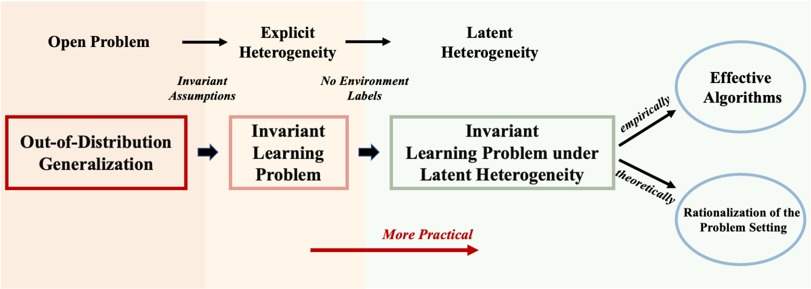
Mining the Data Heterogeneity for Out-of-Distribution Generalization

Jiashuo Liu

TrustWorthy-AI Group, CST, Tsinghua University

2022.06.08

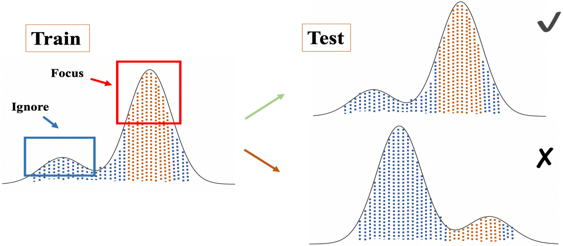
An Overview


Open Problem
Explicit Heterogeneity
Latent Heterogeneity
Invariant Assumptions
No Environment Labels
Out-of-Distribution Generalization
Invariant Learning Problem
Invariant Learning Problem under Latent Heterogeneity
empirically
theoretically
Effective Algorithms
Rationalization of the Problem Setting
More Practical

Data Heterogeneity Hurts the Generalization

Data are collected from multiple sources, which induces latent heterogeneity.

- ERM excessively focuses on the majority and ignores the minor components in data.
- Overall Good = Majority Perfect + Minority Bad
- Majority and Minority can change across different data sources/environments.
- Latent Heterogeneity renders ERM break down under distributional shifts.



Insights: We should leverage the latent heterogeneity in data and develop more rational risk minimization approach to achieve Majority Good and Minority Good, resulting in our Invariant Learning Problem under Latent Heterogeneity.

Out-of-Distribution Generalization Problem(OOD Generalization Problem)

Out-of-Distribution Generalization Problem(OOD Problem) is proposed in order to guarantee the generalization ability under distributional shifts, which can be formalized as:

$$\theta_{OOD} = \arg\min_{\theta} \max_{e \in \mathrm{supp}(\mathcal{E})} \mathcal{L}^e(\theta; X, Y) \tag{1}$$

where

- $\mathcal{E}$ is the random variable on indices of all possible environments, and for each environment $e \in \mathrm{supp}(\mathcal{E})$, the data distribution is denoted as $P^e(X, Y)$.
- The data distribution $P^e(X, Y)$ can be quite different among environments in $\mathrm{supp}(\mathcal{E})$.
- $\mathcal{L}^e(\theta; X, Y)$ denotes the risk of predictor θ on environment e, whose formulation is given by:

$$\mathcal{L}^e(\theta; X, Y) = \mathbb{E}_{X,Y \sim P^e}[\ell(\theta; X, Y)] \tag{2}$$

- OOD problem hopes to optimize the worst-case risk of all possible environments or distributions in $\mathrm{supp}(\mathcal{E})$

Invariance Assumption

To deal with the potential distributional shifts, one common assumption made in invariant learning is the **Invariance Assumption**.

