Learning Decomposed Representation for Counterfactual Inference

Anpeng Wu
Zhejiang University
anpwu2019@gmail.com

Kun Kuang
Zhejiang University
kunkuang@zju.edu.cn

Junkun Yuan
Zhejiang University
yuanjk@zju.edu.cn

Bo Li
Tsinghua University
libo@sem.tsinghua.edu.cn

Pan Zhou
Huazhong University of Science and Technology
panzhou@hust.edu.cn

Jianrong Tao
NetEase Inc.
hztaojianrong@corp.netease.com

Qiang Zhu
Zhejiang University
zhuq@zju.edu.cn

Yueting Zhuang
Zhejiang University
yzhuang@zju.edu.cn

Fei Wu
Zhejiang University
wufei@cs.zju.edu.cn

Abstract

One fundamental problem in the learning treatment effect from observational data is confounder identification and balancing. Most of the previous methods realized confounder balancing by treating all observed variables as confounders, ignoring the identification of confounders and non-confounders. In general, not all the observed variables are confounders which are the common causes of both the treatment and the outcome, some variables only contribute to the treatment and some contribute to the outcome. Balancing those non-confounders would generate additional bias for treatment effect estimation. By modeling the different relations among variables, treatment and outcome, we propose a synergistic learning framework to 1) identify and balance confounders by learning decomposed representation of confounders and non-confounders, and simultaneously 2) estimate the treatment effect in observational studies via counterfactual inference. Our empirical results demonstrate that the proposed method can precisely identify and balance confounders, while the estimation of the treatment effect performs better than the state-of-the-art methods on both synthetic and real-world datasets.

1 Introduction

In [13], causal inference can be defined as the process of inferring causal connections based on the conditions of the occurrence of an effect, which plays an essential role in the decision-making process. One fundamental problem in causal inference is treatment effect estimation. For example, in the medical field, accurately assessing the treatment effect of a particular drug on each patient will help doctors to decide which medical procedure (e.g., taking the drug or not) will benefit a certain patient most. The gold standard approach for treatment effect estimation is to perform a Randomized Controlled Trial (RCT), where different treatments (i.e., medical procedure) are randomly assigned to units (i.e., patients). However, fully RCT is often expensive, unethical or even infeasible.

Preprint. Under review.

Different from the traditional RCT approachs, we focus on treatment effect estimation from observational data $\mathcal{D} = \{x_i, t_i, y_i^{t_i}\}_{i=1}^n$, where n refers to the number of units. For each unit (e.g., patient) indexed by i, we observe its context characteristics $x_i \in \mathcal{X}$, its choice on treatment $t_i \in \mathcal{T}$ from a set of treatment options (e.g., {0:placebo, 1:drug}), and the corresponding outcome (e.g., recovery or not) $y_i^{t_i} \in \mathcal{Y}$ as a result of choosing treatment t_i. In our context, we are interested in the case of a binary treatment (i.e., $t_i \in \{1, 0\}$), and estimating the Individual Treatment Effect (ITE) of each unit i: $ITE_i = y_i^1 - y_i^0$. From the definition of ITE, there are two potential outcomes y_i^0 and y_i^1 for each unit i, however, dataset $\mathcal{D}$ only contains the observed outcome $y_i^{t_i}$ that corresponds to the treatment t_i, and the outcome of the alternative treatment (a.k.a. counterfactual outcome: $y_i^{1-t_i}$) is missing. This is treated as the counterfactual problem of treatment effect estimation with observational data.

In observational studies, we denote the graphical model among the observed variables X, the treatment T and the outcome Y, shown in Figure 1. Without loss of generality, we assume that the observed variables X can be decomposed into three kinds of latent factors $\{I, C, A\}$ under an unknown joint distribution $Pr(X) = Pr(I, C, A)$, where instrumental factor I only causes the treatment, confounding factor C is the common cause of treatment and outcome, and adjustment factor A only determines the outcome. Then, the treatment T follows the distribution $Pr(T|I, C)$, and the outcome Y follows $Pr(Y|C, A)$. Different from RCT, the treatment T in the observational studies is not randomly assigned to units instead depends on some or all attributes of unit x_i (i.e. the latent factors I and C in Figure 1). This change could result in confounding bias: $Pr(T|X) \neq Pr(T)$.

Without the prior knowledge of confounding factors, previous methods, such as propensity score-based methods [4, 5, 21] and variables balancing methods [3, 28, 9], eliminated the confounding bias by treating all observed variables as confounders for balancing. However, back-door criteria [20] demonstrated that the controlling of the confounding factor is sufficient for removing that bias. Recently, Kuang et.al [16] proposed a data-driven variable decomposition method for separating confounding and adjustment factors, while ignoring the decomposition of instrumental factor, which led to entanglement between instrumental and confounding factors. [11] proposed a disentangled representation learning framework, which separated the observed variables into three sets $\{I, C, A\}$ as shown in Figure 1. However, their algorithm cannot guarantee the separation between the instrumental and the confounding factors (discussed in detail in Section 3.2). Moreover, [25] showed that balancing the instrumental factor would generate additional bias for treatment effect estimation. In summary, precise identification of the three factors is vital for confounder balancing and treatment effect estimation in observational studies.

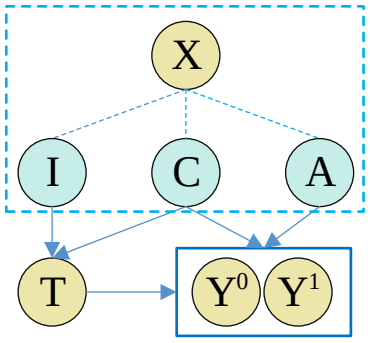


Figure 1: Causal framework with decomposed latent factors. We propose to decompose the observed variables X into three kinds of latent factors: instrumental factor I, which only affect the treatment T; confounding factor C, which is the common cause of treatment T and the outcome Y; and adjustment factor A, which only determine the outcome Y.

With the graphical model illustrated in Figure 1, we generate the following preliminary thoughts to address above problems. (i) Decomposing A from X: (i.a) the adjustment factor A should be independent with the treatment variable T, that is, $A \perp T$; and (i.b) A should predict Y as precisely as possible. Condition (i.a) constraints other factors not be embedded into A, while (i.b) restrains A not be embedded into other factors. (ii) Decomposing I from X: (ii.a) if the confounding factor C is well balanced, one can break the dependency between C and T, and achieve the independence between instrumental factors I and outcome variable Y conditional on the treatment variable T, that is $I \perp Y \mid T$; and (ii.b) I should also predict T as accurately as possible. Condition (ii.a) constraints other factors not be embedded into I, while (ii.b) restrains I not be embedded into other factors. (iii) Predicting factual and counterfactual outcomes $\{y_i^{t_i}, y_i^{1-t_i}\}$: the decomposed representation of confounding factor $C(X)$ and adjustment factor $A(X)$ help to predict both factual $y_i^{t_i}$ and counterfactual outcome $y_i^{1-t_i}$. Inspired by the above thoughts, we propose a synergistic learning algorithm, named Decomposed Representation for CounterFactual Regression (DeR-CFR), to jointly 1) decompose the three latent factors and learn their decomposed representation for confounder identification and balancing, and 2) learn a counterfactual regression model to predict the counterfactual outcome of each unit for individual treatment effect estimation.

Our DeR-CFR algorithm is based on the standard assumptions [14] for treatment effect estimation in observational studies, including stable unit treatment value assumption (SUTVA), unconfoundedness assumption, and overlap assumption. The main contributions in this paper are as follows:

- We solve a critical problem for estimating ITE in observational studies, i.e. the problem of confounder identification and balancing for counterfactual prediction.
- We propose a novel DeR-CFR algorithm to jointly decompose instrumental, confounding, and adjustment factors, and learn counterfactual regression for estimating ITE in observational studies.
- We empirically demonstrate that our algorithm can precisely decompose the latent factors, and the result shows our approach significantly improves the performance of ITE estimation in observational studies with both synthetic and real-world datasets.

2 Related Work

To address the confounding bias in observational studies, most of the previous methods either employ propensity score, including matching, stratification, weighting and doubly robust [18]; or directly optimize sample weight, including entropy balancing [9], residual balancing [3] and stable balancing [28]. Those existing methods focus on confounder balancing alone, while ignoring the importance of confounder identification. Recently, [25] pointed out the necessity of confounder identification/selection for causal inference, due to the fact that the controlling of some non-confounders (e.g., variables related to the instrumental factors) would generate additional bias. Besides, many methods [7, 20, 26] have been proposed for confounder selection, but most of them assume the causal structure is known prior.

