Adaptive Multi-Source Causal Inference from Observational Data

Thanh Vinh Vo [1] Pengfei Wei [2] Trong Nghia Hoang [3] Tze-Yun Leong [1]

[1] National University of Singapore
[2] AI Lab Speech & Audio Bytedance, Singapore
[3] Washington State University

CIKM 2022, Atlanta, Georgia, USA

Outline

1. Motivation
2. Causal quantities of interest
3. AdaTRANS: the proposed method
4. Experimental evaluation
5. Summary: limitations and future work

Motivation

Why do we need causal inference?

- Effect of a ‘new medicine’ on ‘blood pressure’ of patients.

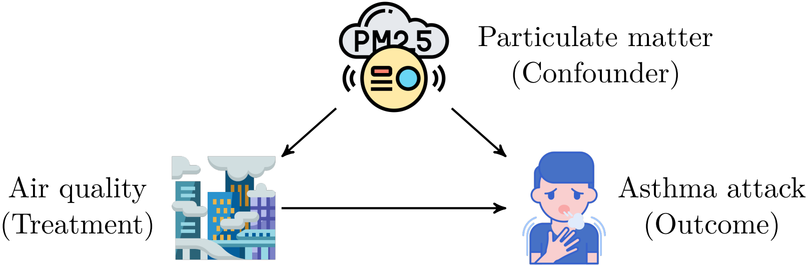
Motivation

Why do we need causal inference?

- Effect of a ‘new medicine’ on ‘blood pressure’ of patients.
- Effect of ‘air quality’ on ‘asthma attack’.
- Effect of ‘smoking’ on ‘cancer’.
- Effect of ‘coronary heart disease’ on ‘mortality’.
- Effect of ‘fertilizer’ on ‘crop yield’.

Typical regression would give a biased estimand because of confounders.

Motivation



Particulate matter
(Confounder)

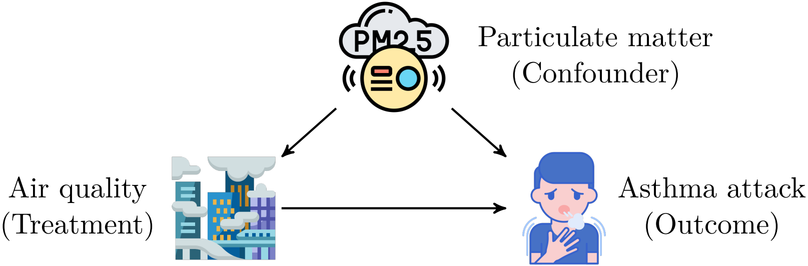
Air quality
(Treatment)

Asthma attack
(Outcome)

Two approaches to estimate causal effects:

- Randomized control trial
- Inference from observational data
  - Potential outcomes framework (PO) (Rubin, 1974, 1975, 1976, 1977, 1978; Rosenbaum & Rubin, 1983)
  - Structural causal model (SCM) (Pearl, 1995, 2000)

Motivation

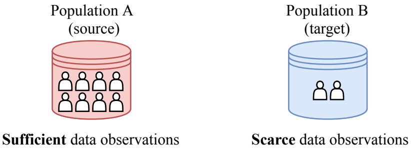


Problem:

- Observational data in a specific population might be **scarce**.
- For example:
  - Vaccination data of the elder in a country might be scarce and much less than the younger.

Motivation

- Estimate causal effects in a target population with scarce data observation



Population A (source)

Sufficient data observations

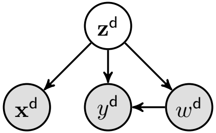
Population B (target)

Scarce data observations

Combining data might lead to poor causal estimands.