Assumption (Invariance Assumption)

There exists random variable $\Phi(X)$ such that for all $e_1, e_2 \in \mathrm{supp}(\mathcal{E})$, we have

$$P^{e_1}(Y|\Phi(X)) = P^{e_2}(Y|\Phi(X)) \tag{3}$$

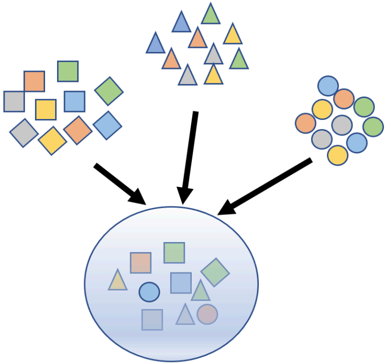
Here we make some demonstrations:

- This assumption is equivalent to $Y \perp \mathcal{E}|\Phi(X)$, indicating that the relationship between $\Phi(X)$ and Y remains invariant across environments, which is also referred to as causal relationship.
- $\Phi^*(X) = \arg\max_{\Phi: Y \perp \mathcal{E}|\Phi} \mathbb{I}(Y; \Phi(X))$ is referred to as **(Maximal) Invariant Predictors**.
- $\mathbb{E}[Y|\Phi^*(X)]$ can achieve OOD optimality[1].

[1] Koyama, Masanori, and Shoichiro Yamaguchi. "Out-of-distribution generalization with maximal invariant predictor." (2020).

Limitation 1: no environment labels

Modern datasets are frequently assembled by merging data from multiple sources without explicit source labels, which means there are not multiple environments but only one pooled dataset.



Limitation 2: quality of environments

- Heterogeneous Enough?
  - whether environments are heterogeneous to reveal the variant relationships
  - for example, all environments are the same $\Rightarrow$ useless
- Homogeneous Enough?
  - whether the invariance holds among the environments
  - for example, some environments are polluted, and only random noises Φ satisfies $Y \perp \mathcal{E}|\Phi \Rightarrow$ useless

Invariant Learning Problem under Latent Heterogeneity

Assumption (Heterogeneity Assumption)

For random variable pair (X, Φ^*) and Φ^* satisfying the Invariance Assumption, using functional representation lemma[2], there exists random variable Ψ^* such that $X = X(\Phi^*, \Psi^*)$, then we assume $P^e(Y|\Psi^*)$ can arbitrary change across environments $e \in \mathrm{supp}(\mathcal{E})$.

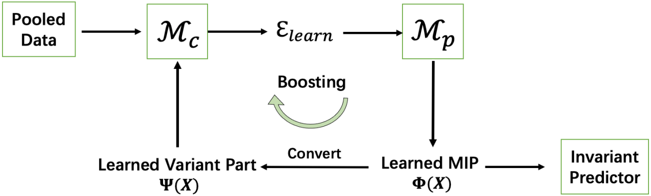
Problem (Invariant Learning Problem under Latent Heterogeneity)

Given heterogeneous dataset $D = \{D^e\}_{e \in \mathrm{supp}(\mathcal{E}_{latent})}$ without environment labels, the task is to generate environments $\mathcal{E}_{learn}$ with minimal $|\mathcal{I}_{\mathcal{E}_{learn}}|$ and learn invariant model under learned $\mathcal{E}_{learn}$ with good OOD performance.

[2] El Gamal, A. and Kim, Y.-H. Network information theory. Network Information Theory, 12 2011.

Empirical Algorithm 1: Heterogeneous Risk Minimization[3]

- This work temporarily focuses on a simple but general setting, where $X = [\Phi^*, \Psi^*]^T$ at the raw feature level.
- The HRM framework contains two modules, named **Heterogeneity Identification** module $\mathcal{M}_c$ and **Invariant Prediction** module $\mathcal{M}_p$.



- The two modules can **mutually promote** each other, meaning that the invariant prediction and the quality of $\mathcal{E}_{learn}$ can both get better and better.
- We adopt feature selection to accomplish the conversion from $\Phi(X)$ to $\Psi(X)$.
- Under our raw feature setting, we simply let $\Phi(X) = M \odot X$ and $\Psi(X) = (1 - M) \odot X$.

[3] Jiashuo Liu, Zheyuan Hu, Peng Cui *et al.* Heterogeneous Risk Minimization. *In ICML 2021*.

The Heterogeneity Identification Module $\mathcal{M}_c$

Recall that for $\mathcal{M}_c$,

$$\Psi(X) \to \mathcal{M}_c \to \mathcal{E}_{learn}$$

we implement it with a convex clustering method. Different from other clustering methods, we cluster the data according to the **relationship** between $\Psi(X)$ and Y.

