Domain Generalization via Heckman-Type Selection Models

Hyungu Kahng, Hyungrok Do, Judy Zhong

Division of Biostatistics
Department of Population Health
New York University

March 22, 2023

Contact: hyungu.kahng@nyulangone.org

Table of Contents

Introduction

Domain Generalization

Domain generalization (DG) aims to incorporate knowledge from *multiple training (source) domains* into a single model that could generalize well on *unseen testing (target) domains*. Using multiple domains may offer more opportunities for a model to discover stable patterns across training domains that are also useful to testing domains.

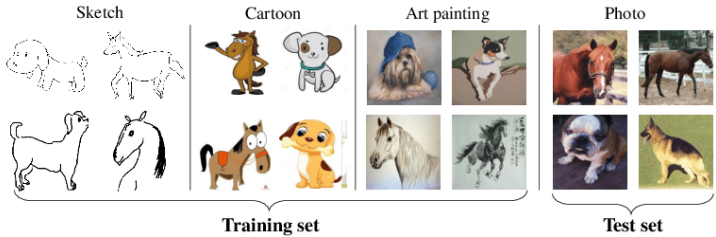


Figure: PACS dataset: PACS is an image dataset for domain generalization. It consists of four domains, namely Photo (1,670), Art Painting (2,048), Cartoon (2,344) and Sketch (3,929). Each domain contains 7 categories. Figure borrowed from [1].

Basic DG Assumptions

- Training data is collected in the form of multiple domains (= multi-source DG).
- The same outcome space $\mathcal{Y}$ is assumed across all domains (= homogeneous DG).
- Both X^{test} and Y^{test} of the testing domain is *inaccessible* during model training.
  - X^{test} is accessible during training $\rightarrow$ domain adaptation
  - Fine-tuning on X^{test} prior to testing $\rightarrow$ test-time adaptation

Examples

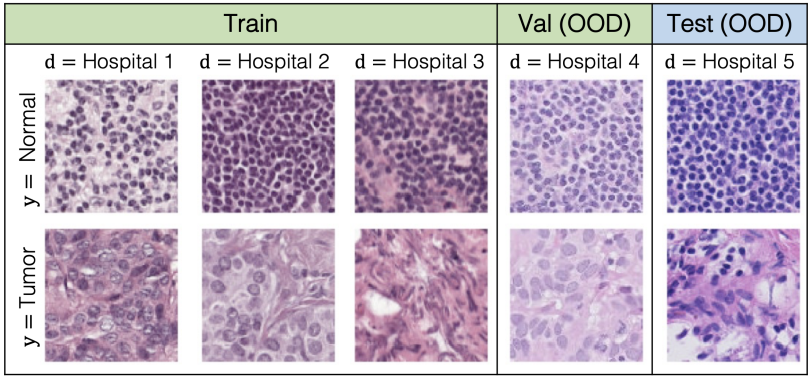


Figure: CAMELYON17: tumor classification. Figure borrowed from [2].

Examples

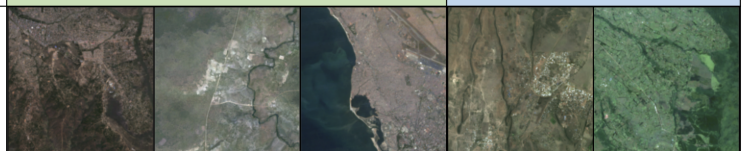
	Train			Test	
Satellite image (x)					
Country / Urban-rural (d)	Angola / urban	Angola / rural	Angola / urban	Kenya / urban	Kenya / rural
Asset index (y)	0.259	-1.106	2.347	0.827	0.130

Figure: PovertyMap: wealth index prediction. Figure borrowed from [2].

