Subspace Learning for Conditional Average Treatment Effect Estimation with Unmeasured Confounding

Han Qiao[1]

[1]South China Normal University
55 Zhongshan Avenue West, Tianhe District
Guangzhou, 510631 China
qiaohan@m.scnu.edu.cn

Abstract

Conditional average treatment effect (CATE) is the average causal effect of a treatment or an intervention (e.g., medication) on the outcome of interest, conditional on subjects' covariates. A key challenge in estimating causal effects from observational (OBS) data is to address unmeasured confounding. Mainstream methods that only rely on OBS data including sensitivity analysis, front door adjustment methods, and instrumental variables methods, may depend on strong assumptions. Recent studies suggest using randomized controlled trial (RCT) data to correct CATE estimates from biased OBS data, but existing methods may fail to efficiently utilize both data. In this paper, we present an end-to-end CATE estimation framework that addresses unmeasured confounding bias from OBS data using insights from limited unbiased RCT data. By learning representations from RCT data accounting for unmeasured confounding, our approach achieves unbiased CATE estimation. Our adaptive model structure mitigates overfitting and ensures performance across different RCT sample sizes. Extensive experiments on different datasets validate the effectiveness of the framework.

Introduction

Conditional mean treatment effect (CATE) refers to the causal effect of a treatment or intervention on a relevant outcome given subject characteristics (Pearl 2009), and it plays an important role in several fields such as e-commerce (Wu et al. 2022c), healthcare (Robins and Hernán 2016), and economics (Huynh, Kreinovich, and Sriboonchitta 2016). In these scenarios, CATE provides a basis for personalized decision-making, making resource allocation more accurate and efficient. The computation of CATE usually relies on causal inference methods, including randomized trials, propensity score matching, dual machine learning, etc. In recent years, with the rapid development of machine learning and deep learning technology, CATE estimation methods based on complex models have been widely used. By dealing with high-dimensional features and nonlinear relationships, these methods can more accurately reveal the causal effects of individual differentiation and provide more powerful tool support for practice in various fields (Huynh, Kreinovich, and Sriboonchitta 2016; Wang et al. 2024a).

To improve the accuracy of CATE estimation, several methods utilizing representation learning have been introduced. Representation learning techniques aim to learn useful features or embeddings from the data that can better capture the underlying structure of the relationships between covariates, treatment assignment, and outcomes. These learned representations are particularly valuable when the data are high-dimensional or involve complex interactions, which traditional methods may struggle to model effectively. Current strategies for obtaining these representations include approaches such as integral probability metric (IPM) regularization (Johansson, Shalit, and Sontag 2016), local similarity preservation (Yao et al. 2018), targeted learning (Zhang, Bellot, and Schaar 2020), and optimal transport (Wang et al. 2024b).

Unmeasured confounding refers to the situation where there are hidden or unobserved variables that influence both the treatment assignment and the outcome, leading to biased estimates of causal effects (Li et al. 2023). In the context of CATE estimation, if such unmeasured confounders are present, it becomes difficult to accurately assess the causal effect of a treatment or intervention because the relationship between the treatment (Charpignon et al. 2022), covariates, and outcome is distorted. Ignoring these unmeasured confounders can lead to misleading conclusions, as the estimated treatment effect may be confounded by factors that were not considered or included in the analysis. This is particularly challenging when trying to make reliable inferences about heterogeneous treatment effects across different subgroups, as the unmeasured confounding can vary across individuals, further complicating the estimation process (Xu and Li 2020) .

Numerous established methods heavily depend on large-scale observational (OBS) data to solve unmeasured confounding, such as instrumental variable techniques (Joshua D. Angrist and Rubin 1996), front-door adjustment (Fulcher et al. 2020), and sensitivity analysis (Imbens 2003). However, these methods are based on unverified assumptions, which raises concerns about their validity if these assumptions are violated (Hartwig et al. 2023). On the other hand, randomized controlled trial (RCT) data are widely considered the gold standard for causal inference (Prosperi et al. 2020). Nevertheless, the high costs and ethical dilemmas associated with RCTs often hinder their use, leading to small

sample sizes (Zabor, Kaizer, and Hobbs 2020). This makes it impractical to directly train causal models on RCT data in many cases (Hoogland et al. 2021). Some approaches attempt to combine RCT and OBS data, using residual correction techniques to address biases (Colnet et al. 2024).

In this paper, we propose a novel end-to-end framework to address unmeasured confounding with limited unbiased RCT data. Specifically, we learn representations through OBS data to obtain measured confounding effects, accurately predict results from OBS, and learn representations through RCT data to capture information about unmeasured confounding bias, thereby calibrating the residuals between OBS biased estimates and RCT unbiased estimates. This method also maximizes mutual information between the two representations to minimize information overlap, ensuring that the representation can effectively capture information of unmeasured confounding effect. The proposed framework adjusts model structures flexibly based on the size of available RCT data to mitigate overfitting issues.