- Assume the j-th cluster centre $P_{\Theta_j}(Y|\Psi)$ parameterized by Θ_j to be a Gaussian around $f_{\Theta_j}(\Psi)$ as $\mathcal{N}(f_{\Theta_j}(\Psi), \sigma^2)$:

$$h_j(\Psi, Y) = P_{\Theta_j}(Y|\Psi) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(Y - f_{\Theta_j}(\Psi))^2}{2\sigma^2}\right) \tag{4}$$

- The empirical data distribution is $\hat{P}_N = \frac{1}{N}\sum_{i=1}^{N} \delta_i(\Psi, Y)$
- The target is to find a distribution in $\mathcal{Q} = \{Q|Q = \sum_{j\in[K]} q_j h_j(\Psi, Y), \mathbf{q} \in \Delta_K\}$ to fit the empirical distribution best.
- The objective function of our heterogeneous clustering is:

$$\min_{Q\in\mathcal{Q}} D_{KL}(\hat{P}_N \| Q) \tag{5}$$

The Invariant Prediction Module $\mathcal{M}_p$

Recall that for $\mathcal{M}_p$,

$$\mathcal{E}_{learn} \to \mathcal{M}_p \to \Phi(X) = M \odot X$$

The algorithm involves two parts, invariant prediction and feature selection.

- For invariant prediction, we adopt the regularizer[4] as:

$$\mathcal{L}_p(M \odot X, Y; \theta) = \mathbb{E}_{\mathcal{E}_{tr}}[\mathcal{L}^e] + \lambda \mathrm{trace}(\mathrm{Var}_{\mathcal{E}_{tr}}(\nabla_\theta \mathcal{L}^e)) \tag{6}$$

  - Restrict the gradient across environments to be the same.
  - Only use invariant features.

- For feature selection, we adopt the continuous feature selection method that allows for continuous optimization of M:

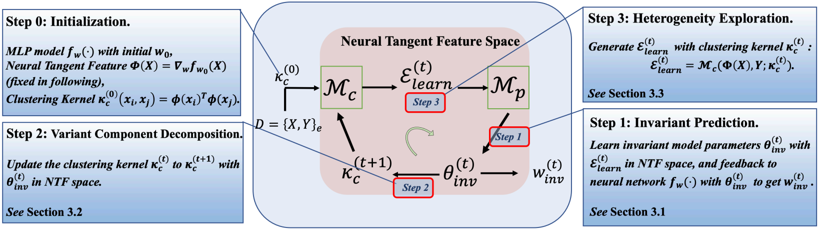
$$\mathcal{L}^e(\theta, \mu) = \mathbb{E}_{P^e}\mathbb{E}_M\left[\ell(M \odot X^e, Y^e; \theta) + \alpha \|M\|_0\right] \tag{7}$$

  - $\|M\|_0$ controls the number of selected features.
  - Conduct continuous optimization as [5].

[4] Koyama, M., & Yamaguchi, S. (2021). When is invariance useful in an Out-of-Distribution Generalization problem ?

[5] Yamada, Y., Lindenbaum, O., Negahban, S., and Kluger, Y. Feature selection using stochastic gates, in ICML2020

Empirical Algorithm 2: Kernelized Heterogeneous Risk Minimization(KerHRM[6])



- Step 0:

$$
\begin{aligned}
f_w(X) &\approx f_{w_0}(X) + \nabla_w f_{w_0}(X)^T(w - w_0) && (8)\\
&= f_{w_0}(X) + \Phi(X)^T(w - w_0) && (9)\\
&\approx f_{w_0}(X) + USV^T(w - w_0) && (10)\\
&= f_{w_0}(X) + \Psi(X)\left(V^T(w - w_0)\right) = f_{w_0}(X) + {\color{red}\Psi(X)\theta} && (11)
\end{aligned}
$$

where $\Psi(X) \in \mathbb{R}^k$ is called the reduced Neural Tangent Features(Reduced NTFs), which convert the complicated data, non-linear setting into raw feature data, linear setting.

[6] Jiashuo Liu, Zheyuan Hu, Peng Cui *et al.* Kernelized Heterogeneous Risk Minimization. *In NeurIPS 2021*.

