An Adaptive Kernel Approach to Federated Learning of Heterogeneous Causal Effects

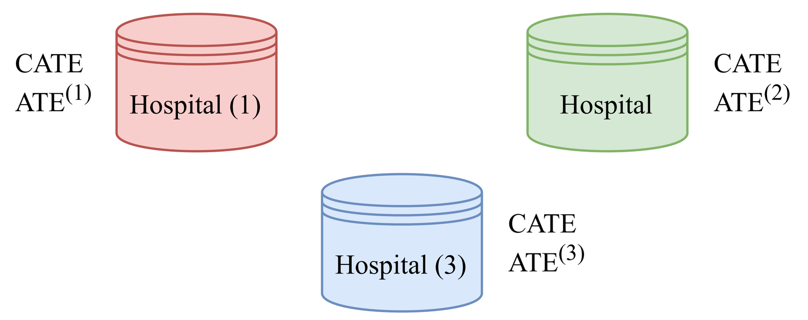
Thanh Vinh Vo [1] Arnab Bhattacharyya [1] Young Lee [2] Tze-Yun Leong [1]
[1]National University of Singapore [2]Roche AG and Harvard University
{votv, arnabb, leongty}@nus.edu.sg

Contributions

- **We propose CausalRFF that learns causal effects from multiple data sources while maintaining the sources at their local sites.**
- **CausalRFF minimizes information transmitted among the sources, thus enabling privacy-preserving causal inference.**
- **CausalRFF adaptively learn similarity of data distributions among the sources and hence reduce negative transfer.**
- **The performance of CausalRFF is competitive to the baselines trained on combined data whose sources are dissimilar.**

Motivation

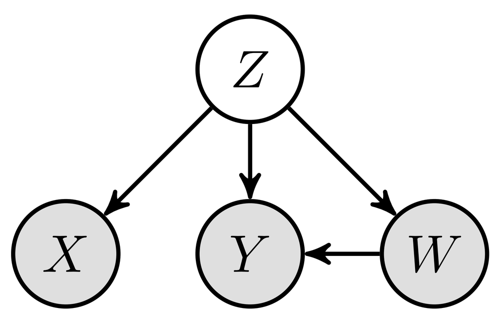
- Multiple data sources cannot be combined or shared due to privacy concern.
- Different data sources might have different data distributions.
- Some sources with sufficient data observations might dominate the ones with fewer data observations.



- For example: Patient data are private and confidential, and they are maintained in multiple hospitals.
- How to estimate causal effects from multiple sources without combining them?

Assumptions & Causal Quantities of Interest

The causal graph



- Z: latent confounder
- Y: the outcome
- Y: the outcome
- X: covariate/proxy variable

(1). *Consistency*: $W = w \implies Y(w) = Y$.
(2). *No interference* + *Positivity*.
(3). Individuals in all sources have the same set of *common* covariates.
(4). Any individual can be uniquely identified across different sources.

We estimate the following quantities:

- **Conditional average treatment effect (CATE):**

$$\tau(\boldsymbol{x}) := E\Big[Y|\mathrm{do}(W{=}1), X{=}\boldsymbol{x}\Big] - E\Big[Y|\mathrm{do}(W{=}0), X{=}\boldsymbol{x}\Big].$$

- **Average treatment effect (ATE):**

$$\tau := E[\tau(X)].$$

The Proposed Method

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

The interventional distribution of the outcome
$p(y_i^s|\mathrm{do}(w_i^s), \mathbf{x}_i^s) = \int p(y_i^s|w_i^s, \mathbf{z}_i^s)\, p(\mathbf{z}_i^s|\mathbf{x}_i^s)\, d\mathbf{z}_i^s$

The conditional distributions to be estimated
$p(\mathbf{z}_i^s|\mathbf{x}_i^s, y_i^s, w_i^s),\ p(y_i^s|w_i^s, \mathbf{z}_i^s),\ p(w_i^s|\mathbf{x}_i^s)$ & $p(y_i^s|\mathbf{x}_i^s, w_i^s)$

Augmented representer theorem estimator

Transfer kernel function
Allow dissimilar data distributions among the sources.

