Robust Inference for Federated Meta-Learning

Zijian Guo

Department of Statistics, Rutgers University

Paper and Code

Guo, Z., Li, X., Han, L., and Cai, Tianxi. (2023). Robust Inference for Federated Meta-Learning. *arXiv preprint arXiv:2301.00718*

R code available at `https://github.com/celehs/RIFL`

Guo, Z. (2020). Statistical Inference for Maximin Effects: Identifying Stable Associations across Multiple Studies. *arXiv preprint arXiv:2011.07568.*

R package `MaximinInfer`

Overview of talk

1. Multi-source Learning
2. Federated Learning under Majority Rule
3. Distributional Robust Models

Multi-source data

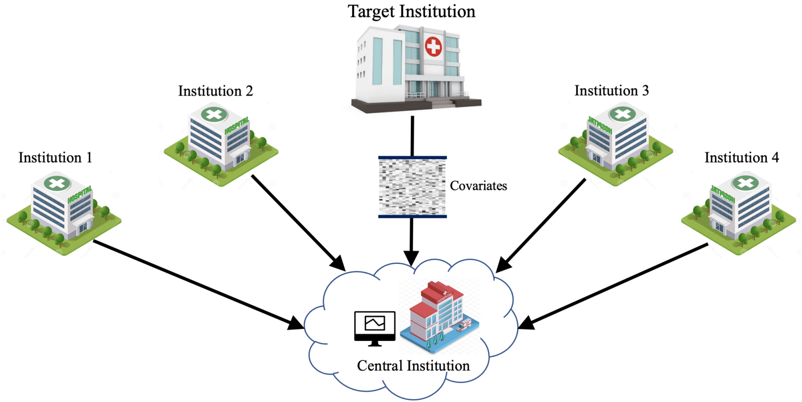


Figure: Figure credited to the Cai Lab at Harvard.

Multi-source data is everywhere

1. Data collected from different sub-populations.
2. Data collected from different healthcare centers.
3. The experiments under different environments.

Challenges

- Data heterogeneity, outlier data
- Privacy-preserving: not sharing individual-level data

Opportunity: learning a generalizable model

- Capture shared information
- Robust prediction for new environments

1. Merging the data
2. Meta-learning and federated learning
   - Inverse variance weighting
   - Federated learning: no passing of individual data
   - Shared components (Cai et al., 2021, Liu et al. 2021, Zhao et al. 2016)
   - Robust to a biased source (Wang, Wang, Miao, 2021)
   - ...
3. Causal inference with invariance principle (Peters, Bühlmann, Meinshausen, 2016, Bühlmann, 2020, Arjovsky, Bottou, Gulrajani, and Lopez-Paz, 2019)
4. **Prevailing models**: majority rule (Sorkin et al., 1998, Kerr et al., 2004, Hastie and Kameda,2005,Burgess et al.,2017, Kang et al., 2016, Maity et al., 2022)
5. **Distributionally robust model** (Meinshausen, Bühlmann, 2015, Rothenhausler et al. 2016, Sagawa et.al. 2019)

Statistical inference for generalizable models

Overview of talk

Majority rule and the prevailing model

We have access to L source populations: for $1 \leq l \leq L$,

- $\theta^{(l)} = \theta(\mathbb{P}^{(l)}) \in \mathbb{R}^d$ denotes the associated model parameter
- $Y^{(l)}_{n_l \times 1} = X^{(l)}_{n_l \times d} \theta^{(l)}_{d \times 1} + \epsilon^{(l)}_{n_l \times 1}$

For any $\theta \in \mathbb{R}^d$, we define the index set $\mathcal{V}(\theta) \subset \{1, \cdots, L\}$ as