In [15], they proposed a representation learning method for confounder balancing by minimizing the distribution difference between different treatment arms in embedding space. Based on this work, [22] proposed to jointly optimize a context-aware importance sampling weight with representation learning. In [27], rather than taking the state-of-the-art ITE estimators balance distribution globally, the authors proposed a local similarity preserving approach for representation learning. In this paper, we propose a decomposed representation learning approach for confounder identification along with a model-free weight schema for confounder balancing.

Our work is related with [16] and [11]. [16] proposed a data-driven variables decomposition algorithm to automatically separate confounder and adjustment factors for treatment effect estimation under a linear setting. The main limitation is they ignored the differentiation between instrumental and confounder factors, leading to imprecise confounder identification and failing to provide estimates of ITE. Aiming at disentangling three latent factors $\{I, C, A\}$ from the observed variables X, [11] proposed a disentangled representations for counterfactual regression. However, the algorithm cannot guarantee the decomposition between I and C. Extremely, $I(X)^\star = \emptyset, C(X)^\star = \{I, C\}, A(X)^\star = A$ could be a possible solution for their algorithm. Moreover, [11] relied on the correct model specification on treatment for confounder balancing. Our proposed algorithm is different from these methods in two ways: (i) Confounder Identification: we propose a series of decomposition regularizers to guarantee the explicit decomposition among the instrumental, confounder and adjustment factors; (ii) Confounder Balancing: we adopt a model-free confounder balancing method to remove the confounding bias in observational data.

3 Factors Decomposition and Representation Learning

In this section, with a brief introduction of some preliminary knowledge regarding variable decomposition, we will present the details of our algorithm.

3.1 Preliminary Propositions

Without loss of generality, we assume that any dataset of the form $\{X, T, Y\}$ is generated from three latent factors $\{I, C, A\}$ as shown in Figure 1. I refers to the instrumental factor that only affects the treatment variable, C denotes the confounding factor which is the common cause of treatment and outcome, A refers to adjustment factor that only determines the outcome. Inspired by

the graphical model, we generate the following preliminary propositions to support decomposition and representation learning of these three latent factors.

1. The adjustment factor would be independent with the treatment variable, that is, $A \perp T$.
2. Under the unconfounderness assumption, controlling confounding factor can help to break the link from confounding factor to treatment variable, that is, $C \perp T$.
3. By controlling the confounding factor, the instrumental factor would become independent with the outcome, given the treatment variable, that is, if $C \perp T$, we have $I \perp Y \mid T$.

Proposition 1 can be easily understood by the definition of adjustment factor, or we can denote the path between adjustment factor and treatment variable as the collider structure at Y: $A \rightarrow Y \leftarrow T$, hence $A \perp T$. Proposition 2 can be guaranteed by the back-door criterion [20]. By controlling the confounder, the path between instrumental factor and outcome can be denoted as $I \rightarrow T \rightarrow Y$, hence $I \perp Y \mid T$ in proposition 3.

Proposition 1 can only constrain that the information of other factors, such as I and C, would not be embedded into A, but A might be embedded into other factors, leading to leaking of A. To address this problem, we propose to simultaneously maximize the predictive power of A on outcome Y.

Similarly, Proposition 3 can only constrain that other factors would not be embedded into I, but cannot guarantee that the information of I would not be represented into other factors. In our context, we propose to jointly maximize the predictive power of I on treatment T for decomposing I.

By decomposing I and A from X, we can also achieve the identification of confounder C. Then, with the decomposed C and A, we can estimate the ITE via potential outcomes regression.

3.2 DeR-CFR Algorithm

Inspired by the above preliminary results and thoughts, we propose a novel model to learn the decomposed representation of instrumental, confounding and adjustment factors for confounder identification and balancing, and simultaneously learn a counterfactual regression model for treatment effect estimation. The architecture of our model consists of the following components:

- Three decomposed representation networks for learning latent factors, one for each underlying factor: $I(X)$, $C(X)$ and $A(X)$.
- Three decomposition and balancing regularizers for confounder identification and balancing: the first is for decomposing A from X with considering $A(X) \perp T$ and $A(X)$ should predict Y as precisely as possible; the second is for decomposing I from X via constraining $I(X) \perp Y \mid T$, and $I(X)$ should be predictive to T; the last is designed for simultaneously balancing confounder $C(X)$ in different treatment arms.
- Two regression networks for potential outcome prediction, one for each treatment arm: $h^0(C(X), A(X))$ and $h^1(C(X), A(X))$.

The core components in our model are the decomposition and balancing regularizers, which help the representation network to learn decomposed representation of I, C, and A for confounder identification, and also to improve the precision of regression networks via accurate confounder balancing with identified C. The decomposition and balancing regularizers are the keys to bridge the representation networks and regression networks for ITE estimation with observational data.

Our framework is inspired by recent work [10], but extends this work in two ways: (i) [10] tried to learn disentangled representation of three underlying factors, but it has no explicit constraints to prevent the information of I and A from being embedded into confounding factor C, extremely $I^\star = \emptyset, C^\star = \{I, C, A\}, A^\star = \emptyset$ could also be a possible solution for its algorithm, which loses the meaning of disentanglement. To address this problem, we propose a joint decomposition and balancing regularizers to explicitly decompose instrumental factor and adjustment factor for confounder identification and balancing with a guarantee from preliminary results and thoughts in Section 3.1. (ii) When doing confounder balancing, [10] needs to estimate the propensity score with two logistic networks, and its performance depends on the correct specification of the propensity score model. In our model, we propose a model-free confounder balancing methods via directly learning sample weights for minimizing the discrepancy of confounder between different treatment arms.

Next, we will describe each component of our DeR-CFR algorithm in detail.

3.2.1 Decomposing Adjustment Factor

From the preliminary proposition, we know the adjustment factor should be independent with the treatment variable, $A(X) \perp T$. Considering the treatment is binary, then we propose to learn the decomposed representation of adjustment factor A by constraining the discrepancy of its distribution between treatment arms $T = 1$ and $T = 0$. Moreover, to prevent the information of adjustment factor from being embedded into other factors, we adopt a regression model g_A to maximize the predictive power of $A(X)$ on Y. Here, we use $\mathcal{L}_A$ to denote the loss of decomposing adjustment factor during the representation learning as:

$$\mathcal{L}_A = \text{disc}\left(\{A\left(x_i\right)\}_{i:t_i=0}, \{A\left(x_i\right)\}_{i:t_i=1}\right) + \textstyle\sum_i l[y_i, g_A(A(x_i))] \tag{1}$$

where $l[y_i, g_A(A(x_i))]$ would be an l_2-loss for continuous outcomes and a log-loss for binary outcomes. $\{A\left(x_i\right)\}_{i:t_i=k}$ denotes the distribution of adjustment factor representation $A(X)$ with respect to the treatment arm $t = k$. $\text{disc}(\cdot)$ denotes the discrepancy of adjustment factor distribution between different treatment arms. Many integral probability metrics (IPMs)[19, 24], such as Maximum Mean Discrepancy (MMD) [8] and Wasserstein distance[2], can be used to measure the discrepancy of distributions. In this paper, we use the MMD to calculate $disc(\cdot)$.

By minimizing this term, our model can ensure the information of instrumental factor I and confounding factor C would not be embedded into $A(X)$, since I and C are associated with the treatment variable. And vice versa. Hence, the regularizer can help to decompose the adjustment factor.

3.2.2 Decomposing Instrumental Factor and Balancing Confounder

From preliminary propositions, we know that if one can control/balance the confounding factor, the instrumental factor would be independent with the outcome variable conditional on the treatment variable.

Firstly, we introduce the loss function of confounder balancing in our model. Most previous work [4, 10, 21] achieved confounder balancing by learning propensity score and their performance relied on the correctness of the specified propensity score model. Here, we propose to adopt a model-free method for confounder balancing. The purpose of confounder balancing is to break the link from confounding factor C to the treatment variable T, that is, to make $C(X)$ become independent with T. Assuming that we have the decomposed representation of confounding factor $C(X)$, we propose to achieve confounder balancing by directly learning sample weight ω with minimizing the following objective function:

$$\mathcal{L}_{C\_B} = \text{disc}\left(\{\omega_i \cdot C\left(x_i\right)\}_{i:t_i=0}, \{\omega_i \cdot C\left(x_i\right)\}_{i:t_i=1}\right) \tag{2}$$

where $\{\omega_i \cdot C\left(x_i\right)\}_{i:t_i=0}$ refers to the weighted distribution of $C(X)$ on the samples with $t = 0$. To avoid all the sample weights to be $zero$, we constrain the sample weight $\sum_{i:t_i=0} \omega_i = \sum_{i:t_i=1} \omega_i = 1$. If $\mathcal{L}_{C\_B}$ can be minimized to be $zero$, one can achieve the independence between $C(X)$ and T by weighting samples with the learned weight ω.