		Observed confounder	Latent confounder	With transfer	Randomized data
Without transfer	Louizos et al. (2017)		✓		
	Madras et al. (2019)		✓		
	Hill (2011)	✓			
	Shalit et al. (2017)	✓			
	Künzel et al. (2019)	✓			
	(to name a few)				
With transfer	Bareinboim & Pearl (2014)			✓	
	Bareinboim & Pearl (2016)			✓	
	Aglietti et al. (2020)			✓	✓
	AdaTRANS (proposed method)		✓	✓	

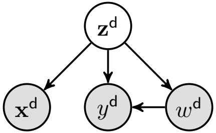
Causal quantities of interest



$\forall \mathsf{d} \in \{\mathsf{t}\} \cup \mathcal{S}$

- t : target population
- $\mathcal{S} = \{\mathsf{s}_1, \mathsf{s}_2, \ldots, \mathsf{s}_m\}$: collection of source populations
- y^{d}: the outcome
- w^{d}: the treatment
- $\mathbf{z}^{\mathsf{d}}$: the latent confounder
- $\mathbf{x}^{\mathsf{d}}$: the covariate

Causal quantities of interest



$\forall \mathsf{d} \in \{\mathsf{t}\} \cup \mathcal{S}$

- t : target population
- $\mathcal{S} = \{\mathsf{s}_1, \mathsf{s}_2, \ldots, \mathsf{s}_m\}$: collection of source populations
- y^{d}: the outcome
- w^{d}: the treatment
- $\mathbf{z}^{\mathsf{d}}$: the latent confounder
- $\mathbf{x}^{\mathsf{d}}$: the covariate

We estimate

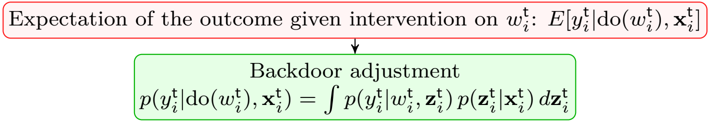
$$\mathtt{ite}(x) = E[y^{\mathsf{t}} | \mathrm{do}(w^{\mathsf{t}} = 1), \mathbf{x}^{\mathsf{t}} = x] - E[y^{\mathsf{t}} | \mathrm{do}(w^{\mathsf{t}} = 0), \mathbf{x}^{\mathsf{t}} = x]$$

$$\mathtt{ate} = E_X[\mathtt{ite}(X)]$$

The proposed method

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

The proposed method

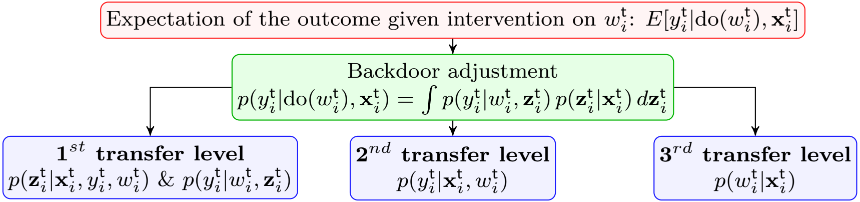


Expectation of the outcome given intervention on w_i^{t}: $E[y_i^{\mathsf{t}} | \mathrm{do}(w_i^{\mathsf{t}}), \mathbf{x}_i^{\mathsf{t}}]$