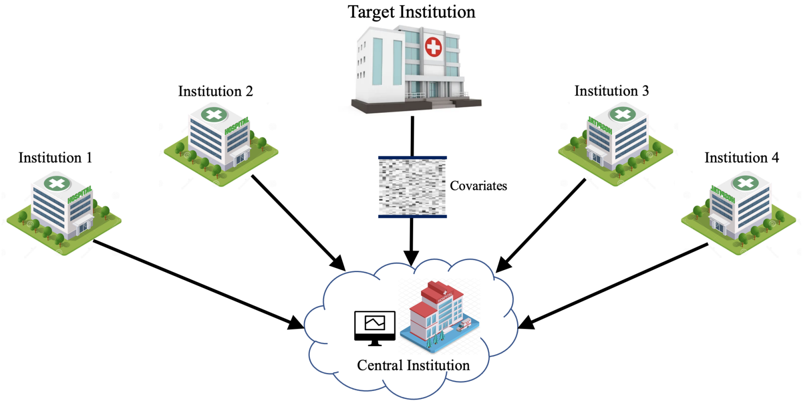
$$\mathcal{V}(\theta) := \{1 \leq l \leq L : \theta^{(l)} = \theta\}.$$

Majority Rule

There exists $\theta^* \in \mathbb{R}^d$ such that $|\mathcal{V}(\theta^*)| > L/2$.

- θ^* agrees with more than half of $\{\theta^{(l)}\}_{1 \leq l \leq L}$.
- prevailing model θ^* and prevailing set $\mathcal{V}(\theta^*)$

Multi-source data



More than half of the institutions are similar.

Research goal

Make inference for $\beta^* = g(\theta^*)$.

Examples for $\beta^* = g(\theta^*)$

1. $\beta^* = \theta_j^*$ for $1 \leq j \leq d$ or $\beta^* = x^\intercal \theta^*$ for any $x \in \mathbb{R}^d$
2. $\beta^* = \|\theta^*\|_2^2$

Site-specific estimators of $\beta^{(l)} = g(\theta^{(l)})$

For $1 \leq l \leq L$, we assume the estimators $\{\widehat{\beta}^{(l)}, \widehat{\sigma}_l\}$ satisfy

$$\frac{1}{\sigma_l}\left(\widehat{\beta}^{(l)} - \beta^{(l)}\right) \xrightarrow{d} N(0,1) \quad \text{and} \quad \frac{\widehat{\sigma}_l}{\sigma_l} \xrightarrow{p} 1,$$

where σ_l denotes the standard error of $\widehat{\beta}^{(l)}$.

Local dissimilarity measure

Define

$$\mathcal{L}_{l,k} = \beta^{(l)} - \beta^{(k)} = g(\theta^{(l)}) - g(\theta^{(k)}).$$

We estimate $\mathcal{L}_{l,k}$ by

$$\widehat{\mathcal{L}}_{l,k} = \widehat{\beta}^{(l)} - \widehat{\beta}^{(k)}.$$

We estimate the standard error of $\widehat{\mathcal{L}}_{l,k}$ by

$$\widehat{\mathrm{SE}}\left(\widehat{\mathcal{L}}_{l,k}\right) = \sqrt{\widehat{\sigma}_l^2 + \widehat{\sigma}_k^2}.$$

Global Dissimilarity Measures

For $1 \leq l < k \leq L$, $\mathcal{D}_{l,k} = \mathcal{D}(\theta^{(l)}, \theta^{(k)})$ is a dissimilarity measure.

- $\mathcal{D}(\theta^{(l)}, \theta^{(k)}) = \|\theta^{(l)} - \theta^{(k)}\|_2^2$
- $\widehat{\mathcal{D}}(\theta^{(l)}, \theta^{(k)}) = \|\widehat{\theta}^{(l)} - \widehat{\theta}^{(k)}\|_2^2$

Global dissimilarity measure

We assume that the dissimilarity estimator $\widehat{\mathcal{D}}_{l,k}$ for $1 \leq l < k \leq L$ satisfies

$$\limsup_{n\to\infty} \mathbf{P}\left(\left|\widehat{\mathcal{D}}_{l,k} - \mathcal{D}_{l,k}\right| / \widehat{\mathrm{SE}}(\widehat{\mathcal{D}}_{l,k}) \geq z_{\alpha}\right) \leq \alpha \quad \text{for} \quad 0 < \alpha < 1$$

where z_{α} denotes the upper quantile of a standard normal distribution.