Based on the sample weight ω, then, we can decompose the instrumental factor by conditional independence $I(X) \perp Y \mid T$. Moreover, to prevent the information of instrumental factor from being embedded into other factors, we adopt a regression model g_I to maximize the predictive power of $I(X)$ on T. Then, the objective function, denoted as $\mathcal{L}_I$ for decomposing instrumental factor is:

$$\mathcal{L}_I = \textstyle\sum_{k=\{0,1\}} \text{disc}\left(\{\omega_i \cdot I\left(x_i\right)\}_{i:y_i=0}, \{\omega_i \cdot I\left(x_i\right)\}_{i:y_i=1}\right)_{i:t_i=k} + \sum_i l[t_i, g_I(I(x_i))] \tag{3}$$

where $\text{disc}(\{\omega_i \cdot I\left(x_i\right)\}_{i:y_i=0}, \{\omega_i \cdot I\left(x_i\right)\}_{i:y_i=1})_{i:t_i=k}$ constrains the learned representation of instrumental factor I to be independent with the outcome Y given the treatment arm $t = k$ and sample weight ω. Here, we assume the outcome variable is binary, i.e., $y_i \in \{0, 1\}$. For continuous or multi-valued outcome, we can approximately achieve the conditional independence $I(X) \perp Y \mid T$ by making outcome become binary during the process of minimizing $\mathcal{L}_I$.

By minimizing the term $\mathcal{L}_I$, our model can ensure the information of confounding factor C and adjustment factor A would not be embedded into $I(X)$, since C and A are associated with the outcome even given the treatment variable. And vice versa with maximizing the predictive power of $I(X)$ on T. Hence, this regularizer can help to decompose the instrumental factor.

3.2.3 Orthogonal Regularizer for Hard Decomposition

Inspired by the orthogonal regularizer in [16] for variable decomposition, in this paper, we employ an orthogonal regularizer among the three representation networks for factors $\{I, C, A\}$ decomposing. Taking the representation network for instrumental factor I as an example. Assuming it is with l layers and let W_k refer to the weight matrix on k^{th} layer of the network. Then, we can approximate the contribution of each variable in X on each dimension of representation $I(X)$ by computing $W_1 \times W_2 \times \cdots \times W_l$, denoted as $W_I \in \mathbb{R}^{m \times d}$, where m and d refer to the dimension of X and $I(X)$, respectively. By averaging each row of W_I, we can obtain $\bar{W}_I \in \mathbb{R}^m$, denoting the average contribution of each variable in X on the representation $I(X)$. Similarity, we can calculate the contribution of each variable in X on $C(X)$ and $A(X)$, denoted as $\bar{W}_C$ and $\bar{W}_A$.

We consider the three representation networks have the same structure. Hence, $\bar{W}_I$, $\bar{W}_C$ and $\bar{W}_A$ are with the same dimension. Then, we propose to achieve hard decomposition by constraining orthogonality on each pair of them. The loss is as follow:

$$\mathcal{L}_O = \bar{W}_I^T \cdot \bar{W}_C + \bar{W}_C^T \cdot \bar{W}_A + \bar{W}_A^T \cdot \bar{W}_I \tag{4}$$

To prevent the representation network from rejecting any input, we constrain the sum of each $\bar{W}_I$, $\bar{W}_C$ and $\bar{W}_A$ to be 1. The orthogonal regularizer ensures the information of each variable in X can only flow into one representation network for a hard decomposition.

3.2.4 Outcome Regression

With the decomposed representations, we propose to learn the outcome regression model for estimating the ITE. Similar to [15, 22, 11], we also train two regression networks for each treatment arm, h^0 and h^1, based on the observed outcomes of samples with $t_i = 0$ and $t_i = 1$, respectively. As guided by the graphical model in Figure 1, we can train these regression models only based on the decomposed representation of $C(X)$ and $A(X)$.

$$\mathcal{L}_R = \textstyle\sum_i \omega_i \cdot l\left[y_i, h^{t_i}\left(C\left(x_i\right), A\left(x_i\right)\right)\right] \tag{5}$$

where the sample weight ω is learned from confounder balancing.

3.2.5 Objective Function

Therefore, we try to minimize the following objective function in our DeR-CFR algorithm:

$$\mathcal{L} = \mathcal{L}_R + \alpha \cdot \mathcal{L}_A + \beta \cdot \mathcal{L}_I + \gamma \cdot \mathcal{L}_{C\_B} + \mu \cdot \mathcal{L}_O + \lambda \cdot Reg \tag{6}$$

where Reg refers to the regularization term on our model parameters (see supplementary for details).

We adopt an alternating training strategy to iteratively optimize the representation for confounder identification and sample weight for confounder balancing as:

$$\mathcal{L}_{-\omega} = \mathcal{L}_R + \alpha \cdot \mathcal{L}_A + \beta \cdot \mathcal{L}_I + \mu \cdot \mathcal{L}_O + \lambda \cdot Reg \tag{7}$$

$$\mathcal{L}_{\omega} = \mathcal{L}_R + \gamma \cdot \mathcal{L}_{C\_B} + \lambda \cdot Reg \tag{8}$$

We minimize $\mathcal{L}_{-\omega}$ using stochastic gradient descent to update the parameters of the representation and hypothesis network, and minimize $\mathcal{L}_{\omega}$ to update ω. The details of pseudo-code and hyper-parameters of our algorithm are provided in the supplementary material.

4 Experiments

4.1 Baselines

We compare the proposed algorithm (**DeR-CFR**) with the following baselines. (1) **CFR-MMD** and **CFR-WASS** [15, 22]: CounterFactual Regression with MMD and Wasserstein metrics; (2) **CFR-ISW** [10]: CounterFactual Regression with Importance Sampling Weights; (3) **SITE** [27]: local Similarity preserved Individual Treatment Effect estimator; and (4) **DR-CFR** [11]: Disentangled Representations for CounterFactual Regression.

4.2 Experiments on Real Dataset

Dataset. The original Randomized Controlled Trial (RCT) data of the Infant Health and Development Program (IHDP) aims to evaluate the effect of a specialist home visits on the future cognitive test scores of premature infants. Hill [12] removed a non-random subset of the treated group and induced selection bias. The dataset comprises 747 units (139 treated, 608 control) with 25 pre-treatment variables related to the children and their mothers. We report the estimation errors on the same benchmark (100 realizations of the outcomes with 63/27/10 proportion of train/validation/test splits) provided by and used in [15, 22, 11]. Experiments on other real datasets, including Jobs [23] and Twins [1], can be found in supplementary.

Results. similar to [15, 22, 11], we adopt the Precision in Estimation of Heterogeneous Effect (PEHE) [12] as the individual-level performance metric, where PEHE =$\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left((\hat{y}_i^1-\hat{y}_i^0)-(y_i^1-y_i^0)\right)^2}$. For population-level, we adopt the bias of the Average Treatment Effect $\epsilon_{\text{ATE}}=|ATE-\widehat{ATE}|$ to evaluate performance, where ATE = $\mathbb{E}(y^1)-\mathbb{E}(y^0)$.

Table 1: Results (mean $\pm$ std) on IHDP.

IHDP				
Mean $\pm$ Std	**Within-sample**		**Out-of-sample**	
Methods	PEHE	ϵ_{ATE}	PEHE	ϵ_{ATE}
CFR-MMD	0.702 ± 0.037	0.284 ± 0.036	0.795 ± 0.078	0.309 ± 0.039
CFR-WASS	0.702 ± 0.034	0.306 ± 0.040	0.798 ± 0.088	0.325 ± 0.045
CFR-ISW	0.598 ± 0.028	0.210 ± 0.028	0.715 ± 0.102	0.218 ± 0.031
SITE	0.609 ± 0.061	0.259 ± 0.091	1.335 ± 0.698	0.341 ± 0.116
DR-CFR	0.657 ± 0.028	0.240 ± 0.032	0.789 ± 0.091	0.261 ± 0.036
DeR-CFR	$\mathbf{0.444 \pm 0.020}$	$\mathbf{0.130 \pm 0.020}$	$\mathbf{0.529 \pm 0.068}$	$\mathbf{0.147 \pm 0.022}$

Table 2: Results (mean $\pm$ std) of ablation studies on our DeR-CFR algorithm.