Examples

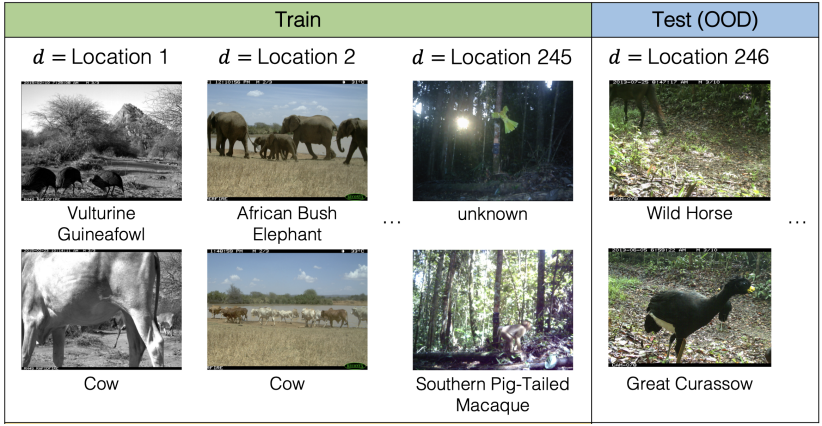


Figure: iWildCam: animal species classification. Figure borrowed from [2].

Notations

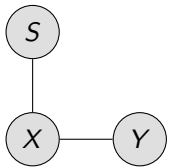
Variable	Definition
X	covariates, predictors, input
Y	outcome, target, output
S, S^k	domain indicator
K	number of training domains
L	number of hypothetically possible domains
f	outcome prediction model
$g, \mathbf{g}$	domain selection model
$\phi(\cdot)$	density of standard normal distribution
$\Phi[\cdot]$	cumulative density of standard normal distribution
$\Phi_2[\cdot, \cdot; a]$	cumulative density of bivariate standard normal with correlation a

Table: Notations

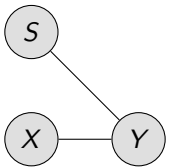
Source of Variations Across Domains

Why is $p(X, Y) \neq p(X, Y \mid S = k) \neq p(X, Y \mid S = k')$?

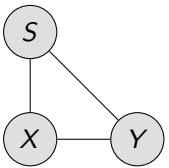
- Covariate Shift: $p(X \mid S = k) \neq p(X \mid S = k')$, while $p(Y \mid X)$ is stable.
  - data acquisition & processing protocols, patient characteristics
- Prevalence Shift: $p(Y \mid S = k) \neq p(Y \mid S = k')$



(a) Covariate Shift



(b) Prevalence Shift



(c) Both

Formal Definition of DG

Some DG papers [3, 4, 5] aim to solve the problem of learning a predictor $f \in \mathcal{F}$ that minimizes the *worst-case* risk (i.e., expected loss) over all possible domains $k \in \{1, \ldots, L\}$:

$$\min_{f \in \mathcal{F}} \max_{k \in \{1,\ldots,L\}} \mathbb{E}_{(X,Y) \sim \mathcal{P}^k_{XY}} [\ell(f(X), Y)] \tag{1}$$

while others [6] focus on minimizing the *average* risk:

$$\min_{f \in \mathcal{F}} \mathbb{E}_{k \sim \{1,\ldots,L\}} \mathbb{E}_{(X,Y) \sim \mathcal{P}^k_{XY}} [\ell(f(X), Y)] \tag{2}$$

Problem Definition

Definition 1 (Domain Generalization [6])

Domain generalization refers to as the problem of learning $f : \mathcal{X} \to \mathcal{Y}$ that has the minimum expected loss across all possible domains, which can be further summarized as the following optimization problem:

$$\min_{f \in \mathcal{F}} \sum_{k=1}^{L} \mathop{\mathbb{E}}_{(X,Y) \sim \mathcal{P}_{XY}^{k}} \Big[\ell(f(X), Y) \Big] P(S^k = 1) \tag{3}$$

where $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ is a loss function and $\mathcal{F}$ is a hypothesis set.

Problem Definition

Proposition 1 (Equivalence of Domain Generalization and Population Risk Minimization)

The problem in Definition 1 is equivalent to the risk minimization under the population distribution $\mathcal{P}_{XY}$. That is:

$$\min_{f\in\mathcal{F}} \sum_{k=1}^{L} \mathop{\mathbb{E}}_{(X,Y)\sim\mathcal{P}^k_{XY}} \Big[\ell(f(X),Y)\Big] P(S^k=1) = \min_{f\in\mathcal{F}} \mathop{\mathbb{E}}_{(X,Y)\sim\mathcal{P}_{XY}} \Big[\ell(f(X),Y)\Big] \tag{4}$$

which straightforwardly follows from the law of total expectation.