Backdoor adjustment

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

The proposed method



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

Backdoor adjustment

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

$\mathbf{1}^{st}$ transfer level

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

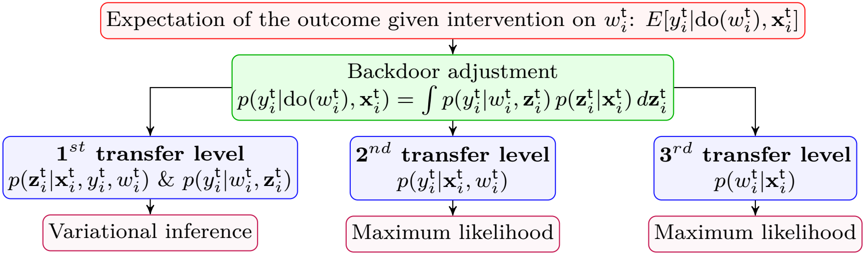
$\mathbf{2}^{nd}$ transfer level

$p(y_i^{\mathrm{t}}|\mathbf{x}_i^{\mathrm{t}},w_i^{\mathrm{t}})$

$\mathbf{3}^{rd}$ transfer level

$p(w_i^{\mathrm{t}}|\mathbf{x}_i^{\mathrm{t}})$

The proposed method



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

Backdoor adjustment

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

1^{st} transfer level

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

Variational inference

2^{nd} transfer level

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

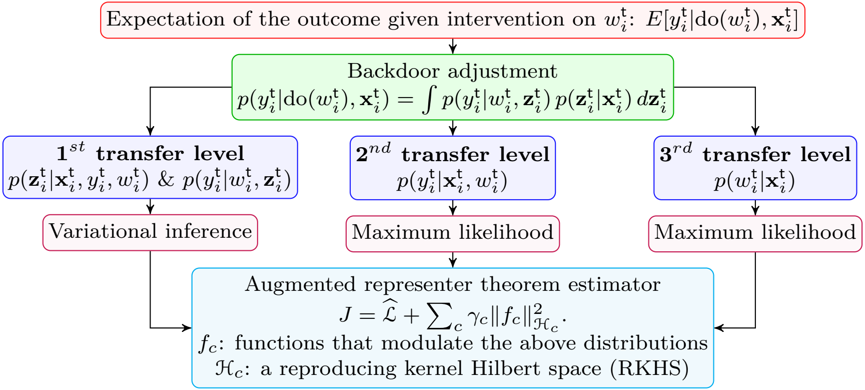
Maximum likelihood

3^{rd} transfer level

$p(w_i^{\mathsf{t}}|\mathbf{x}_i^{\mathsf{t}})$

Maximum likelihood

The proposed method



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

Backdoor adjustment

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

1^{st} transfer level

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

Variational inference

2^{nd} transfer level

$p(y_i^{\mathrm{t}}|\mathbf{x}_i^{\mathrm{t}}, w_i^{\mathrm{t}})$

Maximum likelihood

3^{rd} transfer level

$p(w_i^{\mathrm{t}}|\mathbf{x}_i^{\mathrm{t}})$

Maximum likelihood

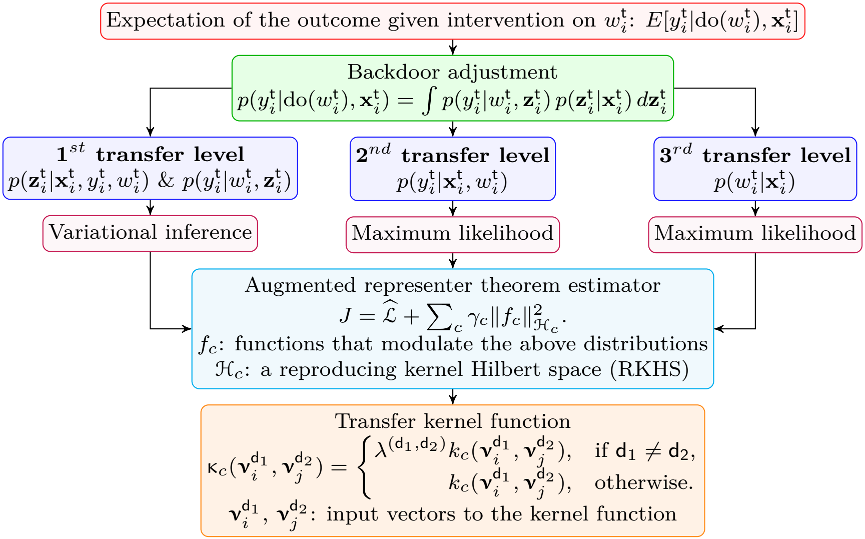
Augmented representer theorem estimator

$J = \widehat{\mathcal{L}} + \sum_c \gamma_c \|f_c\|^2_{\mathcal{H}_c}.$

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

Backdoor adjustment

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

1^{st} transfer level

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

Variational inference

2^{nd} transfer level

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

Maximum likelihood

3^{rd} transfer level

$p(w_i^{\mathsf{t}} | \mathbf{x}_i^{\mathsf{t}})$

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

$$\kappa_c(\boldsymbol{\nu}_i^{\mathsf{d}_1}, \boldsymbol{\nu}_j^{\mathsf{d}_2}) = \begin{cases} \lambda^{(\mathsf{d}_1, \mathsf{d}_2)} k_c(\boldsymbol{\nu}_i^{\mathsf{d}_1}, \boldsymbol{\nu}_j^{\mathsf{d}_2}), & \text{if } \mathsf{d}_1 \neq \mathsf{d}_2, \\ k_c(\boldsymbol{\nu}_i^{\mathsf{d}_1}, \boldsymbol{\nu}_j^{\mathsf{d}_2}), & \text{otherwise.} \end{cases}$$