Similarity Test

We compare $\theta^{(l)}$ and $\theta^{(k)}$ by defining

$$\widehat{S}_{l,k} := \max\left\{\left|\widehat{\mathcal{D}}_{l,k}/\widehat{\mathrm{SE}}(\widehat{\mathcal{D}}_{l,k})\right|, \left|\widehat{\mathcal{L}}_{l,k}/\widehat{\mathrm{SE}}(\widehat{\mathcal{L}}_{l,k})\right|\right\}.$$

For $1 \leq l < k \leq L$, we conduct the following statistical test

$$\widehat{H}_{l,k} = \mathbf{1}\left(\widehat{S}_{l,k} \leq z_{\alpha/[2L(L-1)]}\right), \quad \text{for} \quad 1 \leq l < k \leq L.$$

where $z_{\alpha/[2L(L-1)]}$ is the $\alpha/[2L(L-1)]$ upper quantile of $N(0,1)$.

- $\widehat{H}_{l,k} = 1$ denotes that the l-th and k-th sites vote for each other
- Voting matrix $\widehat{H} = \{\widehat{H}_{l,k}\}_{1 \leq l,k \leq L}$

Prevailing set estimation

We estimate the prevailing set $\mathcal{V}$ by

$$\widetilde{\mathcal{V}} = \{1 \leq l \leq L : \|\widehat{H}_{l,\cdot}\|_0 > L/2\}.$$

Define the graph $\mathcal{G}([L], \widehat{H})$ with vertices $[L] := \{1, 2, \cdots, L\}$ and

$$\widehat{H}_{l,k} = \begin{cases} 1 & \text{l-th and k-th vertices are connected} \\ 0 & \text{otherwise} \end{cases}.$$

Define $\widehat{\mathcal{V}} = \mathcal{MC}([L], \widehat{H})$ as the largest fully connected sub-graph.

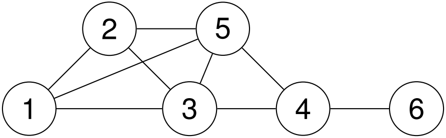


Figure: $\widetilde{\mathcal{V}} = \{1, 2, 3, 4, 5\}$ and $\widehat{\mathcal{V}} = \{1, 2, 3, 5\}$.

Post-selection Problem

Construct the inverse variance estimator with $\{\widehat{\beta}^{(l)}, \widehat{\sigma}_l\}_{l \in \widehat{\mathcal{V}}}$,

$$\widehat{\beta^*} = \frac{\sum_{l \in \widehat{\mathcal{V}}} \widehat{\beta}^{(l)} / \widehat{\sigma}_l^2}{\sum_{l \in \widehat{\mathcal{V}}} 1 / \widehat{\sigma}_l^2}.$$

We construct the $1 - \alpha$ confidence interval

$$\mathrm{CI}_{\mathrm{post}} = \left(\widehat{\beta^*} - z_{\alpha/2} \frac{1}{\sqrt{\sum_{l \in \widehat{\mathcal{V}}} 1/\widehat{\sigma}_l^2}}, \widehat{\beta^*} + z_{\alpha/2} \frac{1}{\sqrt{\sum_{l \in \widehat{\mathcal{V}}} 1/\widehat{\sigma}_l^2}} \right). \quad (1)$$

Undercover if $\widehat{\mathcal{V}} \neq \mathcal{V}(\theta^*)$.

Our proposal for CI construction

Step 1: Resampling

For $1 \leq l < k \leq L$, resample $\{\widehat{\mathcal{D}}_{l,k}^{[m]}\}_{1 \leq m \leq M}$ and $\{\widehat{\mathcal{L}}_{l,k}^{[m]}\}_{1 \leq m \leq M}$ as

$$\widehat{\mathcal{D}}_{l,k}^{[m]} \sim \mathcal{N}\left(\widehat{\mathcal{D}}_{l,k}, \widehat{\mathrm{SE}}^2(\widehat{\mathcal{D}}_{l,k})\right) \quad \text{and} \quad \widehat{\mathcal{L}}_{l,k}^{[m]} \sim \mathcal{N}\left(\widehat{\mathcal{L}}_{l,k}, \widehat{\mathrm{SE}}^2(\widehat{\mathcal{L}}_{l,k})\right)$$