Domain Generalizable Model = Population Model

Key Research Questions

1. How can we estimate the true model $f : \mathcal{X} \to \mathcal{Y}$ that minimizes the risk under the population distribution $\mathcal{P}_{XY}$?
2. How can model the case when selection (S) is correlated with the covariates (X) and the outcome (Y)?

Related Work

Related Work: Sample Selection Bias

The *sample selection bias problem* assumes that the training data is a *non-random* sample ($S = 1$) of the population distribution $p(X, Y)$. Meanwhile, future test data is expected to be a random sample of the true population. The common goal of sample selection bias correction methods (e.g., Heckman correction [7]) is to estimate the true parameters of $f : \mathcal{X} \to \mathcal{Y}$ under sample selection bias.

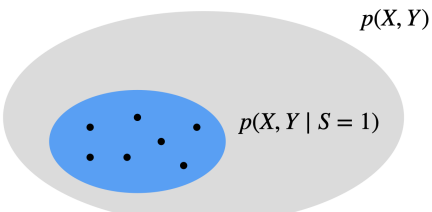


Figure: Sample selection bias

$$p_{\text{train}}(X, Y) \triangleq p(X, Y \mid S = 1)$$
$$p_{\text{test}}(X, Y) \triangleq p(X, Y) \tag{5}$$

Related Work: Heckman Correction

In the regression case, Heckman [7] assumes the following data collection process:

$$
\begin{aligned}
S &= \mathbb{I}[\tilde{S} > 0] = \mathbb{I}[g(X^{\text{sel}}) + \eta > 0] \\
Y &= \begin{cases} f(X^{\text{out}}) + \varepsilon & \text{if } S = 1 \\ \text{-} & \text{otherwise} \end{cases} \\
\begin{bmatrix} \eta \\ \varepsilon \end{bmatrix} &\sim \mathcal{N}\left(0, \begin{bmatrix} 1 & \rho\sigma \\ \rho\sigma & \sigma^2 \end{bmatrix}\right)
\end{aligned} \tag{6}
$$

where X_{sel} and X_{out} are covariates that participate in the selection (S) and outcome (Y) equations. The conditional expectation of Y given the training data is:

$$
\begin{aligned}
\mathbb{E}[Y \mid X_{\text{out}}, S = 1] &= f(X^{\text{out}}) + \mathbb{E}[\varepsilon \mid \eta > -g(X^{\text{sel}})] \\
&= f(X^{\text{out}}) + \rho\sigma \cdot \underbrace{\frac{\phi(g(X^{\text{sel}}))}{\Phi[g(X^{\text{sel}})]}}_{\text{Inverse Mills Ratio (IMR)}}
\end{aligned} \tag{7}
$$

Related Work: Heckman Correction

The two-step estimator proposed by Heckman [7]:

1. Estimate $\hat{g}$ by solving a binary classification problem distinguishing the training data ($S = 1$) and the external data ($S = 0$).

$$\hat{g} = \underset{g}{\operatorname{argmin}} - \left[\sum_{x^{\text{sel}}: S=1} \log \Phi[g(x^{\text{sel}})] + \sum_{x^{\text{sel}}: S=0} \log \Phi[-g(x^{\text{sel}})] \right] \tag{8}$$

2. Compute $\text{IMR}(\boldsymbol{x}^{\text{sel}}) = \frac{\phi[\hat{g}(\boldsymbol{x}^{\text{sel}})]}{\Phi[\hat{g}(\boldsymbol{x}^{\text{sel}})]}$, treat it as an additional covariate, and solve ordinary least squares: If $f(\boldsymbol{x}) = \boldsymbol{\beta}^\top \boldsymbol{x}$, we learn $\tilde{\boldsymbol{\beta}} = \{\boldsymbol{\beta}, \beta_{\text{IMR}}\}$, where $\beta_{\text{IMR}} := \rho\sigma$.