$\boldsymbol{\nu}_i^{\mathsf{d}_1}$, $\boldsymbol{\nu}_j^{\mathsf{d}_2}$: input vectors to the kernel function

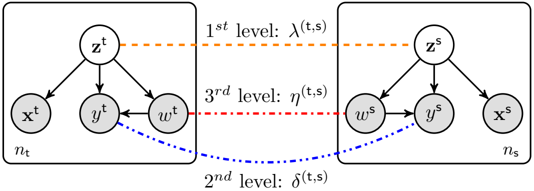
The proposed method

The transfer kernel function

$$\kappa_c(\boldsymbol{\nu}_i^{\mathsf{d}_1}, \boldsymbol{\nu}_j^{\mathsf{d}_2}) = \begin{cases} \lambda^{(\mathsf{d}_1,\mathsf{d}_2)} k_c(\boldsymbol{\nu}_i^{\mathsf{d}_1}, \boldsymbol{\nu}_j^{\mathsf{d}_2}), & \text{if } \mathsf{d}_1 \neq \mathsf{d}_2, \\ k_c(\boldsymbol{\nu}_i^{\mathsf{d}_1}, \boldsymbol{\nu}_j^{\mathsf{d}_2}), & \text{otherwise.} \end{cases}$$

The transfer factor $\lambda^{(\mathsf{d}_1,\mathsf{d}_2)}$ is learned.

- $\lambda^{(\mathsf{d}_1,\mathsf{d}_2)} = 0$: no transfer
- $\lambda^{(\mathsf{d}_1,\mathsf{d}_2)} = 1$: fully transfer
- $0 < \lambda^{(\mathsf{d}_1,\mathsf{d}_2)} < 1$: partially transfer



1^{st} level: $\lambda^{(\mathsf{t},\mathsf{s})}$
3^{rd} level: $\eta^{(\mathsf{t},\mathsf{s})}$
2^{nd} level: $\delta^{(\mathsf{t},\mathsf{s})}$
$\mathbf{z}^{\mathsf{t}}$
$\mathbf{x}^{\mathsf{t}}$
y^{t}
w^{t}
n_{t}
$\mathbf{z}^{\mathsf{s}}$
w^{s}
y^{s}
$\mathbf{x}^{\mathsf{s}}$
n_{s}

Experiments

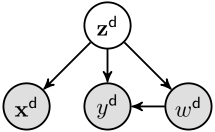
The aims of experiments

- We study the performance when the sources' distributions and the target's distributions are similar/different.
- We illustrate the performance when the similarity of sources and target slowly changes.

Experiments

Datasets

- Synthetic data
  - Simulate based on a ground truth causal graph.



$\forall \mathsf{d} \in \{\mathsf{t}\} \cup \mathcal{S}$

$$\begin{aligned}
\mathbf{z}_i^{\mathsf{d}} &\sim \mathsf{N}(\mathbf{0}, \sigma_z^2 \mathbf{I}_2),\\
x_{ij}^{\mathsf{d}} &\sim \mathsf{Bern}(\varphi(a_{0j} + (\mathbf{z}_i^{\mathsf{d}})^\top \boldsymbol{a}_{1j})),\\
w_i^{\mathsf{d}} &\sim \mathsf{Bern}(\varphi(b_0 + (\mathbf{z}_i^{\mathsf{d}})^\top \boldsymbol{b}_1^{\mathsf{d}})),\\
y_i^{\mathsf{d}}(0) &\sim \mathsf{N}(\zeta(c_0 + (\mathbf{z}_i^{\mathsf{d}})^\top \boldsymbol{c}_1^{\mathsf{d}}), \sigma_y^2),\\
y_i^{\mathsf{d}}(1) &\sim \mathsf{N}(\zeta(d_0 + (\mathbf{z}_i^{\mathsf{d}})^\top \boldsymbol{d}_1^{\mathsf{d}}), \sigma_y^2).
\end{aligned}$$

