清華大學
Tsinghua University

数据异质性的发现和利用

2023年3月14日　**AI-Time**

刘家硕

清华大学 计算机系

刘家硕

清华大学计算机系清华大学计算机系博士生，导师为崔鹏老师，主要研究方向为分布外泛化问题、分布鲁棒优化方法与数据异质性度量和利用。曾获研究生国家奖学金、清华大学优良毕业生等荣誉。目前已在ICML、NeurIPS、ICLR等国际会议发表多篇一作论文，并长期担任ICML、CVPR、UAI等国际会议审稿人。

个人主页：ljsthu.github.io
邮箱：liujiashuo77@gmail.com
微信：jiashuo200819

Outline

- **Motivation**
- Quantifying and measuring data heterogeneity
- Exploiting heterogeneity in prediction
- Future directions

数据异质性（Data Heterogeneity）

Data Science

hard to ***define, measure, exploit***

General Data Heterogeneity

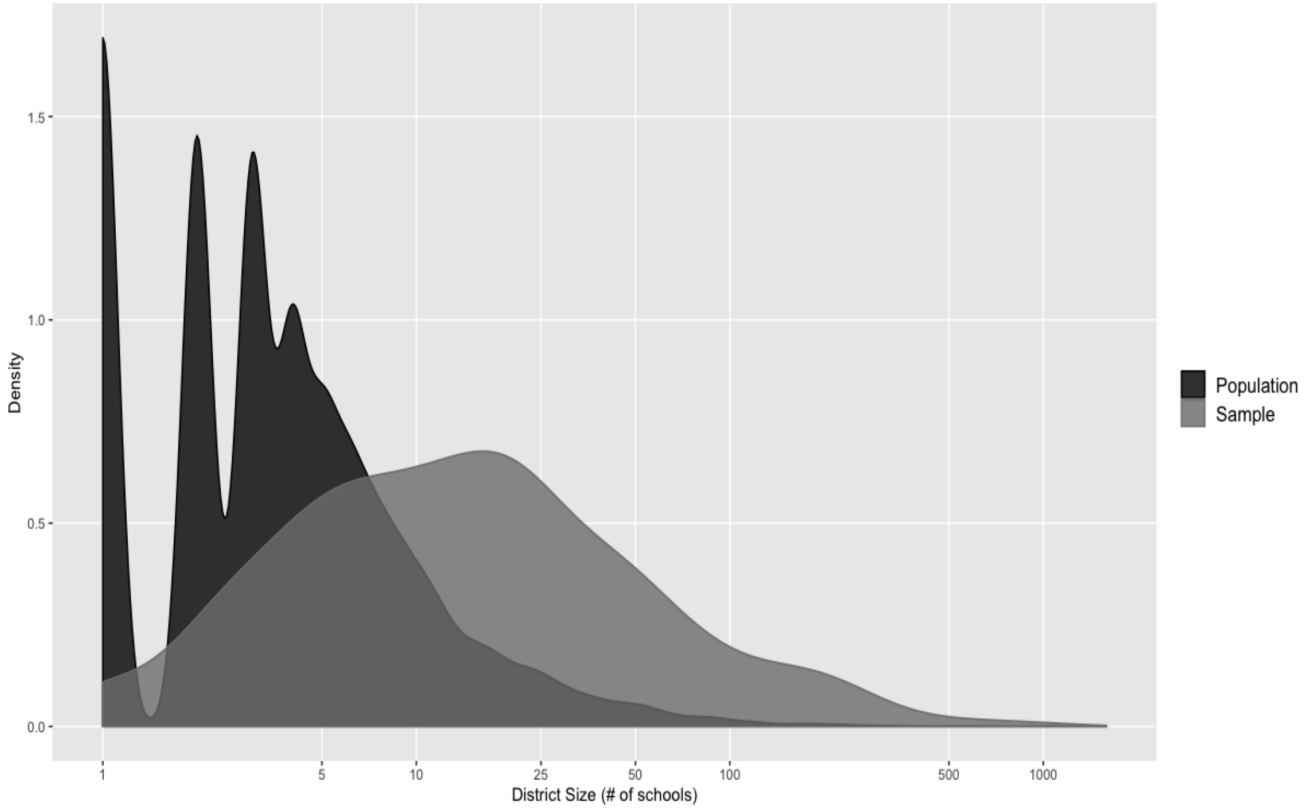
different ***prediction mechanisms***, ***label*** *distributions,* ***covariate*** *distributions,* ***data types***, ***noises***

