Learning Decomposed Representation for Counterfactual Inference

Anpeng Wu, Kun Kuang, Junkun Yuan, Qiang Zhu, Yueting Zhuang, Fei Wu
Zhejiang University
Hangzhou, China
(anpwu;kunkuang;yuanjk;zhuq;yzhuang;wufei)@cs.zju.edu.cn

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

Runze Wu
NetEase Inc.
Hangzhou, China
wurunze1@corp.netease.com

ABSTRACT

The fundamental problem in treatment effect estimation from observational data is confounder identification and balancing. Most of the previous methods realized confounder balancing by treating all observed pre-treatment variables as confounders, ignoring further identifying confounders and non-confounders. In general, not all the observed pre-treatment variables are confounders that refer to the common causes of the treatment and the outcome, some variables only contribute to the treatment and some only contribute to the outcome. Balancing those non-confounders, including instrumental variables and adjustment variables, would generate additional bias for treatment effect estimation. By modeling the different causal relations among observed pre-treatment variables, treatment and outcome, we propose a synergistic learning framework to 1) identify confounders by learning decomposed representations of both confounders and non-confounders, 2) balance confounder with sample re-weighting technique, and simultaneously 3) estimate the treatment effect in observational studies via counterfactual inference. Empirical results on synthetic and real-world datasets demonstrate that the proposed method can precisely decompose confounders and achieve a more precise estimation of treatment effect than baselines.

KEYWORDS

Treatment Effect, Decomposed Representation, Confounder Identification and Balancing, Counterfactual Inference

1 INTRODUCTION

Causal inference is a powerful statistic modeling tool for explanatory analysis and plays an essential role in the decision-making process [27]. One fundamental problem in causal inference is treatment effect estimation. For example, in the medical scenario, accurately assessing a particular drug's treatment effect on each patient will help doctors decide which medical procedure (e.g., taking the drug or not) will benefit a specific patient most. The gold standard approach for treatment effect estimation is to perform Randomized Controlled Trials (RCTs), where different treatments (i.e., medical procedures) are randomly assigned to units (i.e., patients). However, fully RCTs are often expensive [17], unethical or even infeasible [7]. Hence, it is incredibly imperative and highly demanding to develop automatic statistical approaches to infer treatment effect in observational studies.

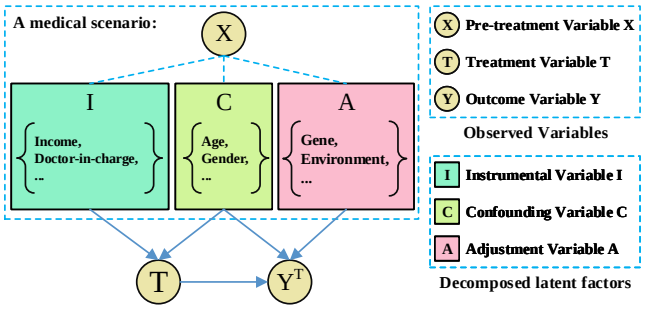


Figure 1: The intuitive illustration of our proposed causal framework w.r.t a medical scenario. Here, the historical data includes patients' pre-treatment variables X, the treatment T and the final outcome Y. Among these historical data, age and gender would simultaneously affect the treatment (doctors' decision) and the outcome (patients' physical differences), hence belonging to the set of confounding factor C; while the income and doctor-in-charge would only affect the treatment, hence belonging to the set of instrumental factor I; gene and environment belong to the set of adjustment factor A, since they would only affect the outcome. Our proposed algoritm intends to decomepose the pre-treatment variables X into the three kinds of latent factor $\{I, C, A\}$ for confounder identification and balancing.

In observational studies, we denote the causal framework among the observed pre-treatment variables X, the treatment T and the outcome Y, shown in Figure 1. Without loss of generality, we assume that the pre-treatment 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 the treatment and the outcome, and adjustment factor A only determines the outcome. Taking the medical scenario as an example, we might collect lots of historical data from each patient, including the treatment T (taking a particular drug or not), the outcome Y (state of health) and patient's features X (e.g., age, gender, income, gene).