$$\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \sum_{(\boldsymbol{x}, y)} y - \left\{ \boldsymbol{\beta}^\top \boldsymbol{x}^{\text{out}} + \beta_{\text{IMR}} \frac{\phi(\hat{g}(\boldsymbol{x}^{\text{sel}}))}{\Phi[\hat{g}(\boldsymbol{x}^{\text{sel}})]} \right\}^2 \tag{9}$$

Related Work: Heckman Correction

We can also formulate a joint likelihood function and solve for g and f simultaneously:

$$\min_{g,f,\sigma,\rho} \sum_{x^{\text{out}},y,x^{\text{sel}}:S=1} \underbrace{-\log \Phi\left[\frac{g(x^{\text{sel}}) + \rho\frac{y-f(x^{\text{out}})}{\sigma}}{\sqrt{1-\rho^2}}\right] + \frac{1}{2}\left(\frac{y-f(x^{\text{out}})}{\sigma}\right)^2 + \log\sqrt{2\pi}\sigma}_{-\log p(y,S=1|x^{\text{sel}},x^{\text{out}})} \tag{10}$$

$$+ \sum_{x^{\text{sel}}:S=0} \underbrace{-\log \Phi[-g(x^{\text{sel}})]}_{-\log p(S=0|x^{\text{sel}})}$$

Method

Proposed Method: HeckmanDG

Motivated by Heckman's bias correction method [7], we formulate DG as a non-random sample selection problem:

- Sample selection (S) correlates with both the outcome (Y) and the covariates (X).
- $X^{\text{sel}} = X^{\text{out}}$: the selection and outcome models are a function of the same covariates.
- An independent selection model $g_k : \mathcal{X} \to [0, 1]$ is associated with each domain k.
- The selection equation g_k and the outcome equation f can take arbitrary forms (e.g., neural networks).

Proposed Method: HeckmanDG (continuous)

For continuous outcomes where $\mathcal{Y} = \mathbb{R}$, we assume probit models for the K selection equations:

$$\begin{aligned} S^k &= \mathbb{I}[\tilde{S}^k > 0] = \mathbb{I}[g_k(X) + \eta_k > 0] \\ Y &= f(X) + \varepsilon \\ \begin{bmatrix} \eta_k \\ \varepsilon \end{bmatrix} &\sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} 1 & \rho_k \sigma \\ \rho_k \sigma & \sigma^2 \end{bmatrix} \right). \end{aligned} \tag{11}$$

Given training data $\mathcal{D} = \{\boldsymbol{x}_i, \boldsymbol{s}_i, y_i\}_{i=1}^N$ where $\boldsymbol{s}_i = [s_{i1}, \ldots, s_{iK}]^\top \in \{0,1\}^K$ is a binary vector indicating domain membership, the data likelihood is defined as:

$$\mathcal{L}(\theta; \mathcal{D}) = \prod_{i=1}^{N} \prod_{k=1}^{K} \Big\{ \underbrace{p(y_i \mid s_{ik} = 1, \boldsymbol{x}_i) \cdot p(s_{ik} = 1 \mid \boldsymbol{x}_i)}_{p(y_i, s_{ik}=1 \mid \boldsymbol{x}_i)} \Big\}^{s_{ik}} \cdot p(s_{ik} = 0 \mid \boldsymbol{x}_i)^{1 - s_{ik}} \tag{12}$$

Proposed Method: HeckmanDG (continuous)

$$
\begin{aligned}
p(s_{ik} = 0 \mid \boldsymbol{x}_i) &= 1 - p(s_{ik} = 1 \mid \boldsymbol{x}_i) \\
&= 1 - p(g_k(\boldsymbol{x}_i) + \eta_k > 0) \\
&= 1 - p(\eta_k > -g_k(\boldsymbol{x}_i)) \\
&= 1 - (1 - \Phi[-g_k(\boldsymbol{x}_i)]) \\
&= \Phi[-g_k(\boldsymbol{x}_i)]
\end{aligned} \tag{13}
$$

$$
\begin{aligned}
p(y_i, s_{ik} = 1 \mid \boldsymbol{x}_i) &= p(s_{ik} = 1 \mid y_i, \boldsymbol{x}_i) \cdot p(y_i \mid \boldsymbol{x}_i) \\
&= \Phi \left[\frac{g_k(\boldsymbol{x}_i) + \rho_k \frac{y_i - f(\boldsymbol{x}_i)}{\sigma}}{\sqrt{1 - \rho_k^2}} \right] \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2} \left(\frac{f(\boldsymbol{x}_i) - y_i}{\sigma} \right)^2 \right\}
\end{aligned} \tag{14}
$$