Random Fourier feature: – enable federated training
$k(\boldsymbol{u}, \boldsymbol{u}') \simeq \phi(\boldsymbol{u})^\top \phi(\boldsymbol{u}')$,
$\phi(\boldsymbol{u}) = B^{-\frac{1}{2}}[\cos(\boldsymbol{\omega}_1^\top \boldsymbol{u}),..., \cos(\boldsymbol{\omega}_B^\top \boldsymbol{u}), \sin(\boldsymbol{\omega}_1^\top \boldsymbol{u}),..., \sin(\boldsymbol{\omega}_B^\top \boldsymbol{u})]^\top$,
where $\boldsymbol{\omega}_b\}_{b=1}^B$ are drawn i.i.d from spectral distribution of the kernels.

Federated inference:

Repeat the following steps until convergence:

(1). Compute gradients using local data and send to a server.
(2). The sever collects all *local* gradients and updates the model.
(3). The server sends the new model to all sources.

Minimax lower bound: for learning distributions with latent confounders

$$\inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P\Big[\|\hat{\theta} - \theta(P)\|_2\Big] \geq \frac{\sqrt{m(d_x+3)}\log(2\sqrt{m})}{64\sqrt{B}\sum_{s\in\mathcal{S}} n_s\Big(1 + \sum_{v\in\mathcal{S}_{\setminus s}} \lambda^{s,v}\Big)^2}.$$

Lower bound for the worst case of the best estimator. The bound shows how the sources are incorporated through the transfer factors $\lambda^{s,v}$.

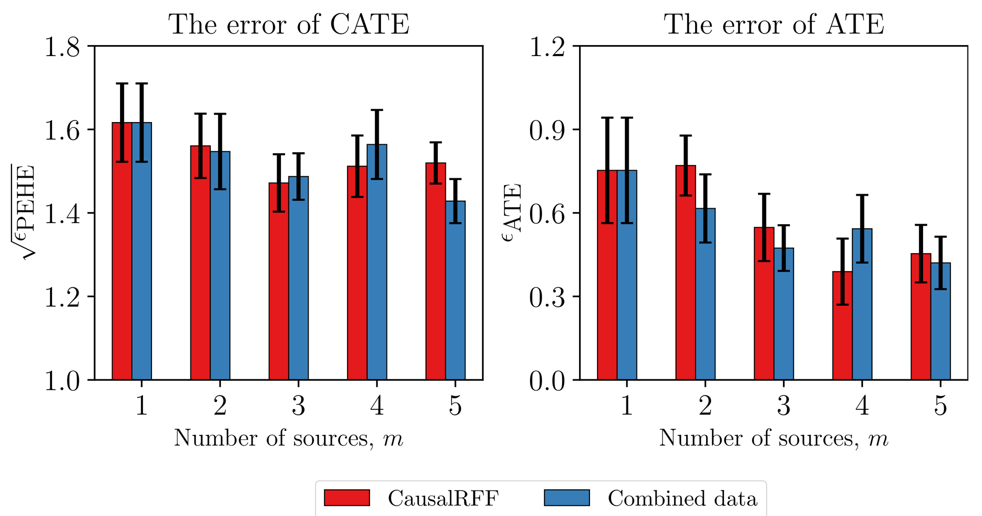
Experiments

Synthetic data

- The data are simulated with a ground truth causal model.
- We use a factor Δ to control for dissimilar data distributions.