*Under review, ** ****, **, **, ***
© 2021 Association for Computing Machinery.
ACM ISBN xxx-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/xx.xxxx/nnnnnnn.nnnnnnn

Among the patient's features, age and gender would simultaneously affect the treatment (doctor would consider the patient's age and gender when choosing the treatment) and the outcome (patient's age and gender would also affect his/her recovery rate), hence belonging to the set of confounding factor C; while the income and doctor-in-charge would only affect the treatment, but have no effect on the outcome, hence belonging to the set of instrumental factor I; gene and environment belong to the set of adjustment factor A, since they would only affect the outcome but have no effect on the treatment.

Different from RCTs, the treatment T in the observational studies is not randomly assigned; instead depends on some or all attributes of unit X (i.e., the factors I and C in Figure 1). This change could result in confounding bias: $Pr(T|X) \neq Pr(T)$. To eliminate the bias, previous methods, such as propensity score-based methods [4, 5, 29] and variables balancing methods [3, 37], simply treated all observed pre-treatment variables as confounding factor (confounders) for balancing. However, back-door criteria [25, 26] demonstrated that controlling the confounding factor is sufficient for removing that bias. In contrast, the instrumental factor's control invariably causes an increase in confounding bias if it exists. Moreover, [19] demonstrated that separating confounding factor and adjustment factor would reduce the estimated treatment effect variance. Overall, balancing the variables that mixed with non-confounders (i.e., instrumental and adjustment factors in Figure 1) would increase the bias and variance of treatment effect estimation [18, 24, 33]. Hence, it is indispensable to decompose the three factors for reducing the bias and variance of treatment effect estimation.

Recently, [19] proposed a data-driven variable decomposition method to separate adjustment variables from all observed pre-treatment variables and achieved lower variance on treatment effect estimation. Nevertheless, it ignored the decomposition of instrumental factor, which led to entanglement between instrumental and confounding factors. Moreover, it only focused on the settings with linear assumptions. [12] proposed to roughly separate the pre-treatment variables into three sets $\{I, C, A\}$ with a disentangled representations learning framework (like Figure 1). However, it could not guarantee the separation between the instrumental and the confounding factors (discussed in detail in the following section), leading to the entanglement among those three factors. Hence, precisely decomposing the instrumental, confounding and adjustment factors for confounder balancing and treatment effect estimation is still an open problem in observational studies.

To address this problem, we propose the following preliminary propositions for decomposing latent factors $\{I, C, A\}$ from pre-treatment variables X as shown in Figure 1: (i) **Decomposing** A **from** X: (i.a) the adjustment factor A should be independent of the treatment variable T, i.e., $A \perp T$; 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 factor I and outcome variable Y conditional on the treatment variable T, i.e., $I \perp Y \mid T$; 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 representations of confounding factor C and adjustment factor A help to predict both factual $y_i^{t_i}$ and counterfactual outcome $y_i^{1-t_i}$.

Guided by these preliminary propositions, we further propose a synergistic learning algorithm, named Decomposed Representations for CounterFactual Regression (DeR-CFR), to jointly 1) learn and decompose the representations of the three latent factors for feature decomposition, 2) optimize sample weights for confounder balancing, and 3) learn a counterfactual regression model to predict the counterfactual outcome for treatment effect estimation in observational studies. 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 study the problems of confounder identification and balancing for counterfactual prediction, which is critical for accurate treatment effect estimation in observational studies.
- We propose a novel DeR-CFR algorithm to jointly decompose instrumental, confounding, and adjustment factors accurately, and learn counterfactual regression to estimate treatment effect in observational studies.
- We empirically demonstrate that our algorithm can precisely decompose the latent factors, and the results show our approach achieves a better performance of treatment effect 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 [21, 22, 28, 29]; or directly optimize sample weight, including entropy balancing, residual balancing and stable balancing [3, 37]. Those existing methods focus on confounder balancing alone, while ignoring the importance of confounder identification. Recently,[18, 33] pointed out the necessity of confounder identification and selection for causal inference, due to the fact that the control of some non-confounders (e.g., variables related to the instrumental factor) would generate additional bias and amplify the variance. Besides, many methods [25, 34] have been proposed for confounder selection, but most assume the causal structure is known prior.

[15, 30] proposed a representations learning method for confounder balancing by minimizing the distribution difference between different treatment arms in embedding space. Based on these works, [11] proposed to optimize a context-aware importance sampling weight with representations learning jointly. Rather than taking the state-of-the-art ITE estimators to balance distribution globally, [35] proposed a local similarity preserving approach for representations learning. In this paper, we propose a decomposed representations learning approach for confounder identification along with a model-free weight schema for confounder balancing.