Proposed Method: HeckmanDG (continuous)

The loss function is formulated as the negative log data likelihood:

$$\ell(\mathcal{D};\theta) = -\log \mathcal{L}(\theta;\mathcal{D}) \tag{15}$$

Proposed Method: HeckmanDG (binary)

For binary outcomes where $\mathcal{Y} = \{0, 1\}$, we assume probit models for both selection and outcome:

$$\begin{aligned} S^k &= \mathbb{I}[\tilde{S}^k > 0] = \mathbb{I}[g_k(X) + \eta_k > 0] \\ Y &= \mathbb{I}[\tilde{Y}^k > 0] = \mathbb{I}[f(X) + \varepsilon > 0] \\ \begin{bmatrix} \eta_k \\ \varepsilon \end{bmatrix} &\sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} 1 & \rho_k \\ \rho_k & 1 \end{bmatrix} \right). \end{aligned} \tag{16}$$

Given training data $\mathcal{D} = \{\boldsymbol{x}_i, \boldsymbol{s}_i, y_i\}_{i=1}^N$ where $\boldsymbol{s}_i = [s_{i1}, \ldots, s_{iK}]^\top \in \{0, 1\}^K$ is a binary vector indicating domain membership, the data likelihood is defined as:

$$\mathcal{L}(\theta; \mathcal{D}) = \prod_{i=1}^{N} \prod_{k=1}^{K} \left\{ p(y_i = 1, s_{ik} = 1 \mid \boldsymbol{x}_i)^{y_i} \cdot p(y_i = 0, s_{ik} = 1 \mid \boldsymbol{x}_i)^{(1-y_i)} \right\}^{s_{ik}} \cdot p(s_{ik} = 0 \mid \boldsymbol{x}_i)^{1-s_{ik}} \tag{17}$$

Proposed Method: HeckmanDG (binary)

$$
\begin{aligned}
p(s_{ik} = 0 \mid \boldsymbol{x}_i) &= 1 - p(s_{ik} = 1 \mid \boldsymbol{x}_i) \\
&= 1 - p(g_k(\boldsymbol{x}_i) + \eta_k > 0) \\
&= 1 - p(\eta_k > -g_k(\boldsymbol{x}_i)) \\
&= 1 - (1 - \Phi[-g_k(\boldsymbol{x}_i)]) \\
&= \Phi[-g_k(\boldsymbol{x}_i)]
\end{aligned} \tag{18}
$$

$$
\begin{aligned}
p(y_i = 1, s_{ik} = 1 \mid \boldsymbol{x}_i) &= p(f(\boldsymbol{x}_i) + \varepsilon > 0, g_k(\boldsymbol{x}_i) + \eta_k > 0) \\
&= p(\varepsilon > -f(\boldsymbol{x}_i), \eta_k > -g_k(\boldsymbol{x}_i)) \\
&= \Phi_2[f(\boldsymbol{x}_i), g_k(\boldsymbol{x}_i); \rho_k]
\end{aligned} \tag{19}
$$

$$
\begin{aligned}
p(y_i = 0, s_{ik} = 1 \mid \boldsymbol{x}_i) &= p(s_{ik} = 1 \mid \boldsymbol{x}_i) - p(y_i = 1, s_{ik} = 1 \mid \boldsymbol{x}_i) \\
&= \Phi[g_k(\boldsymbol{x}_i)] - \Phi_2[f(\boldsymbol{x}_i), g_k(\boldsymbol{x}_i); \rho_k]
\end{aligned} \tag{20}
$$

Proposed Method: HeckmanDG (binary)

The loss function is formulated as the negative log data likelihood:

$$\ell(\mathcal{D};\theta) = -\log \mathcal{L}(\theta;\mathcal{D}) \tag{21}$$

Proposed Method: HeckmanDG (multiclass)

Heckman-Type DG Estimator

Definition 2 (Heckman-Type Domain Generalization Estimator)