The main contributions of this paper are summarized as follows:

- We propose an end-to-end framework for estimating CATE, which learns representations to account for effects from unmeasured confounding to calibrate residuals to correct biased estimates derived from OBS data.
- The proposed framework can flexibly adjust its model structure according to the sample size of RCT data, thus solving the over-fitting problem.
- Extensive experiments are conducted on IHDP and ACIC datasets to demonstrate the effectiveness of our proposal.

Related Work

CATE Estimation. CATE (Conditional Average Treatment Effect), also known as heterogeneous treatment effect (HTE), measures treatment effect variation across subgroups based on covariates. Traditional methods include matching (Dehejia and Wahba 2002), stratification (O'Muircheartaigh and Hedges 2014), reweighting (Rosenbaum 1987; Bang and Robins 2005), and tree-based methods (Chipman, George, and McCulloch 2010; Wager and Athey 2018). Recent advances, fueled by deep learning, have introduced representation learning and generative models. Representation learning aims to balance covariate distributions between treatment and control groups, reducing confounding bias using techniques such as IPM (Johansson, Shalit, and Sontag 2016; Shalit, Johansson, and Sontag 2017), similarity preservation (Yao et al. 2018, 2019), and optimal transport (Wang et al. 2024b; Torous, Gunsilius, and Rigollet 2021). Generative models like VAE (Louizos et al. 2017) and GANs (Yoon, Jordon, and Van Der Schaar 2018) estimate counterfactual outcomes by modeling the data generation process.

Unmeasured Confounding. Unmeasured confounding refers to unmeasured variables that influence both treatment and outcome, leading to biased CATE estimates (Ananth and Schisterman 2018). Solutions can be categorized into two types: those using observational (OBS) data and those combining OBS and RCT data. OBS-based methods include sensitivity analysis, which quantifies the impact of unmeasured confounding and provides bounds (Rosenbaum and Rubin 1983; Robins, Caspi, and Moffitt 2000), though the assumption of similar confounder behavior across individuals is often unrealistic (Imbens 2003). Auxiliary methods like instrumental variables (IV) and front-door adjustment use external instruments or causal pathways to address confounding (Imbens 2014; Wu et al. 2022a; Rudolph, Williams, and Díaz 2024; Bellemare, Bloem, and Wexler 2020; Fulcher et al. 2020), but have limitations related to linearity and causal graph knowledge (Frauen and Feuerriegel 2022; Shah, Shanmugam, and Kocaoglu 2024). Data fusion techniques combine RCT and OBS data to correct bias, with some assuming linear bias models (Kallus, Puli, and Shalit 2018; Ilse et al. 2021), while others, such as (Wu et al. 2022b) and (Cheng and Cai 2021), use methods like R-learner or weighted averages. These often require large sample sizes, which are difficult to obtain in RCTs.

Preliminaries

Problem Setup

We analyze two distinct data sources derived from the same target population: one obtained from a randomized controlled trial (RCT) and the other from an observational study (OBS). For each participant in either the RCT or the OBS study, the information is represented by the random tuple (X, Y, T), following the distribution P. Each dataset includes covariates $X \in \mathcal{X} \subseteq \mathbb{R}^d$, binary treatment assignments $T \in \{0, 1\}$ (where $T = 0$ indicates control and $T = 1$ denotes treatment), and outcomes Y. Using the Neyman-Rubin potential outcome framework (Imbens and Rubin 2015), we denote the potential outcomes as $Y(1)$ and $Y(0)$. The RCT data is represented as $D_i = (X_i, T_i, Y_i) : i \in \mathcal{A}$, while the OBS data is denoted as $D_i = (X_i, T_i, Y_i) : i \in \mathcal{B}$. The conditional average outcome is expressed as $\mu(X, G) = \mathbb{E}[Y|X, G]$. Furthermore, the Conditional Average Treatment Effect (CATE) is defined as the difference in the expected potential outcomes under the given conditions:

$$\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x].$$

To estimate the Causal Average Treatment Effect (CATE) from observed data, three additional assumptions are necessary, alongside the Stable Unit Treatment Value Assumption (SUTVA). These assumptions are: (1) Ignorability: $(Y(1), Y(0)) \perp T \mid X$; (2) Consistency: $Y = TY(1) + (1 - T)Y(0)$; and (3) Positivity: $0 < e(X, G) < 1$ for all $X \in \mathcal{X}$.