For $1 \leq m \leq M$, define the resampled test statistics

$$\widehat{S}_{l,k}^{[m]} := \max\left\{\left|\widehat{\mathcal{D}}_{l,k}^{[m]}/\widehat{\mathrm{SE}}(\widehat{\mathcal{D}}_{l,k})\right|, \left|\widehat{\mathcal{L}}_{l,k}^{[m]}/\widehat{\mathrm{SE}}(\widehat{\mathcal{L}}_{l,k})\right|\right\}.$$

Construct the sampled voting matrix $\widehat{H}^{[m]} = \{\widehat{H}_{l,k}^{[m]}\}_{1 \leq l < k \leq L}$ as

$$\widehat{H}_{l,k}^{[m]} = \mathbf{1}\left(\widehat{S}_{l,k}^{[m]} \leq \rho \cdot z_{\alpha/[2L(L-1)]}\right) \quad \text{for} \quad 1 \leq l < k \leq L,$$

for $\rho = c_* (\log n/M)^{1/L(L-1)} \to 0$.

Sampling property

There exists $1 \leq m^* \leq M$ and

$$\rho = c_* \left(\log n/M\right)^{1/L(L-1)} \to 0$$

such that with probability larger than $1-\alpha$,

$$\max_{1 \leq l < k \leq L} \max \left\{ \left| \frac{\widehat{\mathcal{D}}_{l,k}^{[m*]} - \mathcal{D}_{l,k}}{\widehat{\mathrm{SE}}(\widehat{\mathcal{D}}_{l,k})} \right|, \left| \frac{\widehat{\mathcal{L}}_{l,k}^{[m*]} - \mathcal{L}_{l,k}}{\widehat{\mathrm{SE}}(\widehat{\mathcal{L}}_{l,k})} \right| \right\} \leq \rho \cdot z_{\alpha/[2L(L-1)]},$$

where $n = \min_{1 \leq l \leq L} n_l$, $c_* > 0$ is independent of n and p.

$$\widehat{H}_{l,k}^{[m^*]} = \mathbf{1}\left(\widehat{S}_{l,k}^{[m^*]} \leq \rho \cdot z_{\alpha/[2L(L-1)]}\right) \quad \text{for} \quad 1 \leq l < k \leq L.$$

Step 1 (continued): Screening

Compute the maximum clique of $\mathcal{G}([L], \widehat{H}^{[m]})$,

$$\widehat{\mathcal{V}}^{[m]} = \mathcal{MC}([L], \widehat{H}^{[m]}) \quad \text{for} \quad 1 \leq m \leq M.$$

- If $|\widehat{\mathcal{V}}^{[m]}| > L/2$, the m-th resample may be accurate.
- If $|\widehat{\mathcal{V}}^{[m]}| \leq L/2$, the m-th resample must be inaccurate.

Define the index set

$$\mathcal{M} := \left\{1 \leq m \leq M : |\widehat{\mathcal{V}}^{[m]}| > L/2\right\}.$$

Retain the resampled data for $m \in \mathcal{M}$.

Step 2: CI Construction

Define

$$\widetilde{\mathcal{V}}^{[m]} = \{1 \leq l \leq L : \|\widehat{H}^{[m]}_{l,\cdot}\|_0 > L/2\}.$$

Construct $\widehat{\beta}^{[m]} = \frac{\sum_{l \in \widetilde{\mathcal{V}}^{[m]}} \widehat{\beta}^{(l)}/\widehat{\sigma}_l^2}{\sum_{l \in \widetilde{\mathcal{V}}^{[m]}} 1/\widehat{\sigma}_l^2}$ and

$$\mathrm{CI}^{[m]} = \left(\widehat{\beta}^{[m]} - z_{\nu/2}\frac{1}{\sqrt{\sum_{l \in \widetilde{\mathcal{V}}^{[m]}} 1/\widehat{\sigma}_l^2}}, \widehat{\beta}^{[m]} + z_{\nu/2}\frac{1}{\sqrt{\sum_{l \in \widetilde{\mathcal{V}}^{[m]}} 1/\widehat{\sigma}_l^2}}\right),$$

with $\nu = \alpha - \alpha_0$. We construct the CI as

$$\mathrm{CI} = \cup_{m \in \mathcal{M}} \mathrm{CI}^{[m]}$$

Robust Inference for Federated Learning

Define the generalizability score as $\widehat{p}_l = \sum_{m \in \mathcal{M}} \mathbf{1}(l \in \widetilde{\mathcal{V}}^{[m]})/|\mathcal{M}|$