We formulate **HeckmanDG** as a joint learning problem of f and $\mathbf{g} = \{g_k\}_{k=1}^{K}$ with the following learning objective, where the loss function is derived from the negative log likelihood:

$$\min_{f,g,\Sigma} \sum_{i=1}^{N} \sum_{k=1}^{K} s_{ik} \Lambda(f(x_i), g_k(x_i), y_i, s_{ik}; \Sigma) - \Big\{ s_{ik} \log \Phi[g_k(x_i)] + (1 - s_{ik}) \log \Phi[-g_k(x_i)] \Big\} \tag{22}$$

where $\Phi[\cdot]$ is the cumulative distribution function fo the standard normal distribution ϕ, such that $\Phi[g_k(x_i)] = p(S^k = 1 \mid X = \boldsymbol{x}_i)$ is the selection probability w.r.t domain k, and $\Phi[-g_k(\boldsymbol{x}_i)] = p(S^k = 0 \mid X = \boldsymbol{x}_i)$. Meanwhile, $\Lambda(f(\boldsymbol{x}_i), g_k(\boldsymbol{x}_i), y_i, s_{ik}; \Sigma)$ is the conditional negative log probability of y_i given $s_{ik} = 1$, i.e., $-\log p(y_i \mid s_{ik} = 1, \boldsymbol{x}_i)$. Σ denotes the set of covariance-related parameters including σ and ρ_k.

Optimization

Motivated by Heckman's two-step estimation approach [7], we also fit the HeckmanDG estimator in a two-step manner.

1. Train $\mathbf{g} = \{g_k\}_{k=1}^K$ until convergence:

$$\hat{\mathbf{g}} = \underset{\mathbf{g}}{\operatorname{argmin}} - \sum_{i=1}^{N} \sum_{k=1}^{K} \left[s_{ik} \log \Phi[g_k(\boldsymbol{x}_i)] + (1 - s_{ik}) \log \Phi[-g_k(\boldsymbol{x}_i)] \right] \tag{23}$$

2. Train f and $\Sigma = \{\sigma, \{\rho_k\}\}$ until convergence, while keeping the weights of $\hat{\mathbf{g}}$ frozen.

$$\hat{f}, \hat{\Sigma} = \underset{f,\Sigma}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{k=1}^{K} s_{ik} \underbrace{\Lambda(f(x), \hat{g}_k(\boldsymbol{x}_i), y_i, s_{ik}; \Sigma)}_{= -\log p(y_i | s_{ik}=1, \boldsymbol{x}_i)} \tag{24}$$

Optimization

$$\underset{f,\Sigma}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{k=1}^{K} \left[-\log p(y_i \mid s_{ik} = 1, \boldsymbol{x}_i) - \log p(s_{ik} = 1 \mid \boldsymbol{x}_i) \right] \tag{25}$$

Ideas

Given $\{\hat{f}_k\}_{k=1}^K$ for K training domains:

$$\hat{y}_{\text{test}} = \sum_{k=1}^{K} w_k(\boldsymbol{x}_{\text{test}}) \cdot \hat{f}_k(\boldsymbol{x}_{\text{test}}) \tag{26}$$

where $w_k(\cdot) \in [0, 1]$ and $\sum_{k=1}^K w_k = 1$.

Ideas

$$P(Y=1 \mid S^k=1, \boldsymbol{x}) = \frac{P(Y=1, S^k=1 \mid \boldsymbol{x})}{P(S^k=1 \mid \boldsymbol{x})} = \frac{\Phi_2[f(x), g_k(x); \rho_k]}{\Phi_1[g_k(x)]} \tag{27}$$