$\mathcal{L}_A$	$\mathcal{L}_I$	$\mathcal{L}_{C\_B}$	$\mathcal{L}_O$	PEHE	
				Within-sample	Out-of-sample
✓	✓	✓	✓	$\mathbf{0.444 \pm 0.020}$	$\mathbf{0.529 \pm 0.068}$
✓	✓	✓		0.478 ± 0.033	0.542 ± 0.053
✓	✓		✓	0.482 ± 0.039	0.565 ± 0.075
✓		✓	✓	0.479 ± 0.030	0.560 ± 0.071
	✓	✓	✓	0.635 ± 0.035	0.858 ± 0.133

We report the results, mean and standard deviation (std), of 100 replications on IHDP in Table 1 and 2. Tabel 1 shows that DeR-CFR outperforms all the baselines and achieves a significant improvement on PEHE and ϵ_{ATE} measures. Table 2 investigates the effects of each module of the DeR-CFR by conducting ablation experiments. From Tabel 1 and Table 2, we can draw the following conclusions: (i) With explicitly learning the decomposed representation, DeR-CFR achieves better performance than DR-CFR, which cannot guarantee the disentanglement of different factors. (ii) Each component in our DeR-CFR is necessary, since missing any one of them would bring confusion on the decomposed representation learning and damage the performance on ITE estimation.

4.3 Experiments on Synthetic Dataset

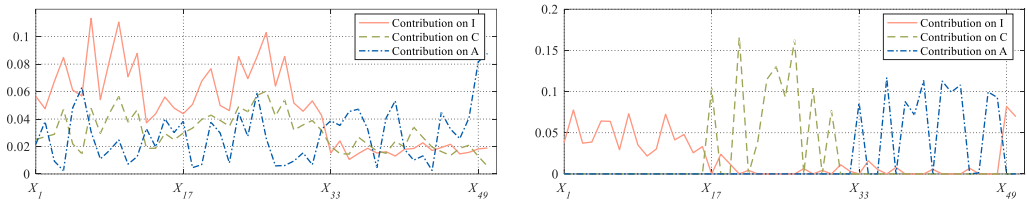
Dataset. To generate synthetic datasets, we design two different sample sizes $n=\{3000, 10000\}$ and two dimensional setting $\{m_I, m_C, m_A\}$=$\{8,8,8\}$ or $\{16,16,16\}$, where m_I, m_C, and m_A denote the dimensions of instrumental variables, confounding variables and adjustment variables, respectively. Thus, the total dimension of observed variables is $m=m_I+m_C+m_A+m_D$, where $m_D=2$ denotes two noise variables. We generate samples from independent Normal distributions $X_1, X_2, \cdots, X_m \sim \mathcal{N}(0,1)$. Then we create selection bias by $t=binomial(1, 1/(1+e^{-z}))$, where $z=\frac{1}{10}\theta_t \times X_{IC}+\varepsilon$, X_{IC} denotes the variables in X that belongs to I and C. The outcomes corresponding to different treatment arms are $y^0=\text{sign}\left(\max\left(0, z^0-\bar{z}^0\right)\right)$ and $y^1=\text{sign}\left(\max\left(0, z^1-\bar{z}^1\right)\right)$, where $z^0=\frac{1}{10}\frac{\theta_{y0}\times X_{CA}}{m_C+m_A}$ and $z^1=\frac{1}{10}\frac{\theta_{y1}\times X_{CA}^2}{m_C+m_A}$. In addition, $\theta_t \sim \mathcal{U}\left((8,16)^{m_I+m_C}\right)$, $\theta_{y0}, \theta_{y1} \sim \mathcal{U}\left((8,16)^{m_C+m_A}\right), \varepsilon \sim \mathcal{N}(0,1)$. We use Syn_$m_I$_$m_C$_$m_A$_$n$ to denote different experimental settings. In each setting, we do experiments with 10 replications, and report the mean and standard deviation (std) on PEHE and ϵ_{ATE}.

Results. We compare the proposed method with the contending baselines under different synthetic settings, and report the results in Table 3. We see that DeR-CFR outperforms other state-of-the-art methods in PEHE and ϵ_{ATE}. Moreover, with constraining the explicit decomposition during representation learning, the performance of DeR-CFR is much better than DR-CFR.

To evaluate the performance of decomposition representation learning, we calculate the average contribution of each variable in X on the representation of each factor, i.e., $\bar{W}_I, \bar{W}_C, \bar{W}_A \in \mathbb{R}^m$ as described in Section 3.2.3. Figure 2 reports the results under setting Syn_16_16_16_3000, more results can be found in supplementary. It's obvious in Figure 2 that our DeR-CFR algorithm can precisely identify the three underlying factors, while the baseline DR-CFR fails to disentangle those

Table 3: Results (mean ± std) on Synthetic Data under different settings Syn_m_I_m_C_m_A_n.

Within-sample								
Datasets	**Syn_8_8_8_3000**		**Syn_8_8_8_10000**		**Syn_16_16_16_3000**		**Syn_16_16_16_10000**	
Methods	PEHE	ϵ_{ATE}	PEHE	ϵ_{ATE}	PEHE	ϵ_{ATE}	PEHE	ϵ_{ATE}
CFR-MMD	0.384 ± 0.004	**0.015 ± 0.006**	0.276 ± 0.004	**0.008 ± 0.003**	0.491 ± 0.005	0.021 ± 0.008	0.399 ± 0.005	0.012 ± 0.005
CFR-WASS	0.378 ± 0.004	0.016 ± 0.006	0.277 ± 0.004	0.008 ± 0.002	0.513 ± 0.007	0.011 ± 0.005	0.408 ± 0.005	0.015 ± 0.005
CFR-ISW	0.383 ± 0.005	0.035 ± 0.007	0.279 ± 0.004	0.013 ± 0.002	0.538 ± 0.003	0.014 ± 0.005	0.441 ± 0.005	0.034 ± 0.005
SITE	0.550 ± 0.007	0.075 ± 0.013	0.497 ± 0.006	0.035 ± 0.012	0.585 ± 0.005	0.035 ± 0.012	0.608 ± 0.006	0.041 ± 0.014
DR-CFR	0.377 ± 0.002	0.027 ± 0.008	0.288 ± 0.005	0.022 ± 0.007	0.544 ± 0.004	0.023 ± 0.010	0.427 ± 0.015	0.043 ± 0.019
DeR-CFR	**0.327 ± 0.002**	0.020 ± 0.006	**0.235 ± 0.002**	0.012 ± 0.002	**0.404 ± 0.003**	**0.011 ± 0.004**	**0.307 ± 0.002**	**0.006 ± 0.002**
Out-of-sample								
Datasets	**Syn_8_8_8_3000**		**Syn_8_8_8_10000**		**Syn_16_16_16_3000**		**Syn_16_16_16_10000**	
Methods	PEHE	ϵ_{ATE}	PEHE	ϵ_{ATE}	PEHE	ϵ_{ATE}	PEHE	ϵ_{ATE}
CFR-MMD	0.465 ± 0.006	0.062 ± 0.021	0.327 ± 0.006	0.021 ± 0.008	0.574 ± 0.007	0.036 ± 0.012	0.463 ± 0.006	**0.018 ± 0.006**
CFR-WASS	0.469 ± 0.011	0.063 ± 0.021	0.320 ± 0.006	0.016 ± 0.007	0.553 ± 0.006	0.028 ± 0.009	0.469 ± 0.005	0.018 ± 0.007
CFR-ISW	0.461 ± 0.005	0.058 ± 0.021	0.334 ± 0.006	0.017 ± 0.007	0.553 ± 0.006	0.034 ± 0.012	0.501 ± 0.005	0.040 ± 0.007
SITE	0.561 ± 0.005	0.077 ± 0.020	0.506 ± 0.006	0.021 ± 0.009	0.588 ± 0.007	0.050 ± 0.016	0.612 ± 0.009	0.049 ± 0.013
DR-CFR	0.469 ± 0.011	0.063 ± 0.024	0.333 ± 0.006	0.030 ± 0.009	0.551 ± 0.008	0.037 ± 0.014	0.486 ± 0.011	0.044 ± 0.019
DeR-CFR	**0.409 ± 0.009**	**0.046 ± 0.017**	**0.286 ± 0.007**	**0.012 ± 0.006**	**0.485 ± 0.006**	**0.028 ± 0.010**	**0.378 ± 0.006**	0.019 ± 0.007