Methodology

Overview

We propose the an end-to-end framework as in figure 1, with the motivation of using a small unbiased data to calibrate the bias in OBS estimates for training the unbiased prediction model. Specifically, we use the treatment and control subspaces to learn the representations of the covariates for the treatment and control groups, respectively, and learn two representations in each subspace to capture the effects of

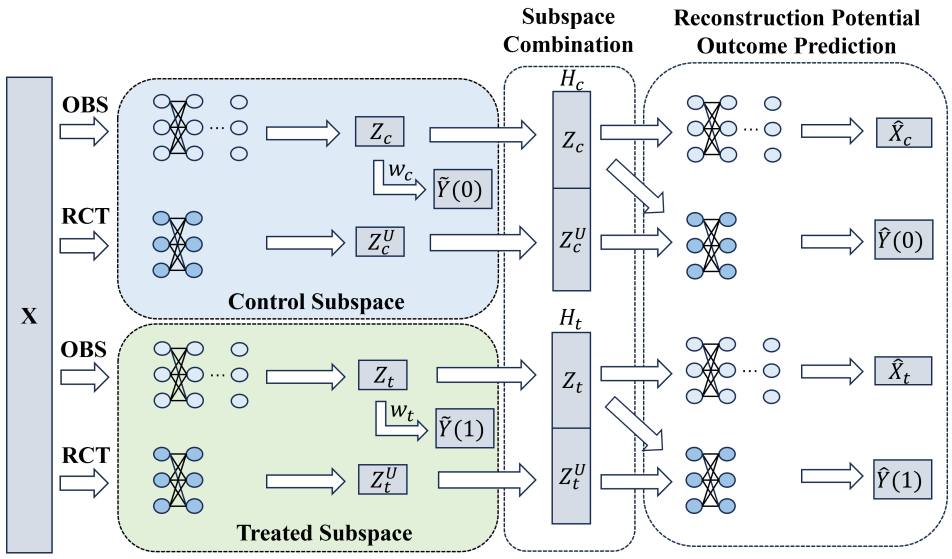

Figure 1: Framework of the proposed subspace learning approach

measurement confounding and unmeasurement confounding on the CATE estimates respectively, with one representation learned through OBS learning to accurately predict the outcome of OBS data, and the other representation learned through RCT to bridge the gap between unbiased RCT estimates and OBS biased estimates. The loss function of the proposed approach is defined as

$$\mathcal{L} = \mathcal{L}_f + \alpha\mathcal{L}_{\text{MI}} + \beta\left(\mathcal{L}_c^{\text{OBS}} + \mathcal{L}_t^{\text{OBS}}\right) + \gamma\mathcal{L}_{\text{rec}} + \lambda\|F\|_2,$$

where $\mathcal{L}_f$ is the RCT outcome prediction loss, $\mathcal{L}_c^{\text{OBS}}$, $\mathcal{L}_t^{\text{OBS}}$ are the OBS outcome prediction losses in the control and treated subspace, $\mathcal{L}_{\text{MI}}$ is mutual information loss in subspace combination, $\mathcal{L}_{\text{rec}}$ is the reconstruction loss, and $\|F\|_2$ is the regularization on all parameters and α, β, γ, λ is the hyperparameters.

Network Architecture

Representation Unlike previous studies that align covariates from different distributions between treatment and control groups using treatment-independent representations, which can lead to significant information loss (Nagalapatti et al. 2024), we employ treatment subspace and control subspace to separately learn representations of covariates for the treatment and control groups.

To address the unmeasured confounding bias, we learn the representation Z_c^U, Z_t^U with RCT data to obtain the information of unmeasured confounding, which helps correct the biased estimates $\tilde{Y}(x)$ to unbiased estimates $\hat{Y}(x)$.

Specifically, in each subspace, we design two feedforward representation network $\Phi_c(X)$, $\Phi_c^U(X)$ with d_c and d_{uc} hidden layers and employ the exponential linear unit (ELU) as the activation function. Specifically, measured confounding are mapped through deep neural networks to the representations $Z_c = \Phi_c(X)$, $Z_t = \Phi_t(X)$ learned from OBS data $Z_c = \Phi_c(X)$, $Z_t = \Phi_t(X)$. The unmeasured confounding are mapped through shallow neural networks to the representations $Z_c^U = \Phi_c^U(X)$, $Z_t^U = \Phi_t^U(X)$ learned from RCT data. It is important that adjustments are made to the shallow neural networks based on the sample size of RCT data to prevent overfitting.

To better represent control information, we introduced an external linear prediction layer to generate OBS outcome estimates in the control space. $\tilde{\text{Y}}(0) = (\mathbf{w}_c)^\top Z_c + b_c$, where $Z_c \in \mathbb{R}^{k_c \times n}$, $\mathbf{w}_c \in \mathbb{R}^{k_c}$, $Z_c^U \in \mathbb{R}^{k_{uc} \times m}$ $Z_c \in \mathbb{R}^{k_c \times n}$, k_c is the dimension of the last hidden layer of Φ_c. The treatment subspace follows a similar approach: $\tilde{\text{Y}}(1) = (\mathbf{w}_t)^\top \mathbf{Z}_t + b_t$, where $Z_t \in \mathbb{R}^{k_t \times n}$, $\mathbf{w}_t \in \mathbb{R}^{k_t}$, $Z_t^U \in \mathbb{R}^{k_{ut} \times m}$, k_t is the dimension of the last hidden layer of Φ_t, k_{ut} is the dimension of the last hidden layer of Φ_t^U.