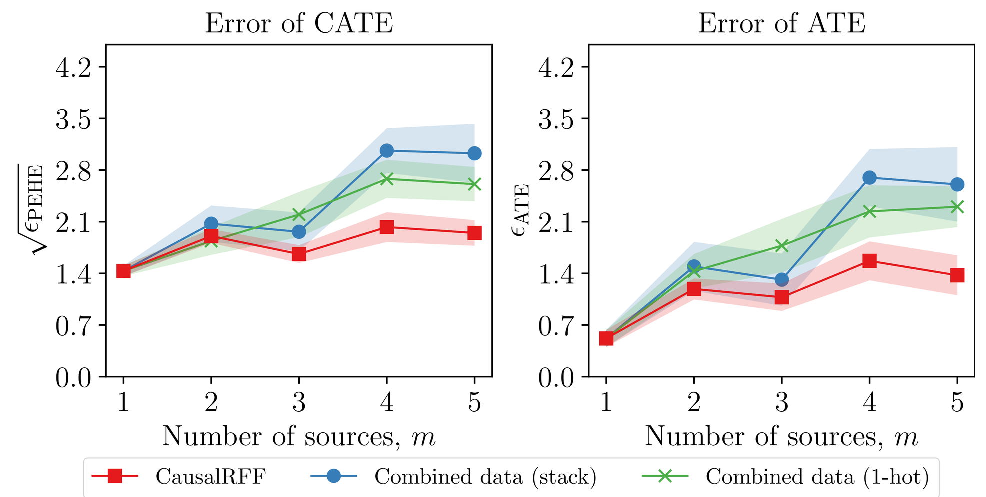
Analysis 1: The sources have the same data distribution ($\Delta = 0$).

Result: The errors of CausalRFF are as low as those of training on combined data. This result verifies the efficacy of CausalRFF for federated learning.



Analysis 2: The sources have different data distributions ($\Delta \neq 0$).

Result: The errors of CausalRFF are lower than those of training on combined data. This verifies the importance of CausalRFF when the sources have different data distributions.



IHDP dataset

- This dataset is from a study on the impact of specialist visits on the cognitive development of children.
- Treatment/control group are children with/without specialist visit.
- The dataset has 747 entries with 25 covariates, it is divided to 3 sources.

We compare with the recent baselines trained on combined data (cb)

Method	The error of CATE ($\sqrt{\epsilon_{PEHE}}$)			The error of ATE (ϵ_{ATE})		
	1 source	2 sources	3 sources	1 source	2 sources	3 sources
$BART_{cb}$	2.2±.22	2.1±.26	2.1±.25	1.0±.16	0.8±.20	0.7±.17
X-$Learner_{cb}$	1.9±.21	1.9±.21	1.8±.18	**0.5±.21**	**0.5±.18**	**0.4±.11**
R-$Learner_{cb}$	2.8±.31	2.6±.23	2.6±.17	1.6±.25	1.6±.26	1.6±.19
$OthoRF_{cb}$	2.8±.16	2.1±.14	1.9±.14	**0.8±.15**	**0.6±.10**	**0.6±.10**
$TARNet_{cb}$	3.5±.59	2.7±.12	2.5±.15	1.6±.61	**0.7±.12**	**0.6±.17**
CFR-$wass_{cb}$	2.2±.15	2.1±.22	2.1±.23	**0.7±.23**	**0.6±.18**	**0.6±.16**
CFR-mmd_{cb}	2.7±.19	2.3±.26	2.2±.10	0.9±.30	**0.7±.17**	**0.5±.17**
$CEVAE_{cb}$	**1.8±.22**	2.0±.11	**1.7±.12**	**0.5±.14**	1.4±.07	0.9±.07
FedCI	**1.6±.10**	**1.6±.12**	**1.7±.09**	**0.5±.10**	**0.5±.24**	**0.5±.09**
CausalRFF	**1.7±.34**	**1.4±.33**	**1.2±.18**	**0.7±.14**	**0.7±.17**	**0.5±.16**

Result: Causal is among top-3 performance. Importantly, it preserves privacy under federated setting while the other baselines violate this constraint.

Conclusion & Future Work

- We proposed CausalRFF that learns causal effects without sharing raw data.
- CausalRFF is an important step towards a privacy-preserving causal learning model.
- Future research direction: Combining CausalRFF with differential privacy to give a stronger privacy guarantee.

Acknowledgement.

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).

AB was supported by an NRF Fellowship for AI grant (NRFFAI1-2019-0002) and an Amazon Research Award.

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