Uniformly valid CI

Theorem

If the shrinkage parameter ρ used in (17) satisfies

$$\mathrm{err}_n(M, \alpha_0) \leq \rho \cdot z_{\alpha/[2L(L-1)]} \ll \sqrt{\frac{\sum_{l \in \mathcal{V}} 1/\widehat{\sigma}_l^2}{L \cdot \sum_{l \in \mathcal{V}^c} 1/\widehat{\sigma}_l^2}}, \tag{2}$$

then our proposed confidence interval satisfies

$$\liminf_{n \to \infty} \mathbf{P}\left(g(\theta^*) \in \mathrm{CI}\right) \geq 1 - \alpha.$$

The above theorem suggests the choice of ρ as

$$\rho = \mathrm{err}_n(M, \alpha_0)/z_{\alpha/[2L(L-1)]} = c_*(\log n/M)^{\frac{1}{L(L-1)}} \to 0.$$

Oracle property

Define the oracle CI with the prior knowledge $\mathcal{V}^* = \mathcal{V}(\theta^*)$,

$$\mathrm{CI}_{\mathrm{ora}} = \left(\widehat{\beta}^{\mathrm{ora}} - z_{\alpha/2}\frac{1}{\sqrt{\sum_{l\in\mathcal{V}^*} 1/\widehat{\sigma}_l^2}}, \widehat{\beta}^{\mathrm{ora}} + z_{\alpha/2}\frac{1}{\sqrt{\sum_{l\in\mathcal{V}^*} 1/\widehat{\sigma}_l^2}}\right),$$

Theorem

If $|\mathcal{V}^*| = \lfloor L/2 \rfloor + 1$ *and any* $k \in \mathcal{V}^c$ *satisfies*

$$|\beta^{(k)} - \beta^*| \geq (2\sqrt{2\log n + 2\log M} + \rho(M) \cdot z_{\alpha/[2L(L-1)]}) \cdot \max_{l\in\mathcal{V}} \widehat{\mathrm{SE}}(\widehat{\mathcal{L}}_{l,k}),$$

then the RIFL CI satisfies $\lim_{n\to\infty}\lim_{M\to\infty} \mathbf{P}\left(\mathrm{CI} = \mathrm{CI}_{\mathrm{ora}}\right) = 1$.

Inputs for RIFL

Site-specific estimators of $\beta^{(l)} = g(\theta^{(l)})$

For $1 \leq l \leq L$, assume the estimators $\{\widehat{\beta}^{(l)}, \widehat{\sigma}_l\}$ satisfy

$$\frac{1}{\sigma_l}\left(\widehat{\beta}^{(l)} - \beta^{(l)}\right) \xrightarrow{d} N(0,1) \quad \text{and} \quad \frac{\widehat{\sigma}_l}{\sigma_l} \xrightarrow{p} 1,$$

Global dissimilarity measure

Assume that the dissimilarity estimator $\widehat{\mathcal{D}}_{l,k}$ for $1 \leq l < k \leq L$ satisfies

$$\limsup_{n\to\infty} \mathbf{P}\left(\left|\widehat{\mathcal{D}}_{l,k} - \mathcal{D}_{l,k}\right| / \widehat{\mathrm{SE}}(\widehat{\mathcal{D}}_{l,k}) \geq z_{\alpha}\right) \leq \alpha \quad \text{for} \quad 0 < \alpha < 1/4$$

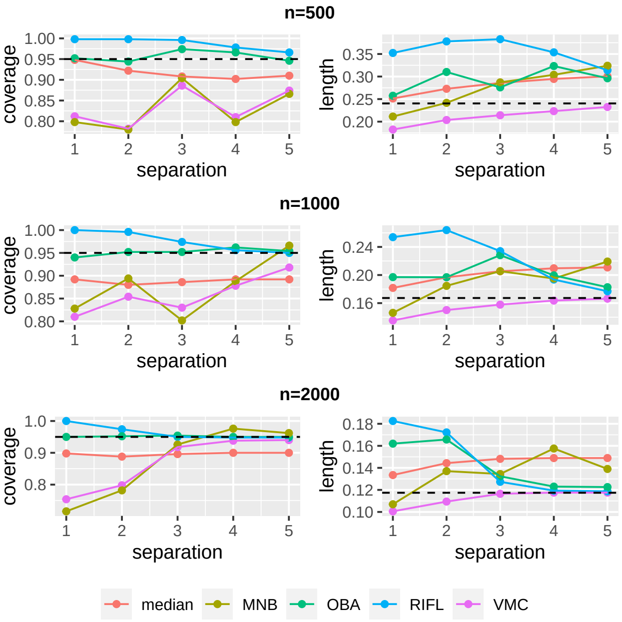
Application 1: multiple parametric models