Since there is hidden confounding in the OBS data, the estimated outcomes are not unbiased. For this reason, we name the prediction result $\tilde{Y}_0$ as the pseudo-control outcome and $\tilde{Y}_1$ as the pseudo-treatment outcome. We use the pseudo-differential $\mathcal{L}_c^{\text{OBS}}$ and $\mathcal{L}_t^{\text{OBS}}$ to measure the distance between the pseudo-outcome and the outcome of the real observed data.

$$\mathcal{L}_c^{\text{OBS}} = \frac{1}{\sum_{i\in\mathcal{B}} \mathbb{I}(T_i = 0)}(1 - T_i)l(Y_i, \tilde{Y}_i(0)),$$
$$\mathcal{L}_t^{\text{OBS}} = \frac{1}{\sum_{i\in\mathcal{B}} \mathbb{I}(T_i = 1)}T_i l(Y_i, \tilde{Y}_i(1)),$$

where $\mathbb{I}(\cdot)$ is the indicator function.

Combination Our approach concatenates the measured confounding representations learned from OBS data and unmeasured confounding representations learned from RCT data to jointly estimate the unbiased outcome.

We define two representation matrices H_c and H_t. The matrix H_c represents the covariate representations for the control group, defined as $H_c = \begin{bmatrix} Z_c \\ Z_c^U \end{bmatrix}$. Similarly, the matrix

H_t denotes the covariate representations for the treatment group, defined as: $H_t = \begin{bmatrix} Z_t \\ Z_t^U \end{bmatrix}$, where $H_c \in \mathbb{R}^{(k_c+k_{uc})\times N}$ and $H_t \in \mathbb{R}^{(k_t+k_{ut})\times N}$ represent the representation matrices for the control and treatment groups, respectively. k_c and k_t are the dimensions of the measured confounding, k_{uc} and k_{ut} are the dimensions of the representations of the unmeasured confounding, and N is the sample size.

To reduce overlap in the information already contained in Z relative to Z^U, we adopt the approach of minimizing the mutual information between Z and Z^U:

$$\begin{aligned}\mathcal{L}_{\mathrm{MI}} = &- \left(\mathbb{E}_{p(Z_c,Z_c^U)}\left[K\left(Z_c, Z_c^U\right)\right]\right. \\ &\left. - \log\left(\mathbb{E}_{p(Z_c)p(Z_c^U)}\left[e^{K\left(Z_c,Z_c^U\right)}\right]\right)\right. \\ &+ \left(\mathbb{E}_{p(Z_c,Z_c^U)}\left[K\left(Z_c, Z_c^U\right)\right]\right. \\ &\left. - \log\left(\mathbb{E}_{p(Z_c)p(Z_c^U)}\left[e^{K\left(Z_c,Z_c^U\right)}\right]\right)\right.,\end{aligned}$$

where $K(\cdot)$ is functions parameterized by neural networks. This approach ensures that Z^U captures distinct aspects of confounding not already captured by Z, which helps Z^U capture the effect of unmeasured confounding.

Reconstruction and Prediction To ensure that Z_c and Z_t retain as much information about the original covariates as possible, we introduce the decoder networks Ψ_c, Ψ_t to reconstruct the original control and treated data: $\hat{X}_c = \Psi_c(Z_c)$, $\hat{X}_t = \Psi_t(Z_t)$. The reconstruction loss is

$$\begin{aligned}\mathcal{L}_{\mathrm{rec}} = &\sum_{i\in\mathcal{B}}((1-T_i)\sum_{j=1}^{d} l(X_{i,j}, \hat{X}_{c,i,j}) \\ &+ T_i\sum_{j=1}^{d} l(X_{i,j}, \hat{X}_{t,i,j})),\end{aligned}$$

where $\hat{X}_{c,i,j}$ and $\hat{X}_{t,i,j}$ are reconstructed value of covariate j for sample i in the control group and treatment group respectively. To ensure that there is no overfitting with limited RCT data and that CATE can be accurately estimated, we avoid directly predicting outcomes with Z_c^U and Z_t^U. Instead, we use the concatenated representations H_c, H_t to predict the unbiased estimates of outcomes in RCT with shallow neural network. We define $f_c(H_c)$ and $f_t(H_t)$ as the predictors for control and treatment outcomes $\hat{Y}_0 = f_c(H_c)$, $\hat{Y}_1 = f_t(H_t)$. The outcome prediction loss as follows.

$$\begin{aligned}L_f = &\frac{1}{\sum_{i\in\mathcal{A}}\mathbb{I}(T_i=1)} T_i l(Y_i, \hat{Y}_i(1)) \\ &+ \frac{1}{\sum_{i\in\mathcal{A}}\mathbb{I}(T_i=0)}(1-T_i) l(Y_i, \hat{Y}_i(0)).\end{aligned}$$

Advantages: We utilize an auto-encoder approach to predict accurate OBS estimates and employ limited unbiased RCT data to learn representations effectively capturing unmeasured confounding biases. This enables us to calibrate residuals between OBS and RCT estimates to achieve unbiased learning with unmeasured confounding. In additional, we design a flexible network architecture that adjusts based on the sample size of RCT data to prevent overfitting. Additionally, considering the limited informational of RCT data for outcome prediction, we focus instead on concatenated representations to predict unbiased CATE estimates.

