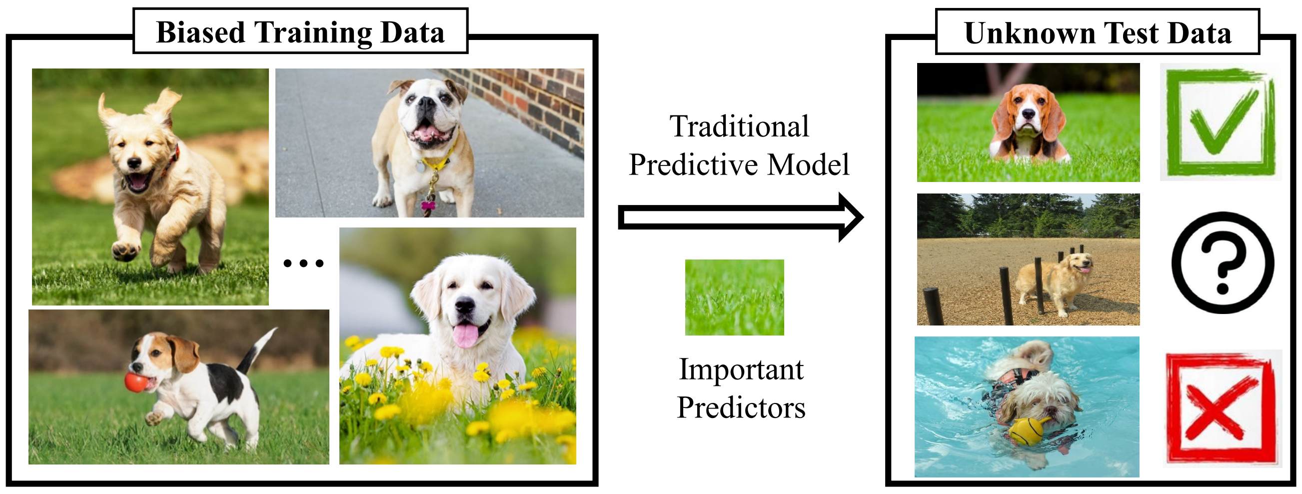
Stable Prediction with Model Misspecification and Agnostic Distribution Shift

Kun Kuang[1,2](kunkuang@zju.edu.cn), Ruoxuan Xiong[3], Peng Cui[2], Susan Athey[3], Bo Li[2]
[1]Zhejiang University, [2]Tsinghua University, [3]Stanford University

INTRODUCTION



Unstable prediction of traditional predictive model, **WHY?**

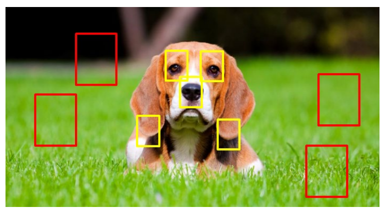
- **Model misspecification and correlation based**
  - Spurious correlation (**Unexplainable**)
- **Agnostic distribution shift**
  - Variant of spurious correlation (**Unstable**)

PROBLEM

Problem 1 (Stable Prediction) ***Given*** *one training environment $e \in \mathcal{E}$ with dataset $D^e = (\mathbf{X}^e, Y^e)$, the task is to* ***learn*** *a predictive model to predict across unknown environment $\mathcal{E}$ with not only small Average_Error but also small Stability_Error.*

$$Average\_Error = \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} RMSE(D^e), \quad (1)$$

$$Stability\_Error = \sqrt{\frac{1}{|\mathcal{E}|-1}\sum_{e\in\mathcal{E}} (RMSE(D^e) - Average\_Error)^2} \quad (2)$$



Human like thinking: What cause the object to be a dog?
Suppose $X = \{S, V\}$, and the label $Y = f(S) + \varepsilon$
We define S as **stable features** and V as **unstable features**
Assumption 1: $P(Y|S)$ is invariant across environments
Assumption 2: All stable features S are observed

Model misspecification: ignoring non-linear term $g(S)$

$$Y^e = f(\mathbf{S}^e) + \mathbf{V}^e\beta_V + \epsilon^e = \mathbf{S}^e\beta_S + g(\mathbf{S}^e) + \mathbf{V}^e\beta_V + \varepsilon^e.$$

where $\beta_V = 0$ and $\varepsilon^e \perp \mathbf{X}^e$.

$$\hat{\beta}_{V_{OLS}} = \beta_V + \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{V}_i^T\mathbf{V}_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{V}_i^T g(\mathbf{S}_i)\right) + \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{V}_i^T\mathbf{V}_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{V}_i^T\mathbf{S}_i\right)(\beta_S - \hat{\beta}_{S_{OLS}}), (4)$$

$$\hat{\beta}_{S_{OLS}} = \beta_S + \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{S}_i^T\mathbf{S}_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{S}_i^T g(\mathbf{S}_i)\right) + \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{S}_i^T\mathbf{S}_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{S}_i^T\mathbf{V}_i\right)(\beta_V - \hat{\beta}_{V_{OLS}}), (5)$$

$\hat{\beta}_{V_{OLS}}$ is biased if $\mathbb{E}(\mathbf{V}^T g(\mathbf{S})) \neq 0$ or $\mathbb{E}(\mathbf{V}^T\mathbf{S}) \neq 0$ in Eq. (4), leading to the biased estimation on $\hat{\beta}_{S_{OLS}}$ in Eq. (5)

Spurious Correlation between S and V might vary across environments, resulting in unstable prediction across unknown environments.

