An Adaptive Kernel Approach to Federated Learning of Heterogeneous Causal Effects

Thanh Vinh Vo [1] Arnab Bhattacharyya [1] Young Lee [2] Tze-Yun Leong [1]

[1] National University of Singapore

[2] Roche AG and Harvard University

NeurIPS 2022, New Orleans, Louisiana, USA

Motivation

Learning causal effects is important for real-life applications:

- Effect of a 'new medicine' on 'blood pressure' of patients.

Motivation

Learning causal effects is important for real-life applications:

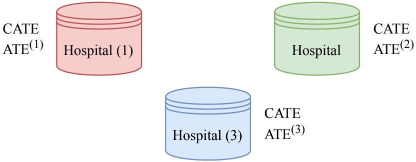
- Effect of a 'new medicine' on 'blood pressure' of patients.
- Effect of 'smoking' on 'cancer'.
- Effect of 'coronary heart disease' on 'mortality'.

Motivation

Learning causal effects is important for real-life applications:

- Effect of a 'new medicine' on 'blood pressure' of patients.
- Effect of 'smoking' on 'cancer'.
- Effect of 'coronary heart disease' on 'mortality'.

The problems:



CATE: Conditional Average Treatment Effect
ATE: Average Treatment Affect

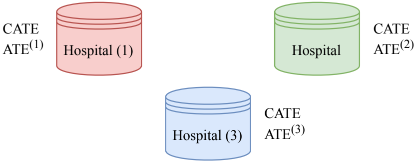
- Multiple data sources cannot be combined or shared due to privacy concern.

Motivation

Learning causal effects is important for real-life applications:

- Effect of a ‘new medicine’ on ‘blood pressure’ of patients.
- Effect of ‘smoking’ on ‘cancer’.
- Effect of ‘coronary heart disease’ on ‘mortality’.

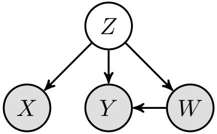
The problems:



CATE: Conditional Average Treatment Effect
ATE: Average Treatment Affect

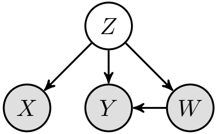
- Multiple data sources cannot be combined or shared due to privacy concern.
- Different data sources might have different data distributions.
  - Some sources with sufficient data observations might dominate the ones with fewer data observations.
  - This might lead to poor causal estimands.

Causal quantities of interest



- Y: the outcome
- W: the treatment
- Z: the latent confounder
- X: the covariate

Causal quantities of interest



- Y: the outcome
- W: the treatment
- Z: the latent confounder
- X: the covariate

We estimate

- Conditional average treatment effect (CATE):

$$\texttt{CATE}(\boldsymbol{x}) = E\big[Y \big| \mathrm{do}(W{=}1), X{=}\boldsymbol{x}\big] - E\big[Y \big| \mathrm{do}(W{=}0), X{=}\boldsymbol{x}\big],$$

- Average treatment effect (ATE):

$$\texttt{ATE} := E[\texttt{CATE}(X)].$$

The proposed method

Expectation of the outcome given intervention on w_i^{s}
$$E[y_i^{\mathsf{s}}|\mathrm{do}(w_i^{\mathsf{s}}), \mathbf{x}_i^{\mathsf{s}}]$$

The proposed method

Expectation of the outcome given intervention on w_i^{s}
$E[y_i^{\mathsf{s}}|\mathrm{do}(w_i^{\mathsf{s}}), \mathbf{x}_i^{\mathsf{s}}]$

$\downarrow$

Backdoor adjustment
$p(y_i^{\mathsf{s}}|\mathrm{do}(w_i^{\mathsf{s}}), \mathbf{x}_i^{\mathsf{s}}) = \int p(y_i^{\mathsf{s}}|w_i^{\mathsf{s}}, \mathbf{z}_i^{\mathsf{s}})\, p(\mathbf{z}_i^{\mathsf{s}}|\mathbf{x}_i^{\mathsf{s}})\, d\mathbf{z}_i^{\mathsf{s}}$

The proposed method

Expectation of the outcome given intervention on w_i^{s}
$E[y_i^{\mathsf{s}} | \mathrm{do}(w_i^{\mathsf{s}}), \mathbf{x}_i^{\mathsf{s}}]$

Backdoor adjustment
$p(y_i^{\mathsf{s}} | \mathrm{do}(w_i^{\mathsf{s}}), \mathbf{x}_i^{\mathsf{s}}) = \int p(y_i^{\mathsf{s}} | w_i^{\mathsf{s}}, \mathbf{z}_i^{\mathsf{s}})\, p(\mathbf{z}_i^{\mathsf{s}} | \mathbf{x}_i^{\mathsf{s}})\, d\mathbf{z}_i^{\mathsf{s}}$

