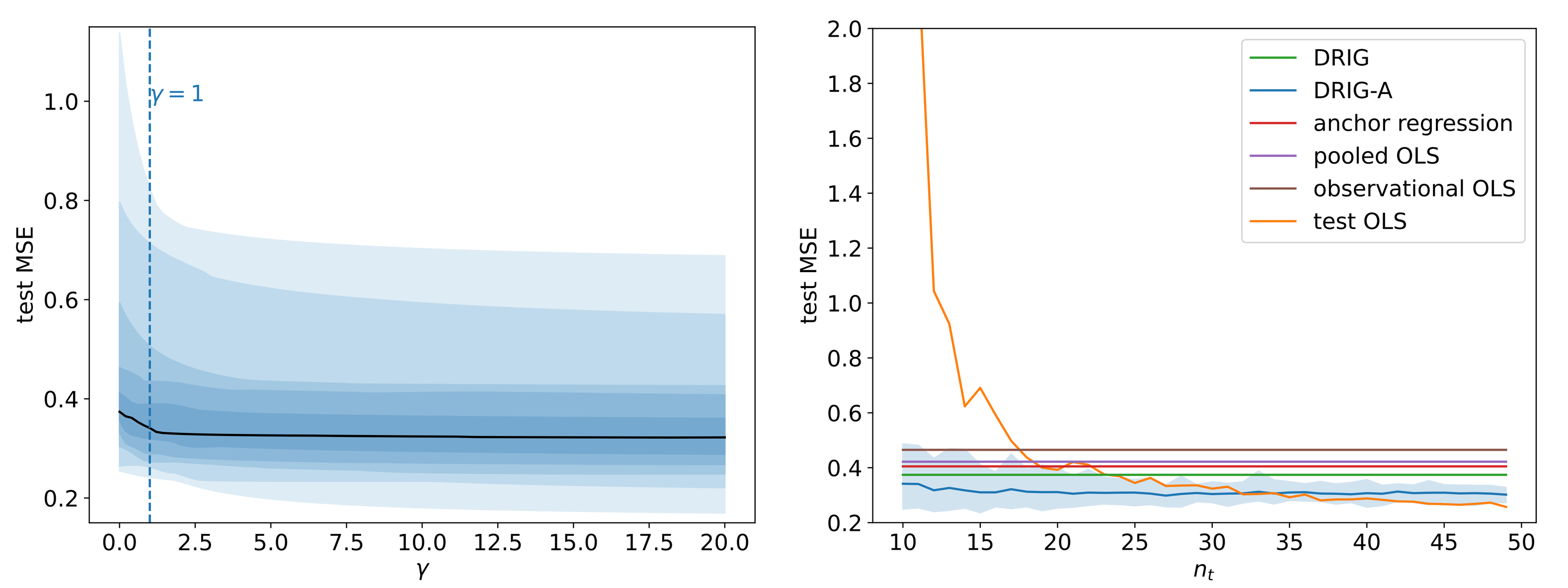
Worst-case MSEs over 20 randomly simulated test environments:



(a) no intervention on Y or H (b) intervention on Y and H

Single-cell data

- 10 genes (1 response), 11,485 observational data, 10 interventional environments.
- Hundreds of test environments; on each of them, one hidden gene is intervened.



(a) Quantiles of MSEs on 50 test environments (b) semi-supervised setting