There exists $\widehat{\theta}^{(l)}$ satisfying

$$\sqrt{n_l}\left(\widehat{\theta}^{(l)} - \theta^{(l)}\right) \xrightarrow{d} N(0, C^{(l)}).$$

Define $\mathcal{D}_{l,k} = \|\theta^{(l)} - \theta^{(k)}\|_2^2$ and $\widehat{\gamma} = \widehat{\theta}^{(l)} - \widehat{\theta}^{(k)}$.
We construct $\widehat{\mathcal{D}}_{l,k} = \|\widehat{\gamma}\|_2^2$ and estimate the standard error of $\widehat{\mathcal{D}}_{l,k}$ by

$$\widehat{\mathrm{SE}}(\widehat{\mathcal{D}}_{l,k}) = \sqrt{4\widehat{\gamma}^{\intercal}\widehat{C}^{(l)}\widehat{\gamma}/n_l + 4\widehat{\gamma}^{\intercal}\widehat{C}^{(k)}\widehat{\gamma}/n_l + 1/\min\{n_l, n_l\}}.$$


n=500
n=1000
n=2000
coverage
length
separation
median
MNB
OBA
RIFL
VMC

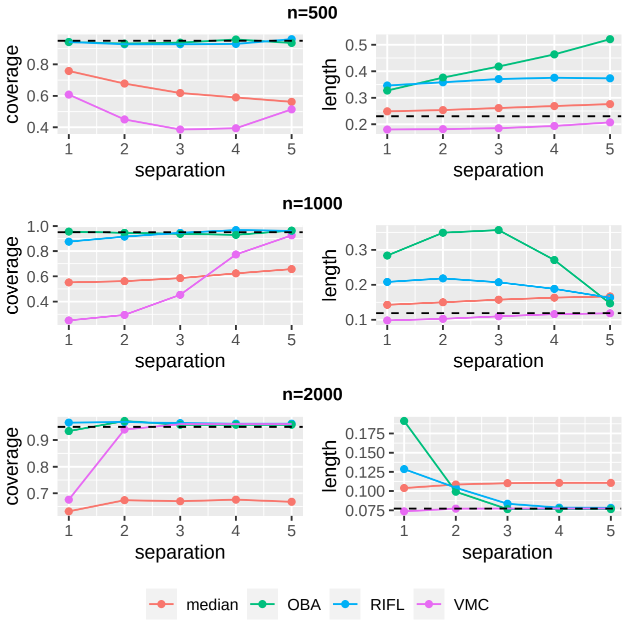
Application 2: high-dimensional prediction models

For the l-th site, we consider the GLM for the data $\{X_i^{(l)}, Y_i^{(l)}\}_{1\leq i\leq n_l}$,

$$\mathbf{E}\left(Y_i^{(l)} \mid X_i^{(l)}\right) = h([X_i^{(l)}]^\intercal \theta^{(l)}) \quad \text{for} \quad 1 \leq i \leq n_l.$$

1. Debiased estimators of $\beta^{(l)} = g(\theta^{(l)})$.
2. Construct debiased estimator $\widehat{\mathcal{D}}_{l,k}$ of $\mathcal{D}_{l,k} = \|\theta^{(l)} - \theta^{(k)}\|_2^2$.
   - Do not require the passing of individual-level data
   - Satisfy