Solution: precisely estimate the $\hat{\beta}_{V_{OLS}}$ by removing spurious correlation

Let $\mathbb{E}(\mathbf{V}^T g(\mathbf{S})) = 0$ and $\mathbb{E}(\mathbf{V}^T\mathbf{S}) = 0$

METHOD

Proposition 1 *If $\mathbf{X}$ are mutually independent with mean 0, then $\mathbb{E}(\mathbf{V}^T g(\mathbf{S})) = 0$ and $\mathbb{E}(\mathbf{V}^T\mathbf{S}) = 0$.*

Making variable independent by *sample reweighting*:

$$\min_W \sum_{a=1}^{\infty}\sum_{b=1}^{\infty} \|\mathbb{E}[\mathbf{X}_{,j}^{a^T}\mathbf{\Sigma}_W\mathbf{X}_{,k}^{b}] - \mathbb{E}[\mathbf{X}_{,j}^{a^T}W]\mathbb{E}[\mathbf{X}_{,k}^{b^T}W]\|_2^2, \quad (6)$$

Variables decorrelation **regularizer**

$$\min_W \sum_{j=1}^{p} \left\|\mathbb{E}[\mathbf{X}_{,j}^T\mathbf{\Sigma}_W\mathbf{X}_{,-j}] - \mathbb{E}[\mathbf{X}_{,j}^T W]\mathbb{E}[\mathbf{X}_{,-j}^T W]\right\|_2^2 \quad (7)$$

Decorrelated **W**eighted **R**egression (DWR)

$$\min_{W,\beta} \sum_{i=1}^{n} W_i \cdot (Y_i - \mathbf{X}_{i,}\beta)^2 \quad (12)$$

$$s.t \quad \sum_{j=1}^{p} \left\|\mathbf{X}_{,j}^T\mathbf{\Sigma}_W\mathbf{X}_{,-j}/n - \mathbf{X}_{,j}^T W/n \cdot \mathbf{X}_{,-j}^T W/n\right\|_2^2 < \lambda_2$$

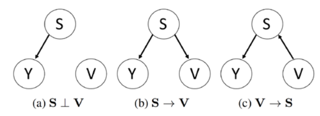
$$|\beta|_1 < \lambda_1, \ \frac{1}{n}\sum_{i=1}^{n} W_i^2 < \lambda_3,$$

$$\left(\frac{1}{n}\sum_{i=1}^{n} W_i - 1\right)^2 < \lambda_4, \ W \succeq 0,$$

EXPERIMENTS

Experiments on Synthetic Data

Generating S and V:



Generating Y:

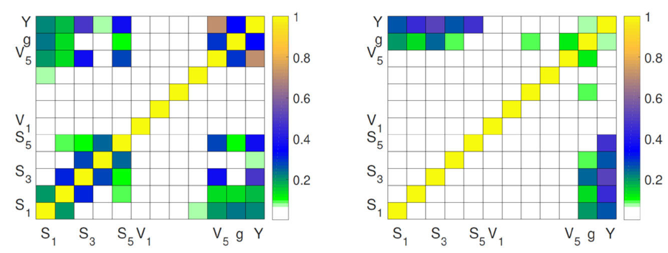
$$Y_{poly} = f(\mathbf{S}) + \varepsilon = [\mathbf{S}, \mathbf{V}] \cdot [\beta_s, \beta_v]^T + \mathbf{S}_{\cdot,1}\mathbf{S}_{\cdot,2}\mathbf{S}_{\cdot,3} + \varepsilon \quad (17)$$

$$Y_{exp} = f(\mathbf{S}) + \varepsilon = [\mathbf{S}, \mathbf{V}] \cdot [\beta_s, \beta_v]^T + e^{\mathbf{S}_{\cdot,1}\mathbf{S}_{\cdot,2}\mathbf{S}_{\cdot,3}} + \varepsilon \quad (18)$$

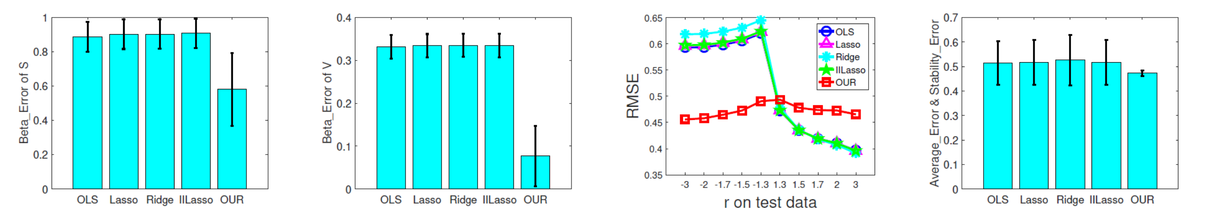
Generating Distributional Shift:

Varying P(Y|V) with bias rate r:

- r > 1: positive correlation between V and Y
- r < 1: negative correlation between V and Y



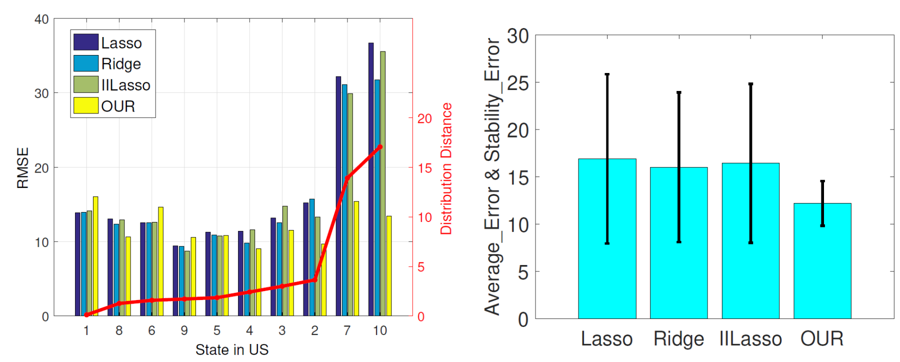
(a) On raw data (b) On the weighted data



(a) β_Error of **S**: Mean (green bar) and Variance (black line) (b) β_Error of **V**: Mean (green bar) and Variance (black line) (c) RMSE over all test environments (d) Average_Error (green bar) & Stability Error (black line)

Scenario 1: varying sample size n															
n, p, r	$n = 1000, p = 10, r = 1.7$					$n = 2000, p = 10, r = 1.7$					$n = 4000, p = 10, r = 1.7$				
	OLS	Lasso	Ridge	IILasso	Our	OLS	Lasso	Ridge	IILasso	Our	OLS	Lasso	Ridge	IILasso	Our
β_S_Error	0.892	0.907	0.907	0.912	**0.578**	0.887	0.903	0.903	0.908	**0.581**	0.906	0.921	0.921	0.926	**0.614**
β_V_Error	0.331	0.333	0.334	0.334	**0.109**	0.332	0.334	0.335	0.335	**0.077**	0.338	0.340	0.341	0.341	**0.078**
Average_Error	0.509	0.511	0.511	0.511	**0.476**	0.514	0.516	0.527	0.516	**0.473**	0.526	0.528	0.531	0.528	**0.480**
Stability_Error	0.084	0.086	0.086	0.086	**0.012**	0.090	0.092	0.103	0.092	**0.012**	0.104	0.105	0.108	0.106	**0.015**
Scenario 2: varying variables' dimension p															
n, p, r	$n = 2000, p = 10, r = 1.5$					$n = 2000, p = 20, r = 1.5$					$n = 2000, p = 40, r = 1.5$				
	OLS	Lasso	Ridge	IILasso	Our	OLS	Lasso	Ridge	IILasso	Our	OLS	Lasso	Ridge	IILasso	Our
β_S_Error	0.618	0.628	0.630	0.632	**0.409**	2.608	2.677	2.670	2.713	**1.761**	8.491	8.846	8.669	8.998	**7.800**
β_V_Error	0.243	0.245	0.246	0.245	**0.052**	0.426	0.433	0.433	0.437	**0.260**	0.661	0.684	0.673	0.694	**0.606**
Average_Error	0.486	0.487	0.487	0.487	**0.476**	0.523	0.527	0.539	0.529	**0.480**	0.532	0.540	0.537	0.543	**0.490**
Stability_Error	0.058	0.059	0.060	0.059	**0.010**	0.116	0.121	0.134	0.123	**0.014**	0.138	0.148	0.145	0.153	**0.073**
Scenario 3: varying bias rate r on training data															
n, p, r	$n = 2000, p = 10, r = 1.5$					$n = 2000, p = 10, r = 1.7$					$n = 2000, p = 10, r = 2.0$				
	OLS	Lasso	Ridge	IILasso	Our	OLS	Lasso	Ridge	IILasso	Our	OLS	Lasso	Ridge	IILasso	Our
β_S_Error	0.618	0.628	0.630	0.632	**0.409**	0.887	0.903	0.903	0.908	**0.581**	1.232	1.249	1.245	1.257	**0.651**
β_V_Error	0.243	0.245	0.246	0.245	**0.052**	0.332	0.334	0.335	0.335	**0.077**	0.441	0.444	0.443	0.445	**0.119**
Average_Error	0.486	0.487	0.487	0.487	**0.476**	0.514	0.516	0.527	0.516	**0.473**	0.568	0.571	0.571	0.571	**0.476**
Stability_Error	0.058	0.059	0.060	0.059	**0.010**	0.090	0.092	0.103	0.092	**0.012**	0.144	0.147	0.147	0.147	**0.008**

Experiments on Real World Data



(a) RMSE v.s. Distribution Distance (b) Average_Error (green bar) & Stability_Error (black line)

Air quality prediction across different States in U.S.