Algorithms

- Step 1: $\mathcal{M}_p$ Invariant Learning with Reduced NTFs $\Psi(X)$[7]:

$$\theta_{inv} = \arg\min_{\theta} \sum_{e \in \mathcal{E}_{learn}} \mathcal{L}^e(\theta; \Psi, Y) + \alpha \mathrm{Var}_{\mathcal{E}_{learn}}(\nabla_\theta \mathcal{L}^e) \tag{12}$$

The obtained θ_{inv} captures the invariant component in data, which can be used to wipe out the invariant part inside data.

- Step 2: Variant Component Decomposition with θ_{inv}.
  - The initial similarity of two data points x_i and x_j:

$$\kappa_c^{(0)}(x_i, x_j) = \phi(x_i)^T \phi(x_j) = \langle U_i S, U_j S \rangle \tag{13}$$

  - Wipe out the invariant component with θ_{inv}:

$$\Psi_V^{(t+1)}(x_i) \leftarrow U_i S - \left\langle U_i S, \theta_{inv}^{(t)} \right\rangle \theta_{inv}^{(t)} / \|\theta_{inv}^{(t)}\|^2 \tag{14}$$

  - Obtain a new kernel for clustering:

$$\kappa_c^{(t+1)}(x_i, x_j) = \Psi_V^{(t+1)}(x_i)^T \Psi_V^{(t+1)}(x_j) \tag{15}$$

[7] Here we adopt the regularizer proposed in 'Masanori Koyama, Shoichiro Yamaguchi. When is invariance useful in an Out-of-Distribution Generalization problem ?'

Algorithms

- Step 3: $\mathcal{M}_c$ Heterogeneity Exploration with κ_c

  - Capture the different relationship between Ψ_V^* and Y.
  - Use $P(Y|\Psi_V)$ as the cluster centre: assume the j-th cluster centre $P_{\Theta_j}(Y|\Psi_V(X))$ to be a Gaussian around $f(\Theta_j; \Psi_V()X)$ as:

$$h_j(\Psi_V(X), Y) = P_{Theta_j}(Y|\Psi_V(X)) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-(Y - f(\Theta_j; \Psi_V(X)))^2/2]\sigma^2) \tag{16}$$

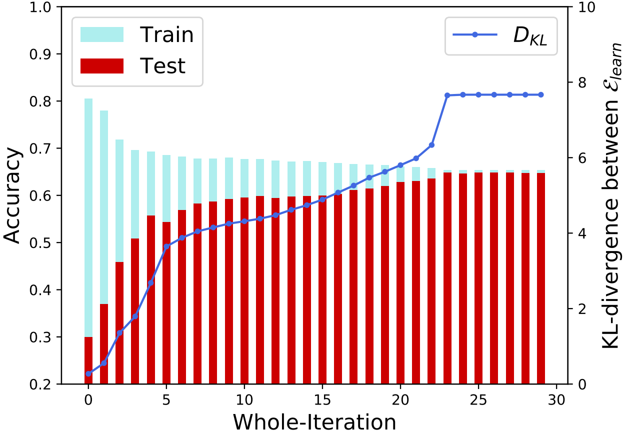
  - Propose on convex clustering algorithm, which finds a mixture distribution in distribution set $\mathcal{Q}$ defined as:

$$\mathcal{Q} = \{Q : Q = \sum_{k\in[K]} q_j h_j\} \tag{17}$$

and gives the objective function:

$$\min_{Q\in\mathcal{Q}} D_{KL}(\hat{P}_N\|Q) \Leftrightarrow \min_{\Theta,\mathbf{q}} \left\{ \mathcal{L}_c = -\frac{1}{N} \sum_{i\in[N]} \log \left[\sum_{j\in[K]} q_j h_j(\psi_V(x_i), y_i) \right] \right\} \tag{18}$$

Surprising Results



- D_{KL} denotes $KL(P_1(Y|C)\|P_2(Y|C))$

α_0-Distributional Invariance Property[8]

Theoretical drawbacks of the invariant learning under latent heterogeneity:

- **Strict invariance should be relaxed.**
  - The environment learning/splitting process is likely to **violate** the underlying strict invariance property.
  - If we still pursue the **strict invariance** that $Y \perp \mathcal{E}_{learn}|\Phi$, we may only obtain random noises.
- **'Invariance to what' should be characterized.**
  - The properties of $\mathcal{E}_{learn}$ are vague.
  - Cannot explain to what the learned model is invariant.