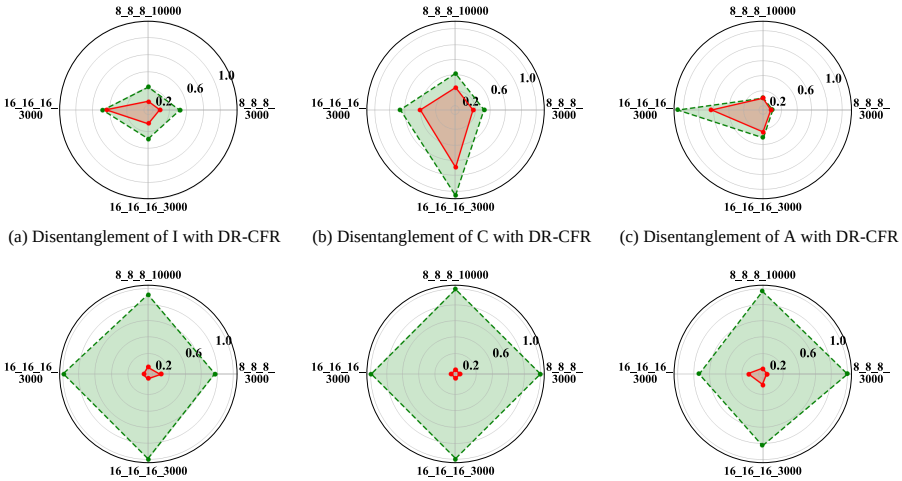


(a) DR-CFR in Syn_16_16_16_3000 (b) DeR-CFR in Syn_16_16_16_3000

Figure 2: Visualization of the contribution of each variable in X on the decomposed representation of I, C and A under the setting Syn_16_16_16_3000, where $X_I = \{X_1 \cdots, X_{16}\}$, $X_C = \{X_{17} \cdots, X_{32}\}$ and $X_A = \{X_{33} \cdots, X_{48}\}$ are the true underlying factors of I, C and A.

factors. This result vilifies the motivation of the proposed DeR-CFR, and is coinsident with our analyses on the comparison of DeR-CFR and DR-CFR algorithms in Section 3.2.

Similar to the setting in DR-CFR [11], we also plot the radar charts on the representation of each factor for compassion in Figure 3. For example, in Figure 3(a), we calculate the average contribution of true factors of I in X, i.e., $X_I = \{X_1, \cdots, X_{16}\}$ on the representation of I (plotted with



(a) Disentanglement of I with DR-CFR (b) Disentanglement of C with DR-CFR (c) Disentanglement of A with DR-CFR

(d) Decomposition of I with DeR-CFR (e) Decomposition of C with DeR-CFR (f) Decomposition of A with DeR-CFR

Figure 3: Radar charts that visualize the disentangled/decomposed representation of all three underlying factors from DR-CFR (sub-figures a,b,c) and DeR-CFR (sub-figures d,e,f) methods. Each vertex on the polygons denotes an experimental setting with form Syn_m_I_m_C_m_A_n. The green and red plots denote the average contribution of true variables and other variables in X on the representation of each factor, respectively

dotted green), compared with the average contribution of other variables in X, i.e., $X \backslash X_I = \{X_{17}, \cdots, X_{48}\}$ on the representation of I (plotted with red) under different settings. From the results, we can conclude that with explicit decomposition regularizers, our DeR-CFR achieves a much better decomposed/disentangled representation of all three underlying factors than DR-CFR. This is the key reason that our DeR-CFR can obtain significant improvement in ITE estimation than DR-CFR as shown in Tables 1 and 3.

5 Conclusion

In this paper, we focus on the problem of estimating individual treatment effect in observational studies. We argue that previous methods mainly focus on confounder balancing, while ignoring the importance of confounder identification. Although some promising algorithms have been proposed for confounder separation/disentanglement, they cannot guarantee the decomposition of instrumental factors and confounding factors. Hence, we propose a Decomposed Representation learning algorithm for CounterFactual Regression (DeR-CFR) with explicit decomposition regularizers for confounder identification and balancing, and simultaneously estimate the ITE via counterfactual inference. Empirical results demonstrate the advantages of DeR-CFR algorithm compared with state-of-the-art methods.

Broader Impact

This paper proposes a decomposed representation learning method for estimating the individual treatment effect in observational studies. The proposed method can first benefit the analysts who need to compare different treatments for decision making with observational data. For example, it can help social marketers to estimate the causal effect of different advertising strategies for choosing a better one to apply without online A/B testing. Our method is proposed to address the confounding bias in data for treatment effect estimation. One limitation of the proposed algorithm is that it only focuses on estimating the causal effect of a binary treatment variable, although previous works also mainly focus on binary treatment.

The effectiveness of the proposed algorithm relies on the standard assumption in causal inference [14], such as SUTVA, unconfounderness and overlap assumptions. When those assumptions are violated in real applications, the proposed algorithm cannot be guaranteed to precisely estimate the treatment effect, but might still reduce the confounding bias to some degree.

References

[1] D. Almond, K. Y. Chay, and D. S. Lee. The costs of low birth weight. *The Quarterly Journal of Economics*, 120(3):1031–1083, 2005.

[2] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[3] S. Athey, G. W. Imbens, and S. Wager. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):597–623, 2018.

[4] P. C. Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424, 2011.

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

[6] J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(Feb):281–305, 2012.

[7] M. A. Brookhart, S. Schneeweiss, K. J. Rothman, R. J. Glynn, J. Avorn, and T. Stürmer. Variable selection for propensity score models. *American journal of epidemiology*, 163(12):1149–1156, 2006.

[8] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

[9] J. Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46, 2012.

[10] N. Hassanpour and R. Greiner. Counterfactual regression with importance sampling weights. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5880–5887, 2019.

[11] N. Hassanpour and R. Greiner. Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*, 2020.

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

[13] P. W. Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

[14] G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

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

[16] K. Kuang, P. Cui, B. Li, M. Jiang, S. Yang, and F. Wang. Treatment effect estimation with data-driven variable decomposition. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[17] R. J. LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, pages 604–620, 1986.

[18] S. L. Y. L. J. G. A. Z. Liuyi Yao, Zhixuan Chu. A survey on causal inference. *arXiv preprint arXiv:2002.02770*, 2020.

[19] Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009.

[20] J. Pearl. *Causality*. Cambridge university press, 2009.

[21] P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[22] U. Shalit, F. D. Johansson, and D. Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3076–3085. JMLR. org, 2017.

[23] J. A. Smith and P. E. Todd. Does matching overcome lalonde's critique of nonexperimental estimators? *Journal of econometrics*, 125(1-2):305–353, 2005.

[24] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. Lanckriet. On integral probability metrics,\phi-divergences and binary classification. *arXiv preprint arXiv:0901.2698*, 2009.

[25] T. J. VanderWeele. Principles of confounder selection. *European journal of epidemiology*, 34(3):211–219, 2019.

[26] T. J. VanderWeele and I. Shpitser. A new criterion for confounder selection. *Biometrics*, 67(4):1406–1413, 2011.

[27] L. Yao, S. Li, Y. Li, M. Huai, J. Gao, and A. Zhang. Representation learning for treatment effect estimation from observational data. In *Advances in Neural Information Processing Systems*, pages 2633–2643, 2018.

[28] J. R. Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922, 2015.

Appendix

The Appendix is organized as follows. Section A and B describe the background knowledge of the potential outcome framework and related assumptions. Section C details how to apply orthogonal regularizer to continuous or multi-valued outcome. Section D gives the regularization term on DeR-CFR parameters. In Section E, we present the pseudo-code of DeR-CFR. In Section F, we summarizes the search method of the hyper-parameters. In Section G, we analyze the results of various models on real-world datasets. Finally, in Section H, we show the identification performance of DeR-CFR and DR-CFR.

A The Potential Outcomes Framework

Under the unconfounderness assumption, the potential outcomes $\{y^0, y^1\}$ are thought to be fixed for each unit. Obviously, the assigned treatment t would determine which potential outcome is observed, and another potential outcome y^{1-t} is always missing (i.e., counterfactual). That is:

$$y_i^F = y_i^{t_i} = t_i y_i^1 + (1 - t_i) y_i^0 \tag{9}$$

$$y_i^{CF} = y_i^{1-t_i} = (1 - t_i) y_i^1 + t_i y_i^0 \tag{10}$$

Hence, the core of the potential outcome framework is to estimate such potential outcomes of corresponding treatment arms i (Table 4) and then predict the treatment effect.

Table 4: The Potential Outcomes Framework.

Group	y^1	y^0
Treated ($t = 1$)	y^F (Observed)	y^{CF} (Potential)
Control ($t = 0$)	y^{CF} (Potential)	y^F (Observed)

B Assumptions

Assumption 1: The Stable Unit Treatment Value Assumption. The potential outcomes for one unit are unaffected by the treatment assignment of other units, and, for each unit, there are no different forms of each treatment level, which lead to different potential outcomes.