Experiment

Dataset and Prepossessing

Following previous studies (Shalit, Johansson, and Sontag 2017; Louizos et al. 2017; Yoon, Jordon, and Van Der Schaar 2018), we conduct experiments on two semi-synthetic dataset, **IHDP** (Hill 2011), and **ACIC** (Dorie et al. 2019). The **IHDP** introduced a semi-synthetic dataset for causal effect estimation. The dataset was based on the Infant Health and Development Program (IHDP), in which the covariates were generated by a randomized experiment investigating the effect of home visits by specialists on future cognitive scores. it consists of 747 units (19% treated, 81% control) and 25 covariates measuring the children and their mothers. The **ACIC** is a common benchmark dataset developed for the 2016 Atlantic Causal Inference Conference competition data. It comprises 4,802 units (28% treated, 72% control) and 82 covariates measuring aspects of the linked birth and infant death data (LBIDD).

Data Preprocessing

Since the dataset does not contain a separate RCT dataset, it is necessary to separate the RCT data. First, we slice the training, validation, and test sets in the ratio of 80/10/10. Second, for all samples in the validation set, we randomly assign treatments according to the following formula and replace the original treatment T and factual outcome Y_f:

$T_{new} = \mathrm{Bern}(0.5),\ \ Y_{new} = \mathbb{I}\{T_{new} = T\}(Y_f - Y_{cf}) + Y_{cf}$,

where the $\mathrm{Bern}(\cdot)$ is the Bernoulli distribution, Y_f is the factual outcome, and Y_{cf} is the counterfactual outcome. For the training set, we split 10% of the samples as RCT data, and replace the treatment T and factual outcome Y using the above formula. In addition, to simulate the unmeasured confounding effect, we randomly mask 50% features in our experiment.

Baselines and Evaluation Metrics

We compare our method with the following baselines: T-learner (Künzel et al. 2019), X-learner (Künzel et al. 2019), Causal Forest (Wager and Athey 2018) BNN (Johansson, Shalit, and Sontag 2016), TARNet (Shalit, Johansson, and Sontag 2017),CEVAE (Louizos et al. 2017), GANITE (Yoon, Jordon, and Van Der Schaar 2018), DragonNet (Shi, Blei, and Veitch 2019), DESCN (Zhong et al. 2022), and ESCFR (Wang et al. 2023). Following the previous studies (Shalit, Johansson, and Sontag 2017; Yao et al. 2018), we evaluation the performance of CATE estimation using the *Precision in Estimation of Heterogeneous Effects* (PEHE), which is defined as

$$\epsilon_{\mathrm{PEHE}} = \frac{1}{N}\sum_{i=1}^{N}((\hat{Y}_i(1) - \hat{Y}_i(0)) - (Y_i(1) - Y_i(0)))^2,$$

Table 1: The experiment results on the IHDP dataset and ACIC dataset. The best result is bolded.

	IHDP				**ACIC**			
	In-sample		Out-sample		In-sample		Out-sample	
Methods	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}
T-learner	2.06 ± 0.04	1.27 ± 0.05	2.23 ± 0.04	1.48 ± 0.07	2.76 ± 0.13	1.05 ± 0.40	2.95 ± 0.11	0.93 ± 0.39
X-learner	2.07 ± 0.04	1.15 ± 0.04	2.15 ± 0.05	1.34 ± 0.05	2.80 ± 0.11	0.42 ± 0.26	3.18 ± 0.13	0.53 ± 0.26
BNN	2.40 ± 0.22	1.29 ± 0.43	2.93 ± 0.24	1.48 ± 0.47	2.68 ± 0.30	0.97 ± 0.34	2.41 ± 0.18	0.88 ± 0.32
Causal Forest	1.87 ± 0.00	1.01 ± 0.00	1.87 ± 0.00	1.08 ± 0.00	2.23 ± 0.00	0.49 ± 0.00	2.31 ± 0.00	$\mathbf{0.42 \pm 0.00}$
TARNet	2.30 ± 0.31	1.28 ± 0.48	2.73 ± 0.40	1.58 ± 0.52	2.94 ± 0.27	1.16 ± 0.37	2.75 ± 0.23	1.03 ± 0.36
CEVAE	4.30 ± 0.38	3.99 ± 0.41	4.37 ± 0.38	4.07 ± 0.41	4.92 ± 0.21	2.59 ± 0.40	5.04 ± 0.18	2.28 ± 0.40
GANITE	6.57 ± 6.32	4.48 ± 1.32	6.60 ± 6.31	5.21 ± 3.23	4.60 ± 1.81	2.69 ± 0.80	4.59 ± 1.81	2.75 ± 0.73
DragonNet	2.54 ± 0.33	1.70 ± 0.51	3.01 ± 0.37	1.99 ± 0.53	2.91 ± 0.24	1.00 ± 0.48	2.69 ± 0.22	0.89 ± 0.46
DESCN	2.68 ± 0.88	1.78 ± 0.97	3.03 ± 1.05	2.06 ± 1.08	3.51 ± 1.84	0.80 ± 0.40	2.93 ± 1.43	0.81 ± 0.57
ESCFR	2.35 ± 0.25	1.11 ± 0.51	2.82 ± 0.32	1.33 ± 0.44	2.67 ± 0.17	0.87 ± 0.33	2.42 ± 0.20	0.71 ± 0.34
Ours	$\mathbf{1.73 \pm 0.15}$	$\mathbf{0.90 \pm 0.09}$	$\mathbf{1.80 \pm 0.12}$	$\mathbf{0.95 \pm 0.23}$	$\mathbf{2.11 \pm 0.51}$	$\mathbf{0.40 \pm 0.11}$	$\mathbf{2.15 \pm 0.12}$	0.68 ± 0.27