Distributions with latent confounders
$p(\mathbf{z}_i^{\mathsf{s}} | \mathbf{x}_i^{\mathsf{s}}, y_i^{\mathsf{s}}, w_i^{\mathsf{s}})$ & $p(y_i^{\mathsf{s}} | w_i^{\mathsf{s}}, \mathbf{z}_i^{\mathsf{s}})$

Distributions without latent confounders
$p(w_i^{\mathsf{s}} | \mathbf{x}_i^{\mathsf{s}})$ & $p(y_i^{\mathsf{s}} | \mathbf{x}_i^{\mathsf{s}}, w_i^{\mathsf{s}})$

The proposed method

Expectation of the outcome given intervention on w_i^{s}

$E[y_i^{\mathsf{s}} | \mathrm{do}(w_i^{\mathsf{s}}), \mathbf{x}_i^{\mathsf{s}}]$

Backdoor adjustment

$p(y_i^{\mathsf{s}} | \mathrm{do}(w_i^{\mathsf{s}}), \mathbf{x}_i^{\mathsf{s}}) = \int p(y_i^{\mathsf{s}} | w_i^{\mathsf{s}}, \mathbf{z}_i^{\mathsf{s}})\, p(\mathbf{z}_i^{\mathsf{s}} | \mathbf{x}_i^{\mathsf{s}})\, d\mathbf{z}_i^{\mathsf{s}}$

Distributions with latent confounders

$p(\mathbf{z}_i^{\mathsf{s}} | \mathbf{x}_i^{\mathsf{s}}, y_i^{\mathsf{s}}, w_i^{\mathsf{s}})$ & $p(y_i^{\mathsf{s}} | w_i^{\mathsf{s}}, \mathbf{z}_i^{\mathsf{s}})$

Variational inference

Distributions without latent confounders

$p(w_i^{\mathsf{s}} | \mathbf{x}_i^{\mathsf{s}})$ & $p(y_i^{\mathsf{s}} | \mathbf{x}_i^{\mathsf{s}}, w_i^{\mathsf{s}})$

Maximum likelihood

The proposed method

Expectation of the outcome given intervention on w_i^{s}
$E[y_i^{\mathsf{s}} | \mathrm{do}(w_i^{\mathsf{s}}), \mathbf{x}_i^{\mathsf{s}}]$

Backdoor adjustment
$p(y_i^{\mathsf{s}} | \mathrm{do}(w_i^{\mathsf{s}}), \mathbf{x}_i^{\mathsf{s}}) = \int p(y_i^{\mathsf{s}} | w_i^{\mathsf{s}}, \mathbf{z}_i^{\mathsf{s}})\, p(\mathbf{z}_i^{\mathsf{s}} | \mathbf{x}_i^{\mathsf{s}})\, d\mathbf{z}_i^{\mathsf{s}}$

Distributions with latent confounders
$p(\mathbf{z}_i^{\mathsf{s}} | \mathbf{x}_i^{\mathsf{s}}, y_i^{\mathsf{s}}, w_i^{\mathsf{s}})$ & $p(y_i^{\mathsf{s}} | w_i^{\mathsf{s}}, \mathbf{z}_i^{\mathsf{s}})$

Variational inference

Distributions without latent confounders
$p(w_i^{\mathsf{s}} | \mathbf{x}_i^{\mathsf{s}})$ & $p(y_i^{\mathsf{s}} | \mathbf{x}_i^{\mathsf{s}}, w_i^{\mathsf{s}})$

Maximum likelihood

Augmented representer theorem estimator
$J = \widehat{\mathcal{L}} + \sum_c \gamma_c \|f_c\|^2_{\mathcal{H}_c}$.
f_c: functions that modulate the above distributions
$\mathcal{H}_c$: a reproducing kernel Hilbert space (RKHS)

The proposed method

Expectation of the outcome given intervention on w_i^{s}
$E[y_i^{\mathsf{s}} | \mathrm{do}(w_i^{\mathsf{s}}), \mathbf{x}_i^{\mathsf{s}}]$

Backdoor adjustment
$p(y_i^{\mathsf{s}} | \mathrm{do}(w_i^{\mathsf{s}}), \mathbf{x}_i^{\mathsf{s}}) = \int p(y_i^{\mathsf{s}} | w_i^{\mathsf{s}}, \mathbf{z}_i^{\mathsf{s}})\, p(\mathbf{z}_i^{\mathsf{s}} | \mathbf{x}_i^{\mathsf{s}})\, d\mathbf{z}_i^{\mathsf{s}}$

Distributions with latent confounders
$p(\mathbf{z}_i^{\mathsf{s}} | \mathbf{x}_i^{\mathsf{s}}, y_i^{\mathsf{s}}, w_i^{\mathsf{s}})$ & $p(y_i^{\mathsf{s}} | w_i^{\mathsf{s}}, \mathbf{z}_i^{\mathsf{s}})$