  - We keep $\mathbf{x}_i^{\mathsf{d}}$, w_i^{d} and either $y_i^{\mathsf{d}}(0)$ or $y_i^{\mathsf{d}}(1)$ as observed data.
  - $\boldsymbol{b}_1^{\mathsf{d}}$, $\boldsymbol{c}_1^{\mathsf{d}}$, $\boldsymbol{d}_1^{\mathsf{d}}$ are set differently on each population d.

Experiments

Datasets

- Twins data
  - Study the impact of twins' weight (treatment) on their mortality (outcome).
  - Source data: 1594 entries, Target data: 457 entries

Experiments

Evaluation metrics

- Precision in estimation of heterogeneous effects (PEHE) (Hill, 2011)

$$\varepsilon_{\text{PEHE}} = E\Big[\Big(\underbrace{(y_1 - y_0)}_{\text{True ITE}} - \underbrace{(\hat{y}_1 - \hat{y}_0)}_{\text{Estimated ITE}}\Big)^2\Big]$$

- Absolute error

$$\varepsilon_{\text{ATE}} = \Big|\underbrace{E(y_1 - y_0)}_{\text{True ATE}} - \underbrace{E(\hat{y}_1 - \hat{y}_0)}_{\text{Estimated ATE}}\Big|$$

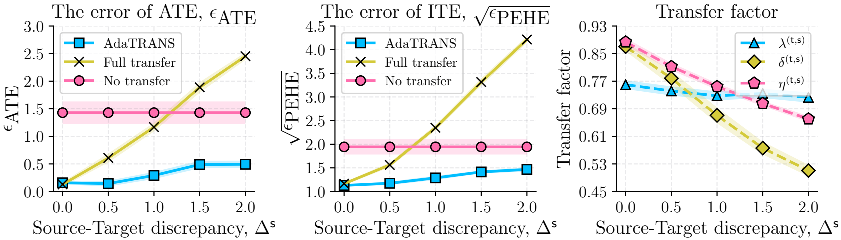
Experiments

The importance of adaptively causal transfer

$$\boldsymbol{b}_1^{\mathsf{t}} = [1.1, 1.7]^\top, \qquad \boldsymbol{c}_1^{\mathsf{t}} = [1.5, 1.8]^\top, \qquad \boldsymbol{d}_1^{\mathsf{t}} = [1.5, 2.8]^\top,$$

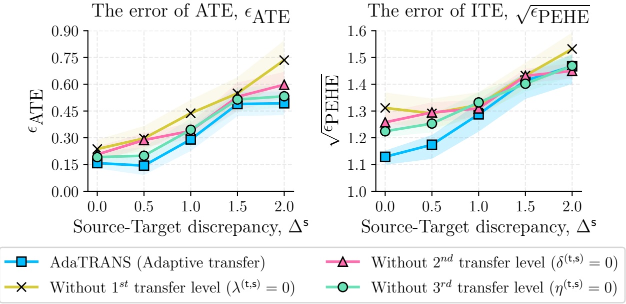
$$\boldsymbol{b}_1^{\mathsf{s}} = \boldsymbol{b}_1^{\mathsf{t}} + \Delta^{\mathsf{s}}[1,1]^\top, \qquad \boldsymbol{c}_1^{\mathsf{s}} = \boldsymbol{c}_1^{\mathsf{t}} + \Delta^{\mathsf{s}}[1,1]^\top, \qquad \boldsymbol{d}_1^{\mathsf{s}} = \boldsymbol{d}_1^{\mathsf{t}} + \Delta^{\mathsf{s}}[1,1]^\top,$$

We vary $\Delta^{\mathsf{s}} \in \{0.0, 0.5, 1.0, 1.5, 2.0\}$ to obtain different instances of the source data.