Table 2: Results (mean$_{\pm\text{std}}$) of PEHE and ATE on IHDP and ACIC Datasets.

	IHDP				**ACIC**			
	PEHE-In	PEHE-Out	ATE-In	ATE-Out	PEHE-In	PEHE-Out	ATE-In	ATE-Out
Ours w/o $\mathcal{L}_{\text{MI}}$	1.89 ± 0.17	1.01 ± 0.21	1.93 ± 0.18	1.09 ± 0.25	2.18 ± 0.31	0.26 ± 0.31	2.48 ± 0.20	0.44 ± 0.30
Ours w/o $\mathcal{L}_e$ and $\mathcal{L}_t$	2.26 ± 0.18	1.35 ± 0.07	2.15 ± 0.16	1.33 ± 0.20	2.35 ± 0.36	0.70 ± 0.20	2.40 ± 0.25	0.77 ± 0.25
Ours w/o $\mathcal{L}_{\text{rec}}$	1.82 ± 0.13	1.04 ± 0.09	1.85 ± 0.16	1.01 ± 0.16	2.21 ± 0.32	0.43 ± 0.14	2.11 ± 0.14	0.69 ± 0.31
Ours w/o $\mathcal{L}_f$	2.15 ± 0.25	1.24 ± 0.26	2.10 ± 0.16	1.10 ± 0.27	2.54 ± 0.27	0.61 ± 0.28	2.37 ± 0.14	0.80 ± 0.20
Ours	1.73 ± 0.15	0.90 ± 0.09	1.80 ± 0.12	0.95 ± 0.23	2.11 ± 0.51	0.40 ± 0.11	2.15 ± 0.12	0.68 ± 0.27

where $\hat{Y}_1(Z)$ and $\hat{Y}_0(Z)$ are the predicted values for the corresponding true potential outcomes. In addition, we also use the absolute error in *Average Treatment Effect* (ATE) to evaluate performance, which is defined as

$$\epsilon_{\text{ATE}} = \frac{1}{N}|\sum_{i=1}^{N}((\hat{Y}_i(1) - \hat{Y}_i(0)) - (Y_i(1) - Y_i(0)))|.$$

Both in-sample and out-of-sample performances are reported in our experiments.

Performance Analysis

The experiment results on the IHDP dataset and ACIC datasets are shown in Table 1. We have the following findings: first, among the baseline models, statistical models such as T-learner, S-learner, and Casual Forest demonstrate competitive performance and methods based on deep representation learning demonstrate sub-optimal performance. In addition, methods based on generative models perform poorly, possibly because the assumptions of the data generation process are violated. Second, our method outperforms the baseline method in all scenarios, showing the effectiveness of the proposed structure.

Ablation Study

In the proposed framework, all the loss components play an important role. Therefore, it is necessary to design an ablation study to explore the effects of each loss. The results are shown in Table 2. First, the method that includes all losses achieves the best performance, moreover, we find that it is the $\mathcal{L}_f$, $\mathcal{L}_e^{OBS}$, and $\mathcal{L}_c^{OBS}$ losses that most affect the estimation performance. This is because when without $\mathcal{L}_e^{OBS}$ and $\mathcal{L}_c^{OBS}$, the $\mathcal{L}_f$ loss alone will have an overfitting problem, meanwhile, there is no way to correct for the confounding bias from the observational data by relying on only $\mathcal{L}_e^{OBS}$ and $\mathcal{L}_c^{OBS}$. Meanwhile, the model without $\mathcal{L}_e^{OBS}$ and $\mathcal{L}_c^{OBS}$ performs worse than the model without $\mathcal{L}_f$, which further illustrates the overfitting problem associated with fitting the model directly on limited RCT data. In addition, we find that the model without $\mathcal{L}_{MI}$ performs worse than the model without $\mathcal{L}_{rec}$. This is due to the fact that $\mathcal{L}_{MI}$ can constrain the overlap in the information between the learned two representations Z_c and Z_c^U, whereas $\mathcal{L}_{rec}$ is not very helpful in learning the representation.