Neural Network Architecture

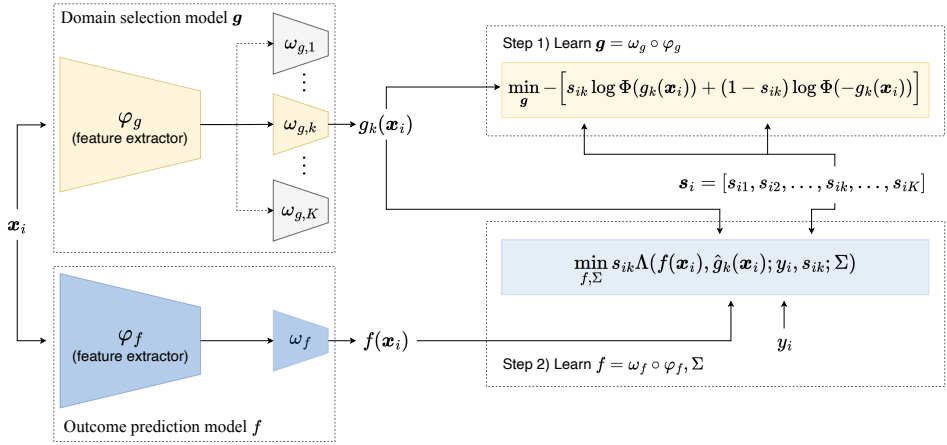


Figure: Neural network architecture of **HeckmanDG**.

Experiments

Benchmarks

Table: A summary on the four datasets from WILDS benchmark [2]. In the *'Domain'* row, the three numbers in parentheses denote the number of train, validation, and test domains.

Dataset	CAMELYON17	POVERTYMAP	IWILDCAM	RXRX1
X	$3 \times 96 \times 96$ (tissue slide)	$8 \times 224 \times 224$ (satellite image)	$3 \times 448 \times 448$ (photo)	$3 \times 256 \times 256$ (cell)
Y	2 (tumor)	continuous (asset wealth)	182 (animal species)	1139 (genetic treatments)
Training examples	302,436	10,000	129,809	40,612
Domain	5 Hospitals (3, 1, 1)	23 countries (13, 5, 5)	323 camera traps (243, 32, 48)	51 batches (33, 4, 14)
Evaluation metric	Average accuracy	Pearson (average, worst-group)	Macro F1	Average accuracy

Experimental Results

Table: CAMELYON17 (binary classification, metric: accuracy)

Method	Validation	Test
ERM (scratch)	84.9 (3.1)	70.8 (7.2)
ERM (ImageNet)	**91.3** (0.2)	84.2 (1.7)
CORAL	86.2 (1.4)	59.5 (7.7)
IRM	86.2 (1.4)	64.2 (8.1)
GroupDRO	85.5 (2.4)	68.4 (7.3)
VREx	82.3 (1.3)	71.5 (8.3)
LISA	81.8 (1.3)	77.1 (6.5)
Fish	82.5 (1.2)	79.5 (6.0)
SWAD	88.1 (1.5)	83.9 (0.9)
HeckmanDG (ours)	*90.6* (2.4)	**87.3** (2.4)

Experimental Results

Table: PovertyMap (regression, metric: Pearson correlation coefficient)

	Average		Worst Group	
Method	Validation	Test	Validation	Test
ERM	0.80 (0.04)	0.78 (0.03)	0.51 (0.06)	0.45 (0.06)
CORAL	0.80 (0.04)	0.77 (0.05)	0.52 (0.06)	0.44 (0.06)
IRM	**0.81** (0.03)	0.77 (0.05)	**0.53** (0.05)	0.43 (0.07)
GroupDRO	0.78 (0.05)	0.75 (0.07)	0.46 (0.04)	0.39 (0.06)
DANN	0.77 (0.04)	0.69 (0.04)	0.44 (0.11)	0.33 (0.10)
Fish	**0.81** (0.01)	**0.81** (0.01)	-	-
SWAD	0.78 (0.03)	0.77 (0.04)	0.48 (0.09)	0.45 (0.11)
HeckmanDG (ours)	**0.81** (0.03)	**0.81** (0.03)	**0.53** (0.06)	**0.51** (0.04)

Experimental Results

(a) RxRx1
(multiclass (1139), metric: accuracy)

Method	Validation	Test
ERM	19.4 (0.2)	29.9 (0.4)
CORAL	18.5 (0.4)	28.4 (0.3)
IRM	5.6 (0.4)	8.2 (1.1)
GroupDRO	15.2 (0.1)	23.0 (0.3)
LISA	*20.1* (0.4)	*31.9* (1.0)
Fish	7.5 (0.6)	10.1 (1.5)
SWAD	14.2 (0.5)	22.9 (0.7)
L2A-OT	17.5 (0.3)	27.8 (0.9)
HeckmanDG (ours)	**20.5** (0.7)	**32.1** (0.8)