affects ***learning, scientific findings***

Motivation 1: Scientific Findings

- even for carefully designed randomized trails, there are huge selection biases

Figure 2. Distribution of log-district size in studies versus total population



Lipton et al. The Convenience of Large Urban School Districts: A Study of Recruitment Practices in 37 Randomized Trials.
Example borrowed from Hong's talk at: https://drive.google.com/file/d/1ApBFWEkzOP39gIVMDBnXbdXXlARWXqDX/view

Motivation 1: Scientific Findings

The New York Times

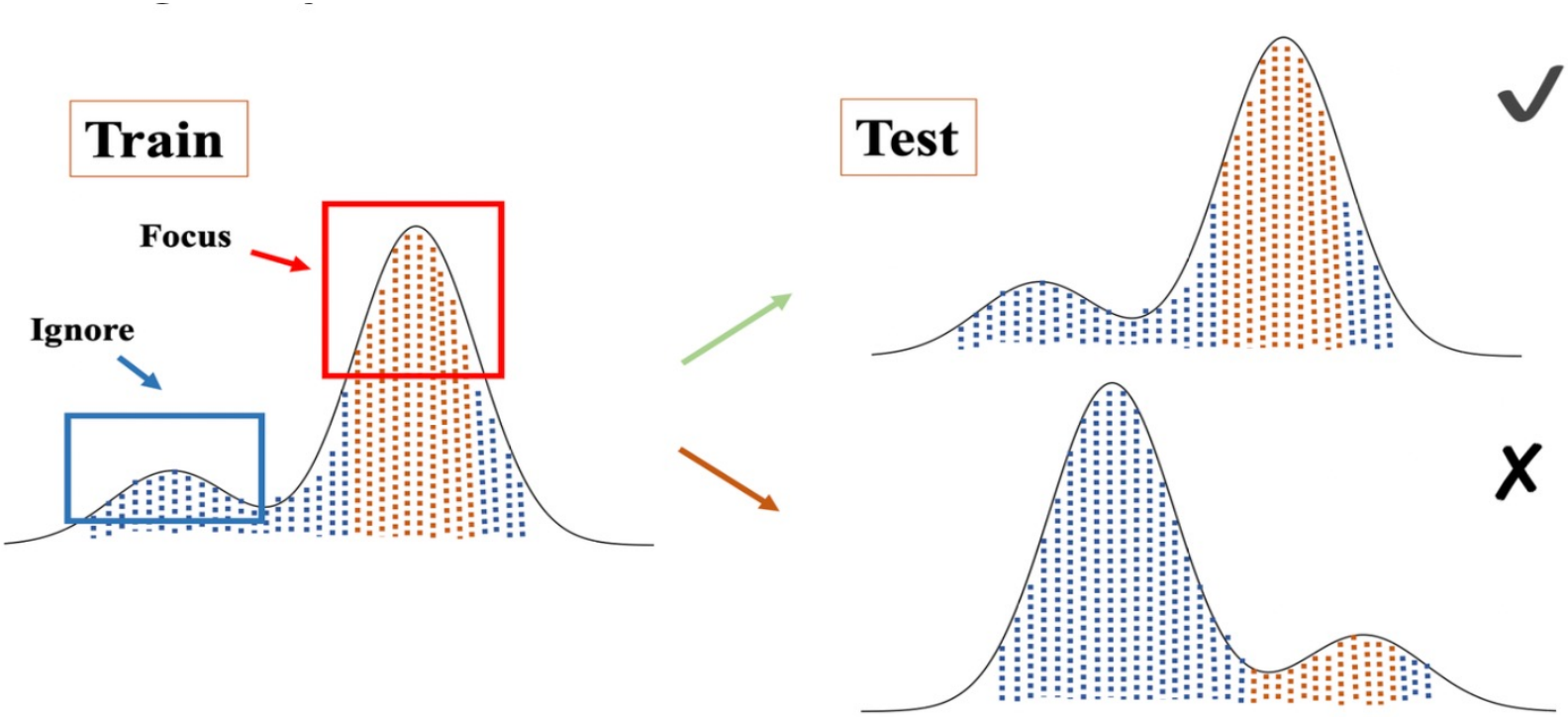
Another Benefit to Going to Museums? You May Live Longer

Researchers in Britain found that people who go to museums, the theater and the opera were less likely to die in the study period than those who didn't.