The assumption can be summarized as two points: independence assumption and certain treatment assumption. The independence assumption means that the outcome of one unit after accept treatment will not be affected by the treatment of the other units; the certain treatment assumption is that treatment must be certain and cannot be broken down into smaller levels of treatment.

Assumption 2: The Ignorable Treatment Assignment Assumption. Given variables, treatment assignment strategy does not affect potential outcomes.

If the variables x are consistent, two units would have the equal probability of being assigned to each treatment arm with the treatment assignment strategy. And, their potential outcomes should be consistent whatever the treatment assignment strategy is.

Assumption 3: The Consistency Assumption. The observed treatment (control) outcomes of the units are consistent with the potential treatment (control) outcomes.

The treatment assignment strategy will determine which treatment arm units in the observation data are assigned to, and then each unit has a treatment assignment and outcome pair. For the assigned treatment, the potential outcome of the unit is consistent with the observed outcome.

Assumption 2 and 3 are also named as **Unconfoundedness Assumption**, that means all the confounders that contribute to both treatment assignment and outcomes are in the observation dataset.

Assumption 4: The Overlap Assumption. For each unit x_i, the probability of being assigned to each treatment arm is positive.

If for some units x_i, the treatment assignment is deterministic; then for these units, one of their treatment outcomes and control outcomes could never be observed. If so, it would be unable and meaningless to estimate the treatment effect.

C For Continuous or Multi-valued Outcome.

In Section 3.2.2, we achieve the independence between $C(X)$ and T by weighting samples with the learned weight ω, that is, $\omega \cdot C(X) \perp T$. Based on standard assumptions and the sample weight ω, we can make the outcome variable independent of the instrumental factor conditional on the treatment variable, $\omega \cdot I(X) \perp Y \mid T$. The objective function $\mathcal{L}_I$ for decomposing instrumental factor is:

$$\mathcal{L}_I = \sum_{k=\{0,1\}} \operatorname{disc}\left(\{\omega_i \cdot I\left(x_i\right)\}_{i:y_i=0}, \{\omega_i \cdot I\left(x_i\right)\}_{i:y_i=1}\right)_{i:t_i=k} + \sum_i l[t_i, g_I(I(x_i))] \quad (11)$$

where $\operatorname{disc}(\{\omega_i \cdot I\left(x_i\right)\}_{i:y_i=0}, \{\omega_i \cdot I\left(x_i\right)\}_{i:y_i=1})_{i:t_i=k}$ constrains the learned representation of instrumental factor I to be independent with the outcome Y given the treatment arm $t = k$ and sample weight ω. Here, we assume the outcome variable is binary, i.e., $y_i \in \{0, 1\}$.

For continuous or multi-valued outcome, we can approximately achieve the conditional independence $I(X) \perp Y \mid T$ by making outcome binary during the process of minimizing $\mathcal{L}_I$:

$$y_i := \begin{cases} 0, & y_i < median(\{y_i\}_{i:t_i=k}) \\ 1, & y_i \geq median(\{y_i\}_{i:t_i=k}) \end{cases} \quad i : t_i = k \quad (12)$$

where $median(\{y_i\}_{i:t_i=k})$ refers to the median of factual outcome $\{y_i\}_{i:t_i=k}$ on the treatment arm $t = k$.

If the instrumental factor is completely independent of the outcome variable Y given treatment T, then we divide the units into two parts based on the outcomes y_i. For the instrumental variable, the distribution of the two parts should be similar to the result of random assignment, based on the law of large numbers. Without loss of generality, we use the median to divide the dataset based on the outcomes y_i. Similarly, other conditional division methods are also applicable.

D The Regularization Term on DeR-CFR Parameters

We try to minimize the following objective function in our DeR-CFR algorithm (Section 3.2.3):

$$\mathcal{L} = \mathcal{L}_R + \alpha \cdot \mathcal{L}_A + \beta \cdot \mathcal{L}_I + \gamma \cdot \mathcal{L}_{C\_B} + \mu \cdot \mathcal{L}_O + \lambda \cdot Reg \quad (13)$$

The regularization term Reg on DeR-CFR parameters is:

$$Reg = \mathcal{R}_{\mathcal{W}} + \mathcal{R}_{C\_B} + \mathcal{R}_O \quad (14)$$

where $\mathcal{R}_{\mathcal{W}}$ is the l_2 regularization on the parameters of networks $\{I, C, A, h^0, h^1, g_I, g_A\}$. $\mathcal{R}_{C\_B}$ restricts the sample weight ω to not be all $zero$. To prevent the representation network from rejecting any input, we use $\mathcal{R}_O$ to constrain the sum of each $\bar{W}_I$, $\bar{W}_C$, and $\bar{W}_A$ to be 1.

Next, we describe each component of Reg in detail.

D.1 The regularization on the network parameters.

$$\mathcal{R}_{\mathcal{W}} = l_2\left(\mathcal{W}(I, C, A, h^0, h^1, g_I, g_A)\right) \quad (15)$$

The regularization term is generally a monotonically increasing function of the model complexity. We believe that the model will have lower complexity and better robustness, when the parameter value of the model is small enough. To prevent overfitting, we penalize the large value in the network parameters $\mathcal{W}(I, C, A, h^0, h^1, g_I, g_A)$ by l2 regularization.

D.2 The regularization on the sample weight.

$$\mathcal{R}_{C\_B} = \left(\sum_{i:t_i=0} \omega_i - 1\right)^2 + \left(\sum_{i:t_i=1} \omega_i - 1\right)^2 \tag{16}$$

To avoid all the sample weights to be $zero$ and maintain quantity allocation on each treatment arm, we constrain the sample weight $\sum_{i:t_i=0} \omega_i = \sum_{i:t_i=1} \omega_i = 1$. As an extreme example, there are only two samples in the treated group, whose ages are 0 and 100 years old, and the two samples in the control group are 20 and 80 years old, respectively. With sample weights $(100, 100)$, the weighted average age of the treated group is 50 years old; sample weights were assigned to the control group (1, 1), the average age of the control group is also 50 years old. In the process of training, the treated group contains 200 samples, while the control group contains only 2 samples, which leads to overfitting. Therefore, it is necessary to softly constrain the sample weights and punish them for not deviating from 1 too much.

D.3 The regularization on the orthogonal regularizer.

In Section 3.2.3, we propose to achieve hard decomposition by constraining orthogonality on each pair of them. The loss is as follow:

$$\mathcal{L}_O = \bar{W}_I^T \cdot \bar{W}_C + \bar{W}_C^T \cdot \bar{W}_A + \bar{W}_A^T \cdot \bar{W}_I \tag{17}$$

By minimizing $\mathcal{L}_O$, on each dimension k, there are at most one element in $\{\bar{W}_I^k, \bar{W}_C^k, \bar{W}_A^k\}$ can not be close to 0, which is equivalent to an orthogonal regularize. However, this constraint may also lead to the result $\bar{W}_I^k = \bar{W}_C^k = \bar{W}_A^k = 0$ for all dimension k. If we strictly require that at least one in $\{\bar{W}_I^k, \bar{W}_C^k, \bar{W}_A^k\}$ is greater than 0, then we will introduce bias in the dimensions corresponding to irrelevant factor, which is independent of both treatment and outcome. Therefore, to prevent the representation network from rejecting any input, we constrain the sum of each $\bar{W}_I$, $\bar{W}_C$ and $\bar{W}_A$ to be 1.

$$\mathcal{R}_O = \left(\sum_{k=1}^m \bar{W}_I^k - 1\right)^2 + \left(\sum_{k=1}^m \bar{W}_C^k - 1\right)^2 + \left(\sum_{k=1}^m \bar{W}_A^k - 1\right)^2 \tag{18}$$

E Pseudo-Code of DeR-CFR

We try to minimize the following objective function in our DeR-CFR algorithm:

$$\mathcal{L} = \mathcal{L}_R + \alpha \cdot \mathcal{L}_A + \beta \cdot \mathcal{L}_I + \gamma \cdot \mathcal{L}_{C\_B} + \mu \cdot \mathcal{L}_O + \lambda \cdot Reg \tag{19}$$

where Reg refers to the regularization term on our model parameters.