Conclusion

This paper introduces a novel subspace learning framework that utilizes limited unbiased RCT data to mitigate unmeasured confounding bias. Our framework leverages OBS data to learn representations that account for measured confounding effects, accurately predicting outcomes. Additionally, it incorporates RCT data to capture information on the unmeasured confounding bias, calibrating residuals between OBS biased estimates and RCT unbiased estimates. Furthermore, our framework maximizes mutual information between these representations to effectively capture the effects of unmeasured confounders. The proposed framework adopts flexible model structures based on the sample size of RCT data to address overfitting issues. We conduct experiments on two datasets to show superiority of our approach.

References

Ananth, C. V.; and Schisterman, E. F. 2018. Hidden biases in observational epidemiology: the case of unmeasured confounding. *BJOG: an international journal of obstetrics and gynaecology*, 125(6): 644.

Bang, H.; and Robins, J. M. 2005. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4): 962–973.

Bellemare, M. F.; Bloem, J. R.; and Wexler, N. 2020. The Paper of How: Estimating Treatment Effects Using the Front-Door Criterion. *Oxford Bulletin of Economics and Statistics*.

Charpignon, M.-L.; Vakulenko-Lagun, B.; Zheng, B.; Magdamo, C.; Su, B.; Evans, K.; Rodriguez, S.; Sokolov, A.; Boswell, S.; Sheu, Y.-H.; et al. 2022. Causal inference in medical records and complementary systems pharmacology for metformin drug repurposing towards dementia. *Nature communications*, 13(1): 7652.

Cheng, D.; and Cai, T. 2021. Adaptive combination of randomized and observational data. *arXiv preprint arXiv:2111.15012*.

Chipman, H. A.; George, E. I.; and McCulloch, R. E. 2010. BART: Bayesian additive regression trees.

Colnet, B.; Mayer, I.; Chen, G.; Dieng, A.; Li, R.; Varoquaux, G.; Vert, J.-P.; Josse, J.; and Yang, S. 2024. Causal inference methods for combining randomized trials and observational studies: a review. *Statistical science*, 39(1): 165–191.

Dehejia, R. H.; and Wahba, S. 2002. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics*, 84(1): 151–161.

Dorie, V.; Hill, J.; Shalit, U.; Scott, M.; and Cervone, D. 2019. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1): 43–68.

Frauen, D.; and Feuerriegel, S. 2022. Estimating individual treatment effects under unobserved confounding using binary instruments. *arXiv preprint arXiv:2208.08544*.

Fulcher, I. R.; Shpitser, I.; Marealle, S.; and Tchetgen Tchetgen, E. J. 2020. Robust inference on population indirect causal effects: the generalized front door criterion. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(1): 199–214.

Hartwig, F. P.; Wang, L.; Smith, G. D.; and Davies, N. M. 2023. Average causal effect estimation via instrumental variables: the no simultaneous heterogeneity assumption. *Epidemiology*, 34(3): 325–332.

Hill, J. L. 2011. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1): 217–240.

Hoogland, J.; IntHout, J.; Belias, M.; Rovers, M. M.; Riley, R. D.; E. Harrell Jr, F.; Moons, K. G.; Debray, T. P.; and Reitsma, J. B. 2021. A tutorial on individualized treatment effect prediction from randomized trials with a binary endpoint. *Statistics in medicine*, 40(26): 5961–5981.

Huynh, V.-N.; Kreinovich, V.; and Sriboonchitta, S. 2016. *Causal inference in econometrics*. Springer.

Ilse, M.; Forré, P.; Welling, M.; and Mooij, J. M. 2021. Combining interventional and observational data using causal reductions. *arXiv preprint arXiv:2103.04786*.

Imbens, G. 2014. Instrumental variables: An econometrician's perspective. Technical report, National Bureau of Economic Research.

Imbens, G. W. 2003. Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, 93(2): 126–132.

Imbens, G. W.; and Rubin, D. B. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press.

Johansson, F.; Shalit, U.; and Sontag, D. 2016. Learning representations for counterfactual inference. In *International conference on machine learning*, 3020–3029. PMLR.

Joshua D. Angrist, G. W. I.; and Rubin, D. B. 1996. Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association*, 91(434): 444–455.

Kallus, N.; Puli, A. M.; and Shalit, U. 2018. Removing hidden confounding by experimental grounding. *Advances in neural information processing systems*, 31.

Künzel, S. R.; Sekhon, J. S.; Bickel, P. J.; and Yu, B. 2019. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10): 4156–4165.

Li, H.; Xiao, Y.; Zheng, C.; and Wu, P. 2023. Balancing unobserved confounding with a few unbiased ratings in debiased recommendations. In *Proceedings of the ACM Web Conference 2023*, 1305–1313.

Louizos, C.; Shalit, U.; Mooij, J. M.; Sontag, D.; Zemel, R.; and Welling, M. 2017. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30.