Variational inference

Distributions without latent confounders
$p(w_i^{\mathsf{s}} | \mathbf{x}_i^{\mathsf{s}})$ & $p(y_i^{\mathsf{s}} | \mathbf{x}_i^{\mathsf{s}}, w_i^{\mathsf{s}})$

Maximum likelihood

Augmented representer theorem estimator
$J = \widehat{\mathcal{L}} + \sum_c \gamma_c \|f_c\|^2_{\mathcal{H}_c}$.
f_c: functions that modulate the above distributions
$\mathcal{H}_c$: a reproducing kernel Hilbert space (RKHS)

Transfer kernel function
Allow dissimilar data distributions among the sources

Random Fourier feature
Enable federated learning

The proposed method

Expectation of the outcome given intervention on w_i^{s}
$E[y_i^{\mathsf{s}}|\mathrm{do}(w_i^{\mathsf{s}}), \mathbf{x}_i^{\mathsf{s}}]$

Backdoor adjustment
$p(y_i^{\mathsf{s}}|\mathrm{do}(w_i^{\mathsf{s}}), \mathbf{x}_i^{\mathsf{s}}) = \int p(y_i^{\mathsf{s}}|w_i^{\mathsf{s}}, \mathbf{z}_i^{\mathsf{s}})\, p(\mathbf{z}_i^{\mathsf{s}}|\mathbf{x}_i^{\mathsf{s}})\, d\mathbf{z}_i^{\mathsf{s}}$

Distributions with latent confounders
$p(\mathbf{z}_i^{\mathsf{s}}|\mathbf{x}_i^{\mathsf{s}}, y_i^{\mathsf{s}}, w_i^{\mathsf{s}})$ & $p(y_i^{\mathsf{s}}|w_i^{\mathsf{s}}, \mathbf{z}_i^{\mathsf{s}})$

Variational inference

Distributions without latent confounders
$p(w_i^{\mathsf{s}}|\mathbf{x}_i^{\mathsf{s}})$ & $p(y_i^{\mathsf{s}}|\mathbf{x}_i^{\mathsf{s}}, w_i^{\mathsf{s}})$

Maximum likelihood

Augmented representer theorem estimator
$J = \widehat{\mathcal{L}} + \sum_c \gamma_c \|f_c\|^2_{\mathcal{H}_c}$.
f_c: functions that modulate the above distributions
$\mathcal{H}_c$: a reproducing kernel Hilbert space (RKHS)

Transfer kernel function
Allow dissimilar data distributions among the sources

Random Fourier feature
Enable federated learning

CausalRFF: federated learning causal effects

The proposed method

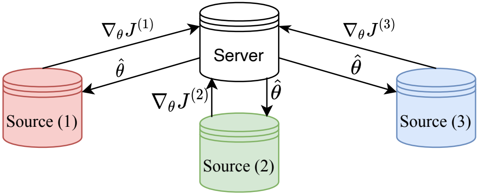
The objective function is decomposed to multiple components, each associated with a source.

$$J \simeq \sum_{\mathsf{s}\in\mathcal{S}} J^{(\mathsf{s})}, \qquad \text{where } J^{(\mathsf{s})} = \widehat{\mathcal{L}}^{(\mathsf{s})} + m^{-1}\sum_{\mathsf{v}\in\mathcal{S}} \zeta\|\theta^{\mathsf{v}}\|_2^2,$$

The proposed method

The objective function is decomposed to multiple components, each associated with a source.

$$J \simeq \sum_{\mathsf{s} \in \mathcal{S}} J^{(\mathsf{s})}, \qquad \text{where } J^{(\mathsf{s})} = \widehat{\mathcal{L}}^{(\mathsf{s})} + m^{-1} \sum_{\mathsf{v} \in \mathcal{S}} \zeta \|\theta^{\mathsf{v}}\|_2^2,$$



Repeat the following steps until convergence:

- Compute the gradients using local data in each source and send to a server.
- The sever collects all local gradients and updates the model.
- The server sends the new model to all sources.

Conclusion & future work

Key contributions:

- We proposed CausalRFF that learns causal effects without sharing raw data.
- CausalRFF minimizes information transmitted among the sources, thus enabling privacy-preserving causal inference.
- CausalRFF allows the dissimilar data distributions among the sources.
- CausalRFF is an important step towards a privacy-preserving causal learning model.
- The performance of CausalRFF is competitive with the baselines trained on combined data.

Future work:

- Preserving privacy is important: combining CausalRFF with differential privacy is an interesting problem to give statistical guarantee.

Acknowledgements & Disclaimer

This research/project is supported by the National Research Foundation Singapore and DSO National Laboratories under the AI Singapore Programme (AISG Award No: AISG2-RP-2020-016).

This work was conducted while YL was at Harvard University and the views expressed here do not necessarily reflect the position of Roche AG.