Experiments

Which transfer level is the most important?



Experiments

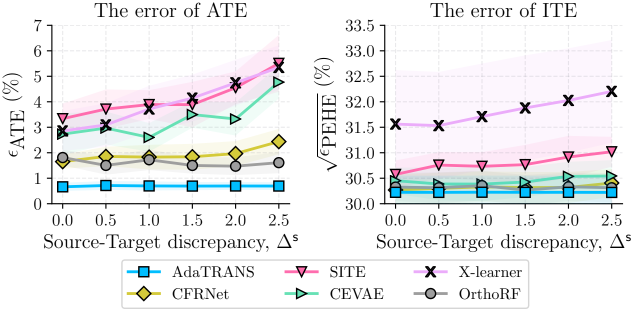
Performance analysis: compare with the baselines (Lower is better)

Method	The error of ITE ($\sqrt{\varepsilon_{\text{PEHE}}}$)			The error of ATE (ε_{ATE})		
	0-source	2-sources	4-sources	0-source	2-sources	4-sources
CEVAE$_{\text{stack}}$	3.1±.30	4.6±.39	4.8±.40	1.7±.29	2.8±.30	2.5±.26
CFRNet$_{\text{stack}}$	4.6±.51	8.9±.50	6.0±.19	1.6±.41	6.1±.48	4.0±.17
SITE$_{\text{stack}}$	6.0±.98	8.9±.61	7.5±.60	3.3±.67	6.4±.79	5.0±.76
BART$_{\text{stack}}$	2.5±.06	2.3±.03	2.2±.06	1.2±.13	0.7±.08	0.6±.09
R-learner$_{\text{stack}}$	3.0±.27	2.2±.11	1.8±.09	1.4±.35	1.2±.17	1.0±.10
X-learner$_{\text{stack}}$	2.0±.13	2.2±.12	1.9±.13	**1.0±.17**	1.0±.11	1.1±.13
OrthoRF$_{\text{stack}}$	6.2±.40	2.4±.03	2.2±.03	1.2±.37	0.5±.08	0.6±.06
CEVAE$_{\text{1-hot}}$	—	5.0±.43	3.3±.12	—	3.1±.42	1.9±.23
CFRNet$_{\text{1-hot}}$	—	4.4±.26	3.3±.21	—	3.3±.26	2.1±.17
SITE$_{\text{1-hot}}$	—	5.8±.99	3.2±.25	—	3.4±.67	2.1±.21
BART$_{\text{1-hot}}$	—	2.3±.03	2.2±.04	—	0.7±.10	0.4±.10
R-learner$_{\text{1-hot}}$	—	2.0±.07	1.7±.15	—	0.8±.15	0.8±.20
X-learner$_{\text{1-hot}}$	—	1.9±.12	1.8±.10	—	0.7±.13	0.6±.12
OrthoRF$_{\text{1-hot}}$	—	5.5±.30	4.1±.16	—	3.9±.22	2.6±.17
AdaTRANS	**1.6±.09**	**1.3±.03**	**1.3±.02**	1.1±.13	**0.2±.05**	**0.1±.03**

- The range of true ATE: (1.51, 1.87).
- The range of true ITE: (-5.69, 13.14).

Experiments

- Twins data
  - Study the impact of twins' weight (treatment) on their mortality (outcome).
  - Source data: 1594 entries, Target data: 457 entries



Summary

- We developed an adaptive method that estimates causal effects for a target population whose data is scarce.
- The proposed method required no prior information about the discrepancy about the source and target population.
- We assume that the source and the target share the same causal graph, but different structural equations.
- Limitations:
  - Causal effects in each population are identifiable.
  - The populations share similar causal graph & data features.
  - The confounders are independent and identically distributed.
- Future direction: allowing in the target population to be unidentifiable and have different set of data features.

Q & A

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).