Heavy selection bias based on unobservables (wealth): decision / treatment is intimately connected with health outcomes

Example borrowed from Hong's talk at: https://drive.google.com/file/d/1ApBFWEkzOP39gIVMDBnXbdXXlARWXqDX/view

Motivation 2: Reliability & Fairness



Outline

- Motivation
- **Quantifying and measuring data heterogeneity**
- Exploiting heterogeneity in prediction
- Future directions

Predictive Heterogeneity

Data Science → Machine Learning

prediction tasks

General Data Heterogeneity

different prediction mechanisms, label distributions, covariate distributions, data types, noises

Predictive Heterogeneity

*predictive **information gain** when considering sub-populations*

*hard to **define**, **measure**, **exploit***

*easy to **define & quantify***

Tsinghua University

Definition (*informal*)

- Interaction Heterogeneity

$$\sup_{\mathcal{E}\ is\ a\ split} \mathbb{I}(Y;X|\mathcal{E}) - \mathbb{I}(Y;X)$$

Maximal additional *predictive* ***information gain*** *when dividing the whole data distribution into sub-populations*

- **Predictive Heterogeneity**

consider the model capacity and computational constraints

$$\sup_{\mathcal{E}\ is\ a\ split} \mathbb{I}_{\mathcal{V}}(Y;X|\mathcal{E}) - \mathbb{I}_{\mathcal{V}}(Y;X)$$

Maximal additional *usable* ***information gain*** *when dividing the whole data distribution into sub-populations*

Jiashuo Liu, Jiayun Wu, Renjie Pi, Renzhe Xu, Xingxuan Zhang, Bo Li and Peng Cui. Measure the Predictive Heterogeneity. ICLR 2023, The 11th International Conference on Learning Representations.

清華大學 Tsinghua University

Predictive $\mathcal{V}$-Information

Definition 1 (Predictive Family (Xu et al., 2020)). *Let* $\Omega = \{f : \mathcal{X} \cup \{\emptyset\} \to \mathcal{P}(\mathcal{Y})\}$. *We say that* $\mathcal{V} \subseteq \Omega$ *is a predictive family if it satisfies:*

$$\forall f \in \mathcal{V},\ \ \forall P \in \text{range}(f),\ \ \exists f' \in \mathcal{V},\quad s.t.\ \forall x \in \mathcal{X}, f'[x] = P, f'[\emptyset] = P. \tag{2}$$

Definition 2 (Predictive $\mathcal{V}$-information (Xu et al., 2020)). *Let* X, Y *be two random variables taking values in* $\mathcal{X} \times \mathcal{Y}$ *and* $\mathcal{V}$ *be a predictive family. The predictive* $\mathcal{V}$*-information from* X *to* Y *is defined as:*

$$\mathbb{I}_{\mathcal{V}}(X \to Y) = H_{\mathcal{V}}(Y|\emptyset) - H_{\mathcal{V}}(Y|X), \tag{3}$$

where $H_{\mathcal{V}}(Y|\emptyset)$, $H_{\mathcal{V}}(Y|X)$ *are the predictive conditional* $\mathcal{V}$*-entropy defined as:*

$$H_{\mathcal{V}}(Y|X) = \inf_{f \in \mathcal{V}} \mathbb{E}_{x,y \sim X,Y}[-\log f[x](y)]. \tag{4}$$

$$H_{\mathcal{V}}(Y|\emptyset) = \inf_{f \in \mathcal{V}} \mathbb{E}_{y \sim Y}[-\log f[\emptyset](y)]. \tag{5}$$