We adopt an alternating training strategy to optimize the algorithm for avoiding the effects of sample weight on representation learning:

$$\mathcal{L}_{-\omega} = \mathcal{L}_R + \alpha \cdot \mathcal{L}_A + \beta \cdot \mathcal{L}_I + \mu \cdot \mathcal{L}_O + \lambda \cdot Reg \tag{20}$$

$$\mathcal{L}_{\omega} = \mathcal{L}_R + \gamma \cdot \mathcal{L}_{C\_B} + \lambda \cdot Reg \tag{21}$$

We minimize $\mathcal{L}_{-\omega}$ using stochastic gradient descent to update the parameters of the representation and hypothesis network and minimize $\mathcal{L}_{\omega}$ to update ω. Algorithm 1 shows the details of the pseudo-code of DeR-CFR.

Algorithm 1 Decomposed Representation for CounterFactual Regression

Input: Observational data $\left\{x_i, t_i, y_i^F\right\}_{i-1}^N$
Output: $\hat{y}_0, \hat{y}_1$
Loss function: $\mathcal{L}_{-\omega}$ and $\mathcal{L}_\omega$
Components: Three representation learning networks $\{I, C, A\}$, two regression networks h^0 and h^1 for potential outcomes, two network g_I, g_A to enforce I, A to predict Treatment and Factual outcome as precisely as possible.
for $i = 0, 1, 2, ...$ **do**
  $\left\{x_i, t_i, y_i^F\right\}_{i-1}^N \rightarrow \{I(X), C(X), A(X)\}$
  $\{I(X)\} \rightarrow g_I(I(X)) \rightarrow \hat{t}$
  $\{A(X)\} \rightarrow g_A(A(X)) \rightarrow \hat{y}$
  $\{C(X), A(X)\} \rightarrow h^1(C(X), A(X)), h^0(C(X), A(X)) \rightarrow \hat{y}^0, \hat{y}^1$
  update $\mathcal{W} \leftarrow Adam\{\mathcal{L}_{-\omega}\}$
  update $\omega \leftarrow Adam\{\mathcal{L}_\omega\}$
end for

where $\mathcal{W}$ is the the trainable parameter of $\{I, C, A, h^0, h^1, g_I, g_A\}$, ω is the trainable parameter of sample weight, and the maximum number of iterations is $\mathcal{I} = 3000$.

F Hyper-parameter Optimization

We adopt Adam optimizer with learning rate of 1e-3 to minimize DeR-CFR's objective function and select ELU as the non-linear activation function. Because we assign each unit with an adaptive weights, so we regard all samples as full-batch and take all training samples as input at each iteration. We set the maximum number of iterations as 3000. Table 5 shows the details on our hyper-parameter search space of DeR-CFR. We optimize our hyper-parameters in DeR-CFR by minimizing objective loss.

Table 5: Hyper-parameters Space of DeR-CFR

Hyper-parameters	**Values**
the number of the constrained layers l	{2, all}
batch norm	{False, True}
rep normalization	{False, True}
depth of layers of [d_R, d_y, d_t]	{1, 2, 3, 5, 7}
hidden state dimension of [h_R, h_y, h_t]	{32, 64, 128, 256}
{α, β, γ, μ, λ}	{1e-3, 1e-2, 1, 5, 10, 20, 50}

[6] showed that random search is more efficient for optimizing hyper-parameter than trials on grid search. In this paper, we randomly choose trails to optimize Hyper-parameters for each Dataset based on Hyper-parameters space (Tabel 5). In addition, we will prioritize to fix model capacity [d_R, d_y, d_t, h_R, h_y, h_t] and select norm operations based on search results with $\alpha = \beta = \gamma = \mu = \lambda = 0, k =$all. And then, we proceed to the other Hyper-parameters search to optimize our model. Tabel 6 shows the search results.

Table 6: Optimal Hyper-parameters

Hyper-parameters	**IHDP**	**Jobs**	**Twins**	**Syn**
l	2	2	all	all
batch norm	False	True	True	True
rep normalization	True	True	True	False
[d_R, d_y, d_t]	[7, 4, 1]	[5, 4, 1]	[7, 7, 3]	[2, 2, 3]
[h_R, h_y, h_t]	[32, 256, 256]	[32, 128, 128]	[64, 64, 64]	[256, 256, 256]
{α, β, γ, μ, λ}	[5, 50, 1, 10, 1e-2]	[1e-2, 1, 1e-2, 5, 1e-3]	[1e-2, 1e-4, 1e-4, 5, 5]	[1e-3, 1e-3, 1, 1, 1e-3]

G Detailed Description and Analysis of Real-world Datasets

G.1 Three real-world datasets

G.1.1 Semi-synthetic Benchmark: IHDP

The original Randomized Controlled Trial (RCT) data of the Infant Health and Development Program (IHDP) aims to evaluate the effect of a specialist home visits on the future cognitive test scores of premature infants. Hill [12] removed a non-random subset of the treated group and induced selection bias. The dataset comprises 747 units (139 treated, 608 control) with 25 pre-treatment variables related to the children and their mothers. We report the estimation errors on the same benchmark (100 realizations of the outcomes with 63/27/10 proportion of train/validation/test splits) provided by and used in [15, 22, 11].

G.1.2 Semi-synthetic Benchmark: Twins

The twins dataset is derived from the all twins birth in the USA between the year of 1989 and 1991 [1]. When a unit is the heavier one of the twins, the treatment is $t_i = 1$, and treatment is $t_i = 0$ if the unit is the lighter one. Besides that, we obtained 28 variables related to the parents, the pregnancy and the birth. The outcome is the children's mortality after one year. We focus on the same sex twins weighing less than 2000g and without missing features. The final dataset contains 5271 units. In order to create the selection bias and instrument variables, we first generate a 10-dimension covariate (iid.) for each sample using a binomial distribution:

$$X_1, X_2, \cdots, X_{10} \sim \mathcal{B}(5, 0.5)$$

So, for each sample there are 38 variables:

$$X = \{X_1, X_2, ..., X_{38}\}$$

where $\{X_{11}, X_{12}, ..., X_{38}\}$ comes from the variables of the original data. Because the data in the twins dataset is discrete, in order to make the variables in each dimension the same status, we normalize the data in each dimension by dividing by their respective maximum values. The treatment assignment strategy is:

$$t_i | x_i \sim \text{Bern}\left(Sigmoid\left(w^T X_{AB} + n\right)\right) \tag{22}$$

where, $w^T \sim \mathcal{U}\left((-0.1, 0.1)^{44\times 1}\right)$ and $n \sim \mathcal{N}(0, 0.1)$.

We conduct our experiments on the 10 realizations with 63/27/10 proportion of train/validation/test splits.

G.1.3 Real-world Data: Jobs

The jobs dataset created by LaLonde [17] is a widely used benchmark in the causal inference community, which is based on the randomized controlled trial samples. The study contains 8 variables such as age, education, and so on. The treatment is whether a unit participated in a job training program, and the outcome is the unit's employment status. Following Smith and Todd [23], we use the LaLonde's data (297 treated, 425 control) and the PSID comparison group (2490 control) to carry out our experiment. We randomly split the data of a total of 3212 samples into train/validation/test with 63/27/10 ratio (10 realizations).

G.2 Results analysis

On IHDP/Twins dataset, we adopt the Precision in Estimation of Heterogeneous Effect ($\mathcal{E}_{\text{PEHE}}$) [12] as the performance metric:

$$\mathcal{E}_{\text{PEHE}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{\text{e}}_i - \text{e}_i\right)^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left((\hat{y}_i^1 - \hat{y}_i^0) - (y_i^1 - y_i^0)\right)^2} \tag{23}$$

On Jobs dataset, there is not ground truth for counterfactual outcomes, so the policy risk [22] is adopted as the metric, which is defined as:

$$\mathcal{R}_{pol} = 1 - \left(\mathbb{E}\left[y^1|\pi_f(x)=1, t=1\right]\mathcal{P}\left(\pi_f(x)=1\right) + \mathbb{E}\left[y^0|\pi_f(x)=0, t=0\right]\mathcal{P}\left(\pi_f(x)=0\right)\right) \tag{24}$$

where, $\pi_f(x) = 1$ if $\hat{y}_1 - \hat{y}_0 > 0$ and $\pi_f(x) = 0$, otherwise. The policy risk measures the expected loss if the treatment is taken according to the ITE estimation. For PEHE and policy risk, the smaller value is, the better the performance.