Nagalapatti, L.; Iyer, A.; De, A.; and Sarawagi, S. 2024. Continuous treatment effect estimation using gradient interpolation and kernel smoothing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 14397–14404.

O'Muircheartaigh, C.; and Hedges, L. V. 2014. Generalizing from unrepresentative experiments: a stratified propensity score approach. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 63(2): 195–210.

Pearl, J. 2009. *Causality*. Cambridge university press.

Prosperi, M.; Guo, Y.; Sperrin, M.; Koopman, J. S.; Min, J. S.; He, X.; Rich, S.; Wang, M.; Buchan, I. E.; and Bian, J. 2020. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence*, 2(7): 369–375.

Robins, J. M.; and Hernán, M. 2016. Causal inference.

Robins, R. W.; Caspi, A.; and Moffitt, T. E. 2000. Two personalities, one relationship: both partners' personality traits shape the quality of their relationship. *Journal of personality and social psychology*, 79(2): 251.

Rosenbaum, P. R. 1987. Model-based direct adjustment. *Journal of the American statistical Association*, 82(398): 387–394.

Rosenbaum, P. R.; and Rubin, D. B. 1983. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B (Methodological)*, 45(2): 212–218.

Rudolph, K. E.; Williams, N.; and Díaz, I. 2024. Using instrumental variables to address unmeasured confounding in causal mediation analysis. *Biometrics*, 80(1): ujad037.

Shah, A.; Shanmugam, K.; and Kocaoglu, M. 2024. Front-door adjustment beyond markov equivalence with limited graph knowledge. *Advances in Neural Information Processing Systems*, 36.

Shalit, U.; Johansson, F. D.; and Sontag, D. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, 3076–3085. PMLR.

Shi, C.; Blei, D.; and Veitch, V. 2019. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, 32.

Torous, W.; Gunsilius, F.; and Rigollet, P. 2021. An optimal transport approach to causal inference. *arXiv preprint arXiv:2108.05858*.

Wager, S.; and Athey, S. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523): 1228–1242.

Wang, F.; Chen, C.; Liu, W.; Fan, T.; Liao, X.; Tan, Y.; Qi, L.; and Zheng, X. 2024a. CE-RCFR: Robust counterfactual regression for consensus-enabled treatment effect estimation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 3013–3023.

Wang, H.; Fan, J.; Chen, Z.; Li, H.; Liu, W.; Liu, T.; Dai, Q.; Wang, Y.; Dong, Z.; and Tang, R. 2023. Optimal Transport for Treatment Effect Estimation. In *Advances in Neural Information Processing Systems*.

Wang, H.; Fan, J.; Chen, Z.; Li, H.; Liu, W.; Liu, T.; Dai, Q.; Wang, Y.; Dong, Z.; and Tang, R. 2024b. Optimal transport for treatment effect estimation. *Advances in Neural Information Processing Systems*, 36.

Wu, A.; Kuang, K.; Li, B.; and Wu, F. 2022a. Instrumental variable regression with confounder balancing. In *International Conference on Machine Learning*, 24056–24075. PMLR.

Wu, A.; Kuang, K.; Li, B.; and Wu, F. 2022b. Instrumental Variable Regression with Confounder Balancing. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvari, C.; Niu, G.; and Sabato, S., eds., *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 24056–24075. PMLR.

Wu, P.; Li, H.; Deng, Y.; Hu, W.; Dai, Q.; Dong, Z.; Sun, J.; Zhang, R.; and Zhou, X.-H. 2022c. On the Opportunity of Causal Learning in Recommendation Systems: Foundation, Estimation, Prediction and Challenges. In Raedt, L. D., ed., *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, 5646–5653. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Xu, Y.; and Li, A. 2020. The relationship between innovative human capital and interprovincial economic growth based on panel data model and spatial econometrics. *Journal of computational and applied mathematics*, 365: 112381.

Yao, L.; Li, S.; Li, Y.; Huai, M.; Gao, J.; and Zhang, A. 2018. Representation learning for treatment effect estimation from observational data. *Advances in neural information processing systems*, 31.

Yao, L.; Li, S.; Li, Y.; Huai, M.; Gao, J.; and Zhang, A. 2019. ACE: Adaptively similarity-preserved representation learning for individual treatment effect estimation. In *2019 IEEE International Conference on Data Mining (ICDM)*, 1432–1437. IEEE.

Yoon, J.; Jordon, J.; and Van Der Schaar, M. 2018. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *International conference on learning representations*.

Zabor, E. C.; Kaizer, A. M.; and Hobbs, B. P. 2020. Randomized controlled trials. *Chest*, 158(1): S79–S87.

Zhang, Y.; Bellot, A.; and Schaar, M. 2020. Learning overlapping representations for the estimation of individualized treatment effects. In *International Conference on Artificial Intelligence and Statistics*, 1005–1014. PMLR.

Zhong, K.; Xiao, F.; Ren, Y.; Liang, Y.; Yao, W.; Yang, X.; and Cen, L. 2022. Descn: Deep entire space cross networks for individual treatment effect estimation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4612–4620.