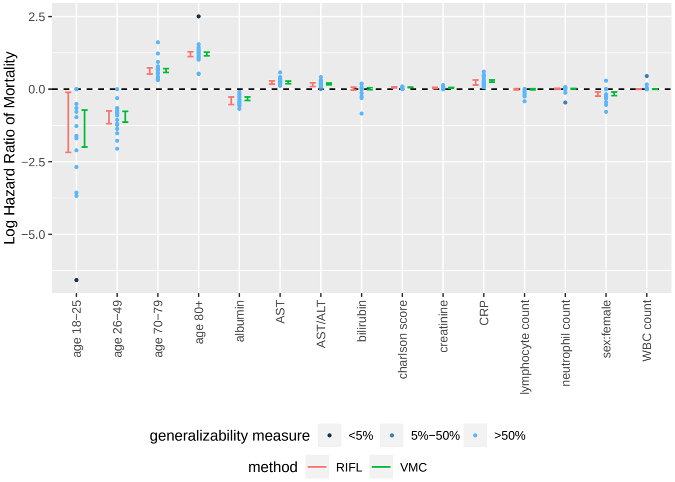
$$\limsup_{n\to\infty} \mathbf{P}\left(\left|\widehat{\mathcal{D}}_{l,k} - \mathcal{D}_{l,k}\right| / \widehat{\mathrm{SE}}(\widehat{\mathcal{D}}_{l,k}) \geq z_\alpha\right) \leq \alpha.$$


n=500
coverage
0.8
0.6
0.4
1
2
3
4
5
separation
length
0.5
0.4
0.3
0.2
1
2
3
4
5
separation
n=1000
coverage
1.0
0.8
0.6
0.4
1
2
3
4
5
separation
length
0.3
0.2
0.1
1
2
3
4
5
separation
n=2000
coverage
0.9
0.8
0.7
1
2
3
4
5
separation
length
0.175
0.150
0.125
0.100
0.075
1
2
3
4
5
separation
median
OBA
RIFL
VMC

Mortality risk for patients hospitalized with COVID-19

Consortium for Clinical Characterization of COVID-19 by EHR (4CE)

- $L = 16$ healthcare centers from 4 countries, $42,655$ patients
- $d = 15$ risk factors: age group (18-25, 26-49, 50-69, 70-79, 80+), sex, pre-admission Charlson comorbidity score, and nine laboratory test values at admission.
- For l-th healthcare center, $\theta^{(l)} \in \mathbb{R}^{15}$ denote the regression vector in a multivariate Cox model.


Log Hazard Ratio of Mortality
2.5
0.0
−2.5
−5.0
age 18−25
age 26−49
age 70−79
age 80+
albumin
AST
AST/ALT
bilirubin
charlson score
creatinine
CRP
lymphocyte count
neutrophil count
sex:female
WBC count
generalizability measure
<5%
5%−50%
>50%
method
RIFL
VMC

Overview of talk

1. Multi-source Learning
2. Federated Learning under Majority Rule
3. Distributional Robust Models

Setups

For $1 \leq l \leq L$, the training data $\{X_i^{(l)}, Y_i^{(l)}\}_{1 \leq i \leq n_l}$ follows,

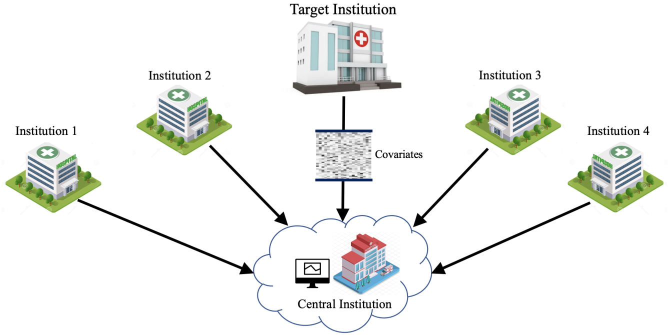
$$X_i^{(l)} \overset{\text{i.i.d.}}{\sim} \mathbb{P}_X^{(l)}, \quad Y_i^{(l)} \mid X_i^{(l)} \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{Y|X}^{(l)} \quad \text{for} \quad 1 \leq i \leq n_l.$$

We consider the target population,

$$X_i^{\mathbb{Q}} \overset{\text{i.i.d.}}{\sim} \mathbb{Q}_X, \quad Y_i^{\mathbb{Q}} \mid X_i^{\mathbb{Q}} \overset{\text{i.i.d.}}{\sim} \mathbb{Q}_{Y|X} \quad \text{for} \quad 1 \leq i \leq N_{\mathbb{Q}},$$

1. No observations of $\{Y_i^{\mathbb{Q}}\}_{1 \leq i \leq N_{\mathbb{Q}}}$.
2. Covariate shift: $\mathbb{Q}_X$ is different from $\{\mathbb{P}_X^{(l)}\}_{1 \leq l \leq L}$.
3. $\mathbb{Q}_{Y|X}$ is different from $\{\mathbb{P}_{Y|X}^{(l)}\}_{1 \leq l \leq L}$.