Notably that $f \in \mathcal{V}$ *is a function* $\mathcal{X} \cup \{\emptyset\} \to \mathcal{P}(\mathcal{Y})$, *so* $f[x] \in \mathcal{P}(\mathcal{Y})$ *is a probability measure on* $\mathcal{Y}$, *and* $f[x](y) \in \mathbb{R}$ *is the density evaluated on* $y \in \mathcal{Y}$. $H_{\mathcal{V}}(Y|\emptyset)$ *is also denoted as* $H_{\mathcal{V}}(Y)$.

$$\textit{when } \mathcal{V} = \Omega, \mathbb{I}_{\mathcal{V}}(X \to Y) = \mathbb{I}(X;Y)$$

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. In 8th International Conference on Learning Representations, ICLR 2020

Tsinghua University

Definition (*formal*)

Definition 4 (Conditional Predictive $\mathcal{V}$-information). *Let X, Y be two random variables taking values in $\mathcal{X} \times \mathcal{Y}$ and $\mathcal{E}$ be an environment variable. The conditional predictive $\mathcal{V}$-information is defined as:*

$$\mathbb{I}_{\mathcal{V}}(X \to Y|\mathcal{E}) = H_{\mathcal{V}}(Y|\emptyset, \mathcal{E}) - H_{\mathcal{V}}(Y|X, \mathcal{E}), \tag{7}$$

where $H_{\mathcal{V}}(Y|\emptyset, \mathcal{E})$ and $H_{\mathcal{V}}(Y|X, \mathcal{E})$ are defined as:

$$H_{\mathcal{V}}(Y|X, \mathcal{E}) = \mathbb{E}_{e\sim\mathcal{E}}\left[\inf_{f\in\mathcal{V}} \mathbb{E}_{x,y\sim X,Y|\mathcal{E}=e}[-\log f[x](y)]\right]. \tag{8}$$

$$H_{\mathcal{V}}(Y|\emptyset, \mathcal{E}) = \mathbb{E}_{e\sim\mathcal{E}}\left[\inf_{f\in\mathcal{V}} \mathbb{E}_{y\sim Y|\mathcal{E}=e}[-\log f[\emptyset](y)]\right]. \tag{9}$$

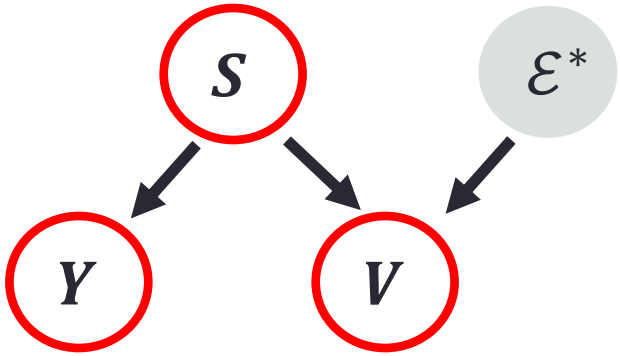
Definition 5 (Predictive Heterogeneity). *Let X, Y be random variables taking values in $\mathcal{X} \times \mathcal{Y}$ and $\mathscr{E}$ be an environment set. The predictive heterogeneity for the prediction $X \to Y$ with respect to $\mathscr{E}$ is defined as:*

$$\mathcal{H}^{\mathscr{E}}_{\mathcal{V}}(X \to Y) = \sup_{\mathcal{E}\in\mathscr{E}} \mathbb{I}_{\mathcal{V}}(X \to Y|\mathcal{E}) - \mathbb{I}_{\mathcal{V}}(X \to Y), \tag{10}$$

Maximal additional *usable* ***information gain*** *when dividing the whole data distribution into sub-populations*