(b) iWildCam
(multiclass (182), metric: macro-F1)

Method	Validation	Test
ERM	**37.4** (1.3)	31.0 (1.3)
CORAL	37.0 (1.2)	**32.8** (0.1)
IRM	20.2 (7.6)	15.1 (4.9)
GroupDRO	26.3 (0.2)	23.9 (2.1)
DANN	-	*31.9* (1.4)
Fish	25.8 (0.5)	24.2 (0.9)
SWAD	31.6 (0.2)	29.1 (0.1)
L2A-OT	22.8 (2.9)	18.1 (3.2)
HeckmanDG (ours)	34.5 (0.9)	*31.8* (0.3)

Discussions

Contributions

- To the best of our knowledge, we are the first paper to formulate the DG problem using a non-random sample selection framework, and to propose a Selection Guided Domain Generalization (SGDG) method under this framework (although not presented in this talk).

- We present a class of parametric SGDG (HeckmanDG) estimators applicable to continuous, binary, and multiclass outcomes.

- We demonstrate the efficacy of our method both theoretically and empirically on simulated data and four challenging benchmarks.

Future Research Directions

1. Improving the estimation of $\mathbf{g}$ to tackle overconfidence issues in $\hat{g}_k(\boldsymbol{x})$.
   - Large $\hat{g}(\boldsymbol{x})$ leads to a Inverse Mills Ratio (IMR) close to zero.
   - The fundamental nature of the data vs. using models with high capacity (NNs).
   - One way to verify: Fit a model on permuted domain labels.
   - ...

2. Heckman Correction focuses on statistical inference (of the true model), while HeckmanDG aims to increase predictive DG performance. We can be more flexible on how we obtain the final predictions (does not have to $\hat{f}$).

3. Leveraging representations $\boldsymbol{z}_g = \varphi_g(\boldsymbol{x})$ learned by the selection feature extractor.
   - What would be the relationship between $\boldsymbol{z}_g$ and $\boldsymbol{z}_f = \varphi_f(\boldsymbol{x})$?
   - Should $(\boldsymbol{z}_g, \boldsymbol{z}_f)$ have low mutual information (i.e., independent)?

Future Research Directions

- Can we share feature extractors? For instance, $\varphi(\cdot) = \varphi_g(\cdot) = \varphi_f(\cdot)$?

Future Research Directions

3. Extensions to domain adaptation (DA) & test-time adaptation (TTA).
   - Given $\mathcal{D}^{\text{test}} = \{\boldsymbol{x}_i\}_{i=1}^{N^{\text{test}}}$, we can estimate the selection model $\hat{g}_{\text{test}}$.

4. Moving beyond the assumption of jointly normal errors.
   - Assuming jointly normal error terms restricts us to use probit models, whose computational complexity increases significantly w.r.t the number of outcomes, in the multiclass case.
   - Can we formulate HeckmanDG with logit models?

5. How will HeckmanDG work on non-image (e.g., EHR, tabular) datasets?

Thank you.

References

[1] Jindong Wang et al. "Generalizing to unseen domains: A survey on domain generalization". In: *IEEE Transactions on Knowledge and Data Engineering* (2022).

[2] Pang Wei Koh et al. "Wilds: A benchmark of in-the-wild distribution shifts". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 5637–5664.

[3] Martin Arjovsky et al. "Invariant risk minimization". In: *arXiv preprint arXiv:1907.02893* (2019).

[4] Shiori Sagawa et al. "Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization". In: *arXiv preprint arXiv:1911.08731* (2019).

[5] Alexander Robey, George J Pappas, and Hamed Hassani. "Model-based domain generalization". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 20210–20229.

References

[6] Gilles Blanchard, Gyemin Lee, and Clayton Scott. "Generalizing from several related classification tasks to a new unlabeled sample". In: *Advances in neural information processing systems* 24 (2011).

[7] James J Heckman. "Sample selection bias as a specification error". In: *Econometrica: Journal of the econometric society* (1979), pp. 153–161.