A class $\mathcal{C}(\mathbb{Q}_X)$ of target populations



$$\mathcal{C}(\mathbb{Q}_X) = \left\{ \mathbb{T} = (\mathbb{Q}_X, \mathbb{T}_{Y|X}) : \mathbb{T}_{Y|X} = \sum_{l=1}^{L} q_l \cdot \mathbb{P}^{(l)}_{Y|X} \quad \text{with} \quad q \in \Delta_L \right\},$$

where $\Delta_L := \left\{ q \succ 0 : \sum_{l=1}^{L} q_l = 1 \right\}$.

Group Distributionally Robust Models

If the test data $\{X_i, Y_i\} \sim \mathbb{T}$, define the reward function for β,

$$\mathbf{E}_{\mathbb{T}} Y_i^2 - \mathbf{E}_{\mathbb{T}}(Y_i - X_i^{\intercal}\beta)^2.$$

The worst-case or adversarial reward,

$$R_{\mathbb{Q}}(\beta) = \min_{\mathbb{T} \in \mathcal{C}(\mathbb{Q}_X)} \left\{ \mathbf{E}_{\mathbb{T}} Y_i^2 - \mathbf{E}_{\mathbb{T}}(Y_i - X_i^{\intercal}\beta)^2 \right\}.$$

The group distributionally robust model

$$\beta^*(\mathbb{Q}) := \arg\max_{\beta \in \mathbb{R}^p} R_{\mathbb{Q}}(\beta).$$

Sagawa, S., Koh, P. W., Hashimoto, T. B., & Liang, P. (2019). Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization.

Covariate-shift Maximin Effect

If $\mathbb{P}_X^{(l)} = \mathbb{Q}_X$, then $\beta^*(\mathbb{Q})$ is reduced to the maximin effect

$$\beta^* := \arg\max_{\beta \in \mathbb{R}^p} R(\beta) \tag{3}$$

$$R(\beta) = \min_{1 \leq l \leq L} \left\{ \mathbf{E}[Y_1^{(l)}]^2 - \mathbf{E}[Y_1^{(l)} - (X_1^{(l)})^\intercal \beta]^2 \right\}.$$

Meinshausen, N., & Bühlmann, P. (2015). Maximin effects in inhomogeneous large-scale data. The Annals of Statistics, 43(4), 1801-1830.

Inference challenge

Construct the estimator

$$\widehat{\beta^*}(\mathbb{Q}) := \arg\max_{\beta \in \mathbb{R}^p} \min_{\mathbb{T} \in \mathcal{C}(\mathbb{Q}_X)} \left\{ \widehat{\mathbf{E}}_{\mathbb{T}} Y_i^2 - \widehat{\mathbf{E}}_{\mathbb{T}} (Y_i - X_i^{\intercal} \beta)^2 \right\}.$$

$\widehat{\beta^*}(\mathbb{Q}) - \beta^*(\mathbb{Q})$ may have a non-standard distribution.

Guo, Z. (2020). Statistical Inference for Maximin Effects: Identifying Stable Associations across Multiple Studies. *arXiv preprint arXiv:2011.07568.*

Youtube link: `https://www.youtube.com/watch?v=XQ-Udp_WaMM`.

Take home message

- Challenge: heterogeneity, outlier, privacy-preserving.
- Construction of generalizable models
- Non-standard statistical inference: CLT does not hold.
- Proposals: resampling methods!

Acknowledgement

- Majority rule: Dylan Small, Hyunseung Kang
- Maximin: Peter Bühlmann, Nicolai Meinshausen, Tianxi Cai
- ReSampling: Minge Xie

Acknowledgment to NSF and NIH for funding.

Thank you!