Tsinghua University

Linear Example



Theorem 2 (Endogeneity with Selection Bias). *For the prediction task $X = [S, V]^T \to Y$ with a latent environment variable $\mathcal{E}^*$, the data generation process with selection bias is defined as:*

$$Y = \beta S + f(S) + \epsilon_Y, \epsilon_Y \sim \mathcal{N}(0, \sigma_Y^2); \quad V = r(\mathcal{E}^*) f(S) + \sigma(\mathcal{E}^*) \cdot \epsilon_V, \epsilon_V \sim \mathcal{N}(0, 1), \tag{12}$$

where $f : \mathbb{R} \to \mathbb{R}$ and $r, \sigma : supp(\mathcal{E}^*) \to \mathbb{R}$ are measurable functions. $\beta \in \mathbb{R}$. Assume that $\mathbb{E}[f(S)S] = 0$ and there exists $L > 1$ such that $L\sigma^2(\mathcal{E}^*) < r^2(\mathcal{E}^*)\mathbb{E}[f^2]$. For the predictive family defined in equation 11 and the environment set $\mathscr{E} = C$, the predictive heterogeneity of the prediction task $[S, V]^T \to Y$ approximates to:

$$\mathcal{H}_{\mathcal{V}}^{\mathcal{C}}(X \to Y) \approx \frac{Var(r_e)\mathbb{E}[f^2] + \mathbb{E}[\sigma^2(\mathcal{E}^*)]}{\mathbb{E}[r_e^2]\mathbb{E}[f^2] + \mathbb{E}[\sigma^2(\mathcal{E}^*)]} \mathbb{E}[f^2(S)], \text{error bounded by } \frac{1}{2}\max(\sigma_Y^2, R(r, \sigma, f)). \tag{13}$$

And $R(r(\mathcal{E}^*), \sigma(\mathcal{E}^*), f) = \mathbb{E}[(\frac{1}{\frac{r^2\mathbb{E}[f^2]}{\sigma^2}+1})^2]\mathbb{E}[f^2] + \mathbb{E}_{\mathcal{E}^*}[(\frac{1}{\frac{r}{\sigma}+\frac{\sigma}{r\mathbb{E}[f^2]}})^2] < \mathbb{E}[f^2](\frac{1}{(L+1)^2} + \frac{1}{L+2+\frac{1}{L}})$.

environment diversity

model misspecification strength

Tsinghua University

PAC Guarantees for Estimation

Theorem 3 (PAC Bound). *Consider the prediction task $X \to Y$ where X, Y are random variables taking values in $\mathcal{X} \times \mathcal{Y}$. Assume that the predictive family $\mathcal{V}$ satisfies $\forall x \in \mathcal{X}$, $\forall y \in \mathcal{Y}$,$\forall f \in \mathcal{V}$, $\log f[x](y) \in [-B, B]$ where $B > 0$. For given $K \in \mathbb{N}$, the environment set is defined as $\mathscr{E}_K = \{\mathcal{E} | \mathcal{E} \in \mathcal{C}, supp(\mathcal{E}) = K\}$. Let $\mathcal{Q}$ be the set of all probability distributions of X,Y,$\mathcal{E}$ where $\mathcal{E} \in \mathscr{E}_K$. Take an $e \in supp(\mathcal{E})$ and define a function class $\mathcal{G}_\mathcal{V} = \{g | g(x, y) = \log f[x](y) Q(\mathcal{E} = e | x, y), f \in \mathcal{V}, Q \in \mathcal{Q}\}$. Denote the Rademacher complexity of $\mathcal{G}$ with N samples by $\mathscr{R}_N(\mathcal{G})$. Then for any $\delta \in (0, 1/(2K + 2))$, with a probability over $1 - 2(K + 1)\delta$, for dataset $\mathcal{D}$ independently and identically drawn from X, Y, we have:*

$$|\mathcal{H}_\mathcal{V}^{\mathscr{E}_K}(X \to Y) - \hat{\mathcal{H}}_\mathcal{V}^{\mathscr{E}_K}(X \to Y; \mathcal{D})| \leq 4(K + 1)\mathscr{R}_{|\mathcal{D}|}(\mathcal{G}_\mathcal{V}) + 2(K + 1)B\sqrt{2 \log \frac{1}{\delta} / |\mathcal{D}|}, \tag{14}$$

where $\mathscr{R}_{|\mathcal{D}|}(\mathcal{G}_\mathcal{V}) = \mathcal{O}(|\mathcal{D}|^{-\frac{1}{2}})$ (Bartlett & Mendelson, 2002).