We propose the α_0-**Distributional Invariance** to address:

- To what extent the invariance holds: we allow for some violations on the relationship $\Phi \rightarrow Y$
- To what the invariance is considered: we only consider sub-populations larger than ratio α_0

[8] Jiashuo Liu, Jiayun Wu, *et al.* Distributionally Invariant Learning: Rationalization and Practical Algorithms.(*under review*)

α-Distributional Invariance Property

Definition (α_0-Distributional Invariance Property)

Given observed data distribution $P_0(X, Y)$ with latent heterogeneity, assume a lower bound $\alpha_0 \in (0, \frac{1}{2})$ on the sub-population proportion α and consider the set of potential minority sub-populations

$$\mathcal{P}_{\alpha_0}(P_0) = \{Q : P_0 = \alpha Q + (1-\alpha)Q_0, \text{ for } \alpha \in [\alpha_0, 1) \text{ and distribution } Q_0 \ll {}^9 P_0\} \tag{19}$$

Then a representation Φ is α_0-distributionally invariant if

$$\underbrace{\sup_{Q \in \mathcal{P}_{\alpha_0}(P_0(X,Y))}}_{\text{to what it is invariant}} \rho(Q(Y|\Phi), P_0(Y|\Phi)) \leq \underbrace{\delta}_{\text{to what extent}} \quad \text{with some } \delta > 0 \tag{20}$$

where $\rho(\cdot, \cdot)$ is some distance metric between two distributions (e.g., MMD distance, KL divergence). For simplicity, for representation Φ that is α_0-distributionally invariant, we denote it as $Y \perp^{\delta} \mathcal{E}_{\alpha_0}(P_0)|\Phi$, where $\mathcal{E}_{\alpha_0}(P_0)$ denotes the random variable on indices of distributions in $\mathcal{P}_{\alpha_0}(P_0)$.

- Based on this, we propose the Distributionally Invariant Learning (DIL) framework[10].
- We could derive the generalization error bound for our method.

[9] $Q_0 \ll P_0$ means the support of Q_0 is no larger than P_0

[10] Jiashuo Liu, Jiayun Wu, *et al.* Distributionally Invariant Learning: Rationalization and Practical Algorithms. https://arxiv.org/abs/2206.02990

Conclusion

For the **Invariant Learning Problem under Latent Heterogeneity**, we introduce

- Empirical Algorithms:
  - Heterogeneous Risk Minimization[11]
  - Kernelized Heterogeneous Risk Minimization[12]
  - Invariant Preference Learning in Recommendation[13]
- Theoretical Rationalization:
  - Distributionally Invariant Learning[14]

Some other materials for OOD Generalization:

- Anual Progress Report on Out-of-Distribution Generalization[15]
- Stable Learning and its Causal Implication[16]

[11] Jiashuo Liu, Zheyuan Hu, Peng Cui *et al.* Heterogeneous Risk Minimization. *In ICML 2021*.

[12] Jiashuo Liu, Zheyuan Hu, Peng Cui *et al.* Kernelized Heterogeneous Risk Minimization. *In NeurIPS 2021*.

[13] Zimu Wang, Yue He, Jiashuo Liu, Wenchao Zou, Philip Yu, Peng Cui. Invariant Preference Learning for General Debiasing in Recommendation. *In KDD 2022*.

[14] Jiashuo Liu, Jiayun Wu, *et al.* Distributionally Invariant Learning: Rationalization and Practical Algorithms.(*under review*)

[15] http://pengcui.thumedialab.com/papers/OOD_APR_valse2021.pdf

[16] http://pengcui.thumedialab.com/papers/Stable%20Learning-tutorial-valse2021.pdf

Contact

Jiashuo Liu

(+86) 13015155336
@liujiashuo77
liujiashuo77@gmail.com
ljsthu.github.io
https://github.com/LJSthu