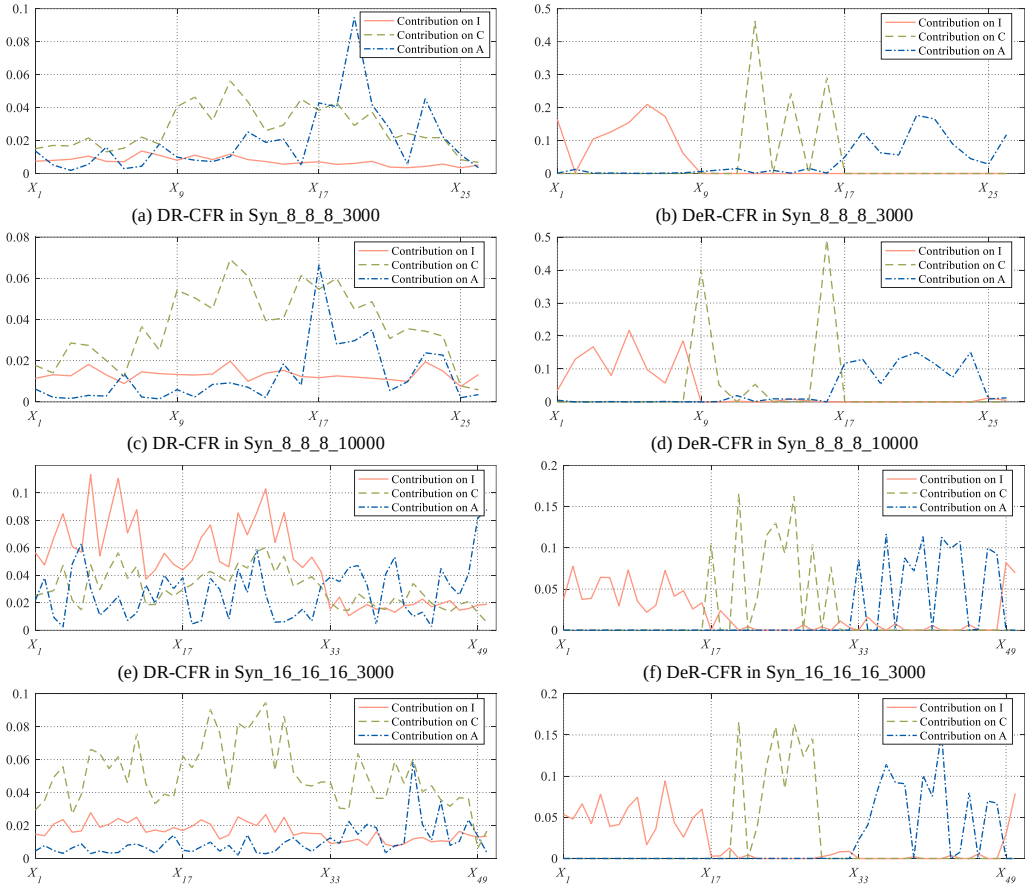
Table 7: The results of Real-world Data.

Within-sample						
Datasets	**IHDP(Mean $\pm$ Std)**		**Jobs(Mean $\pm$ Std)**		**Twins-28(Mean $\pm$ Std)**	
Methods	PEHE	ϵ_{ATE}	$\mathcal{R}_{pol}(\pi)$	ϵ_{ATT}	PEHE	ϵ_{ATE}
CFR-MMD	0.702 ± 0.037	0.284 ± 0.036	0.253 ± 0.004	0.062 ± 0.023	0.279 ± 0.001	0.010 ± 0.004
CFR-WASS	0.702 ± 0.034	0.306 ± 0.040	0.241 ± 0.005	$\mathbf{0.036 \pm 0.013}$	0.277 ± 0.001	0.021 ± 0.001
CFR-ISW	0.598 ± 0.028	0.210 ± 0.028	0.241 ± 0.003	0.037 ± 0.013	0.279 ± 0.001	0.036 ± 0.002
SITE	0.609 ± 0.061	0.259 ± 0.091	0.241 ± 0.007	0.052 ± 0.019	0.279 ± 0.001	0.037 ± 0.003
DR-CFR	0.657 ± 0.028	0.240 ± 0.032	$\mathbf{0.235 \pm 0.004}$	0.047 ± 0.015	0.276 ± 0.001	$\mathbf{0.006 \pm 0.002}$
DeR-CFR	$\mathbf{0.444 \pm 0.020}$	$\mathbf{0.130 \pm 0.020}$	0.246 ± 0.003	0.053 ± 0.084	$\mathbf{0.276 \pm 0.001}$	0.008 ± 0.003
Out-of-sample						
Datasets	**IHDP(Mean $\pm$ Std)**		**Jobs(Mean $\pm$ Std)**		**Twins-28(Mean $\pm$ Std)**	
Methods	PEHE	ϵ_{ATE}	$\mathcal{R}_{pol}(\pi)$	ϵ_{ATT}	PEHE	ϵ_{ATE}
CFR-MMD	0.795 ± 0.078	0.309 ± 0.039	$\mathbf{0.222 \pm 0.019}$	0.084 ± 0.028	0.284 ± 0.005	0.010 ± 0.004
CFR-WASS	0.798 ± 0.088	0.325 ± 0.045	0.237 ± 0.022	0.101 ± 0.036	0.281 ± 0.005	0.023 ± 0.003
CFR-ISW	0.715 ± 0.102	0.218 ± 0.031	0.225 ± 0.024	0.089 ± 0.033	0.283 ± 0.006	0.039 ± 0.004
SITE	1.335 ± 0.698	0.341 ± 0.116	0.229 ± 0.023	$\mathbf{0.074 \pm 0.028}$	0.283 ± 0.006	0.040 ± 0.004
DR-CFR	0.789 ± 0.091	0.261 ± 0.036	0.235 ± 0.015	0.088 ± 0.030	0.280 ± 0.005	$\mathbf{0.009 \pm 0.003}$
DeR-CFR	$\mathbf{0.529 \pm 0.068}$	$\mathbf{0.147 \pm 0.022}$	0.225 ± 0.022	0.093 ± 0.032	$\mathbf{0.280 \pm 0.005}$	0.010 ± 0.004

Table 7 shows that in comparison with state-of-the-art methods, DeR-CFR has achieved a significant improvement on PEHE and ϵ_{ATE} measures on the IHDP dataset. On Twins, the performance of DeR-CFR has hardly improved. We believe that this is due to the insufficient variable information related to the results of Twins. Each variable in the Twins data set is discrete, such as state of care, race, place of residence, etc. and most of the data are similar, because parents tend to adopt positive behaviors to protect their babies. Therefore, the lack of information leads to the low improvement of DeR-CFR performance (It may be close to the upper bound of prediction performance). On Jobs, due to the missing counterfactual outcomes, we have to adopt policy risk [22] to evaluate the quality of ITE estimation. But we find an abnormal phenomenon (Table 7) that some results in out-of-sample case is better than in within-sample case. In addition, the results reported in [22, 27] show that the performance of the traditional methods and state-of-the-Art methods in policy risk is almost the same. Thus, we hardly conclude that one model is better than the other. And if the eight variables selected manually are all confounding variables, DeR-CFR and DR-CFR are equivalent to CFR. At least, our results are almost the same as those of CFR.

H The Identification Performance of DeR-CFR and DR-CFR

Figure 4 reports the identification results for instrumental, confounding, and adjustment factors in a series of synthetic datasets. It's obvious in Figure 4 that our DeR-CFR algorithm can precisely identify the three underlying factors, while the baseline DR-CFR fails to disentangle those factors. This result vilifies the motivation of the proposed DeR-CFR, and is coninsident with our analyses on the comparison of DeR-CFR and DR-CFR algorithms in our paper.



(a) DR-CFR in Syn_8_8_8_3000 (b) DeR-CFR in Syn_8_8_8_3000

(c) DR-CFR in Syn_8_8_8_10000 (d) DeR-CFR in Syn_8_8_8_10000

(e) DR-CFR in Syn_16_16_16_3000 (f) DeR-CFR in Syn_16_16_16_3000

(g) DR-CFR in Syn_16_16_16_10000 (h) DeR-CFR in Syn_16_16_16_10000

Figure 4: Visualization of the contribution of each variable in X on the decomposed representation of I, C and A under different synthetic settings. (a) DR-CFR in Syn_8_8_8_3000, (b) DeR-CFR in Syn_8_8_8_3000, (c) DR-CFR in Syn_16_16_16_3000, (d) DeR-CFR in Syn_16_16_16_3000, (e) DR-CFR in Syn_8_8_8_10000, (f) DeR-CFR in Syn_8_8_8_10000, (g) DR-CFR in Syn_16_16_16_10000, (h) DeR-CFR in Syn_16_16_16_10000. Under the setting Syn_m_I_m_C_m_A_n, where $X_I = \{X_1 \cdots, X_{m_I}\}$, $X_C = \{X_{m_I+1} \cdots, X_{m_I+m_C}\}$ and $X_A = \{X_{m_I+m_C+1} \cdots, X_{m_I+m_C+m_A}\}$ are the true underlying factors of I, C and A.