Tsinghua University

Algorithm to find $\mathcal{E}^*$

- Objective Function:

$$\min_{W \in \mathcal{W}_K} \mathcal{R}_\nu(W, \theta_1^*(W), \ldots, \theta_K^*(W)) = \left\{ \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{K} w_{ij} \ell_\nu(f_{\theta_j^*}(x_i), y_i) + U_\nu(W, Y_N) \right\},$$

$$\text{s.t.} \quad \theta_j^*(W) \in \arg\min_{\theta} \left\{ \mathcal{L}_\nu(W, \theta) = \sum_{i=1}^{N} w_{ij} \ell_\nu(f_\theta(x_i), y_i) \right\}, \quad \text{for } j = 1, \ldots, K,$$

- Penalties reflect the difficulty of each 'sub-task'
  - regression:

$$U_{\nu_1}(W, Y_N) = \text{Var}_{j \in [K]}(\overline{Y_N^j}) = \sum_{j=1}^{K} \left(\sum_{i=1}^{N} w_{ij} y_i \right)^2 \frac{1}{N \sum_{i=1}^{N} w_{ij}} - \left(\frac{1}{N} \sum_{i=1}^{N} y_i \right)^2$$

  - classification:

$$U_{\nu_2}(W, Y_N) = -\sum_{j=1}^{K} \frac{1}{N} \left(\sum_{i=1}^{N} w_{ij} \right) H(Y_N^j),$$

Tsinghua University

Algorithm to find $\mathcal{E}^*$

- Relationship with EM algorithm:
  - EM: learn latent variable to maximize likelihood
  - ours: learn latent variable to maximize usable predictive information

- Optimization:
  - bi-level optimization:

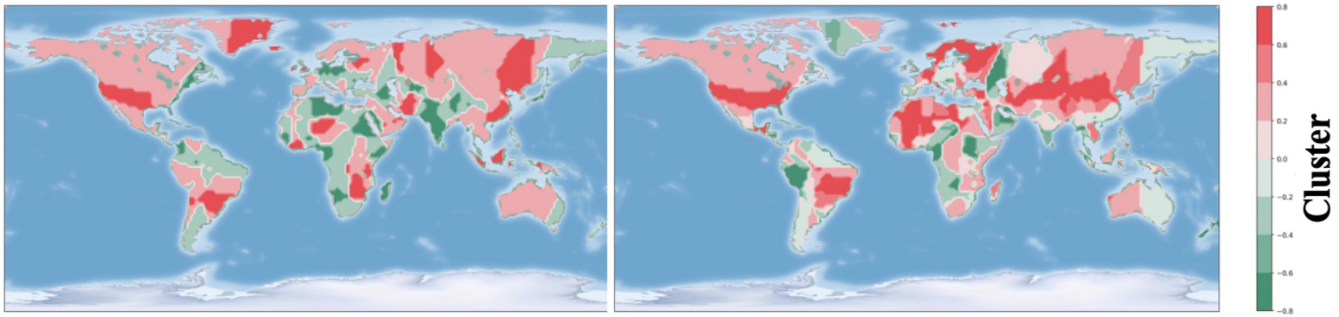
$$\nabla_W \mathcal{R} = \nabla_W U + \left[\ell(f_{\theta_j}(x_i), y_i)\right]_{i,j}^{N\times K} + \sum_{j=1}^{K} \boxed{\nabla_{\theta_j}\mathcal{R}|_{\theta_j^*} \nabla_W \theta_j^*}, \tag{19}$$

$$\text{where } \boxed{\nabla_{\theta_j}\mathcal{R}\big|_{\theta_j^*} \nabla_W \theta_j^*} \approx \nabla_{\theta_j}\mathcal{R}\big|_{\theta_j^t} \sum_{h\le t} \left[\prod_{k<h} (I - \frac{\partial^2 \mathcal{L}}{\partial \theta_j \partial \theta_j^{\mathrm{T}}}\Big|_{\theta_j^{t-k-1}})\right] \frac{\partial^2 \mathcal{L}}{\partial \theta_j \partial W^{\mathrm{T}}}\Big|_{\theta_j^{t-h-1}} \tag{20}$$

$$\approx \nabla_{\theta_j}\mathcal{R}\big|_{\theta_j^t} \frac{\partial^2 \mathcal{L}}{\partial \theta_j \partial W^{\mathrm{T}}}\Big|_{\theta_j^{t-1}} \quad \text{, for } j = 1, \ldots, K. \tag{21}$$

Tsinghua University

Application in Agriculture



(a) Division of wheat and rice cultivation areas (b) Division learned by our algorithm

Figure 1: Results on the crop yield data. We color each region according to its main crop type, and the shade represents the proportion of the main crop type after smoothing via k-means ($k = 3$).

learned sub-populations
correspond to different crop types

Tsinghua University

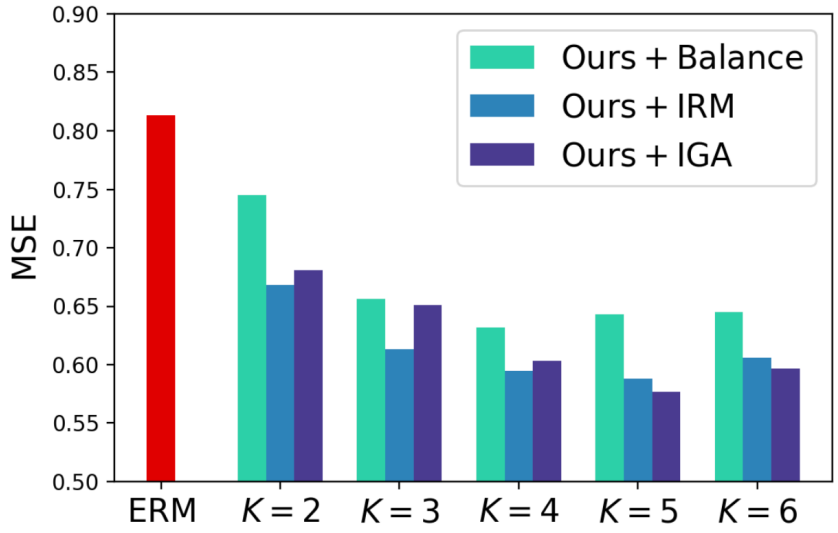
Application in Object Detection

learned sub-populations correspond to different spurious correlations

Tsinghua University

Application in OOD Generalization

Method		1. Simulated Data				2. Colored MNIST	
		Training Sub-population Error		Test Error			
		Major ($r = 1.9$)	Minor ($r = -1.9$)	$r = -2.3$	$r = -2.7$	Train Accuracy	Test Accuracy
ERM		0.255(±0.024)	0.740(±0.022)	0.738(±0.035)	0.737(±0.023)	0.998(±0.001)	0.406(±0.019)
EIIL		**0.164**(±0.014)	1.428(±0.035)	1.431(±0.061)	1.431(±0.046)	0.812(±0.006)	0.610(±0.016)
KMeans	Balance	0.231(±0.022)	0.847(±0.024)	0.846(±0.039)	0.845(±0.026)	**0.999**(±0.001)	0.328(±0.021)
	IRM	0.231(±0.022)	0.845(±0.024)	0.844(±0.039)	0.843(±0.026)	0.947(±0.004)	0.259(±0.021)
	IGA	0.235(±0.022)	0.840(±0.023)	0.839(±0.038)	0.838(±0.027)	0.997(±0.001)	0.302(±0.021)
Ours	Balance	0.403(±0.041)	**0.423**(±0.016)	**0.416**(±0.022)	**0.416**(±0.014)	0.749(±0.012)	**0.692**(±0.039)
	IRM	0.391(±0.039)	**0.432**(±0.016)	**0.430**(±0.022)	**0.430**(±0.014)	0.759(±0.014)	**0.727**(±0.047)
	IGA	0.449(±0.037)	**0.426**(±0.017)	**0.417**(±0.022)	**0.417**(±0.014)	0.759(±0.012)	**0.713**(±0.034)



Outline

- Motivation
- Quantifying and measuring data heterogeneity
- **Exploiting heterogeneity in prediction**
- Future directions

清華大學 Tsinghua University

Direction 1: Distributionally Robust Optimization

- The objective function of DRO:

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

  - $\mathcal{P}(P_{tr})$ is the distribution set defined via some distance metric as:

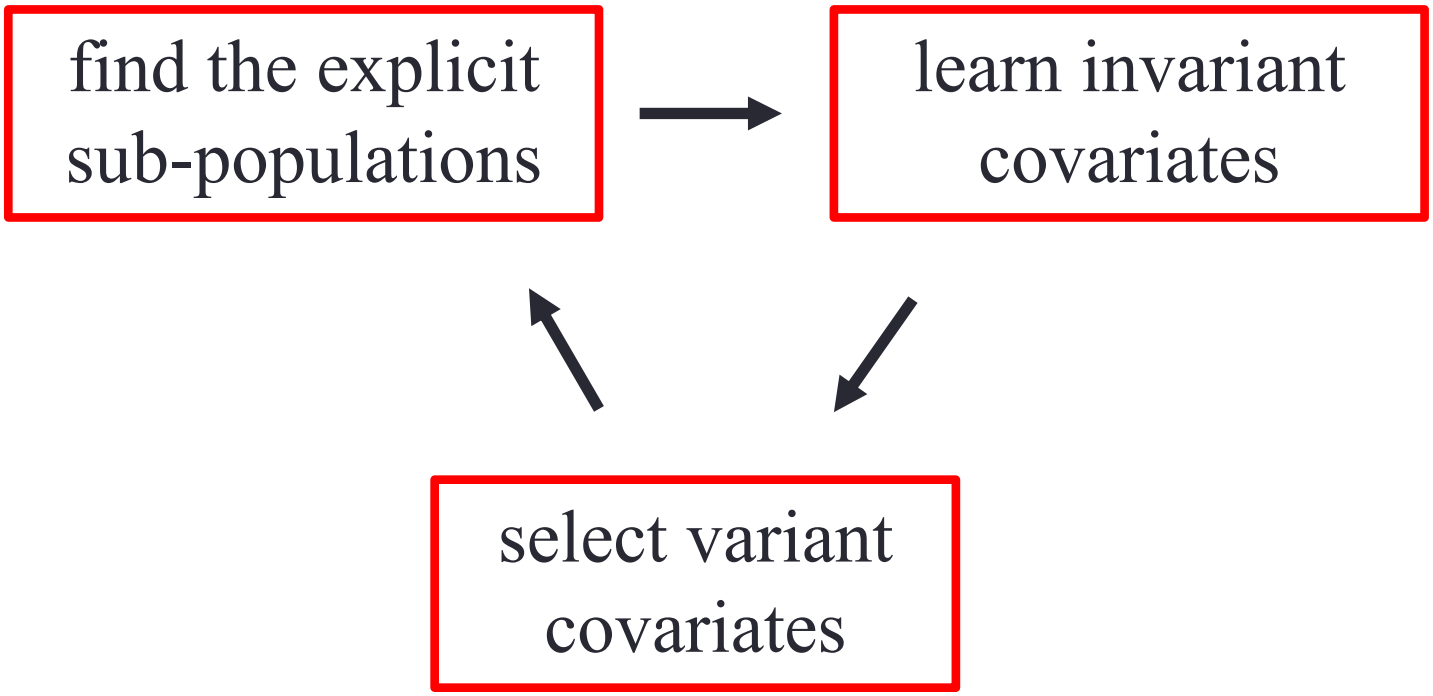
$$\mathcal{P}(P_{tr}) = \{Q : Dist(Q, P_{tr}) \leq \rho\}$$

latent

Optimize the worst-case ***sub-population*** to avoid the effects of latent heterogeneity

Tsinghua University

Direction 2: Heterogeneous Risk Minimization



Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, Zheyan Shen. Heterogeneous Risk Minimization. ICML 2021, In International Conference on Machine Learning.
Jiashuo Liu*, Zheyuan Hu*, Peng Cui, Bo Li, Zheyan Shen. Kernelized Heterogeneous Risk Minimization. NeurIPS 2021, In Neural Information Processing Systems.

清華大學
Tsinghua University

Thanks for listening!