Our work is related to [19] and [12]. [19] 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 that they ignored the differentiation between instrumental and confounder factors, leading to imprecise confounder identification and failing to provide an estimation of ITE. Aiming at disentangling the three latent factors $\{I, C, A\}$ from the pre-treatment variables X, [12] proposed disentangled representations for counterfactual regression. However, the algorithm cannot guarantee to clearly decompose I, C and A. Extremely, $I(X)^\star = \emptyset, C(X)^\star = \{I, C, A\}, A(X)^\star = \emptyset$ could be a possible solution of their algorithm. They cannot guarantee accurate learning disentangled representations of the instrumental factor and confounding factor, which may introduce additional bias. Moreover, [12] relied on the correct model specification on treatment and the importance sampling weights 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 NOTATIONS AND PROPOSITIONS

In this section, we first give the notations and assumptions for treatment effect estimation in observational data, then propose a series of propositions to decompose instrumental, confounding and adjustment factors with representation leaning for treatment effect estimation.

3.1 Notations and Assumptions

In this paper, 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 first focus on the case of the binary treatment (for the continuous treatments, we will discuss in Section 5), and estimating the Individual Treatment Effect (ITE) of each unit i:

$$ITE_i = y_i^1 - y_i^0 \tag{1}$$

With ITE of each unit, one can easily estimate the Average Treatment Effect (ATE) as:

$$ATE = \mathbb{E}[y^1 - y^0] = \frac{1}{n}\sum_{i=1}^n ITE_i \tag{2}$$

From the definition of ITE and ATE, 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. To address this problem, we propose a counterfactual inference framework for predicting the counterfactual outcome.

Our analysis in this paper relies on the following standard assumptions [14] for treatment effect estimations.

Assumption 1: Stable Unit Treatment Value. The distribution of the potential outcome of one unit is assumed to be independent of the treatment assignment of another unit.

Assumption 2: Unconfoundedness. The distribution of treatment is independent of the potential outcome when given the pre-treatment variables. Formally, $T \perp \left(Y^0, Y^1\right) | X$.

Assumption 3: Overlap. Every unit should have a nonzero probability to receive either treatment status. Formally, $0 < p(T = 1|X) < 1$.

3.2 Preliminary Propositions

As shown in Figure 1, we assume that any dataset of the form $\{X, T, Y\}$ is generated from three latent factors $\{I, C, A\}$. Inspired by the causal framework, we further generate the following preliminary propositions to support decomposition and representations learning of these three latent factors.

Proposition 1: The adjustment factor would be independent of the treatment variable. Formally, $A \perp T$.

Proposition 2: Under the unconfounderness assumption, controlling confounding factor can help to break the relationship between the confounding factor and the treatment variable. That is, $C \perp T$.

Proposition 3: By controlling the confounding factor, the instrumental factor would become independent of 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. 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 [25]. 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.

Decomposing A: Proposition 1 can only constrain that the information of other factors (i.e., I and C) would not be embedded into A, but A might be embedded into other factors, resulting in information leaking of A. To address this problem, we propose to simultaneously maximize the predictive power of A on outcome Y to precisely decompose the adjustment factor A.

Decomposing I: Similarly, proposition 3 can only constrain that other factors (i.e., C and A) 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 the precise decomposition of instrumental factor I.

By decomposing I and A from X, we can identify confounder C. Then, with the decomposed C and A, we can accurately estimate the treatment effect via potential outcomes regression.

4 DER-CFR ALGORITHM

Guided by the above preliminary propositions and analyses, we propose a novel model, named Decomposed Representations for CounterFactual Regression (DeR-CFR), to learn the decomposed representations of instrumental, confounding, and adjustment factors for confounder identification and balancing, and simultaneously

learn a counterfactual regression model for treatment effect estimation. The overall 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 $h^1(C(X), A(X))$.

Our model's core components are the decomposition and balancing regularizers, which help the representation networks learn the decomposed representations 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.

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

4.1 Decomposing A

From the preliminary proposition, we know the adjustment factor should be independent of the treatment variable, $A(X) \perp T$. Considering the treatment is binary, we propose to learn the decomposed representation of adjustment factor $A(X)$ 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 as:

$$\mathcal{L}_A = disc(\{A(x_i)\}_{i:t_i=0}, \{A(x_i)\}_{i:t_i=1}) + \textstyle\sum_i l[y_i, g_A(A(x_i))] \quad (3)$$

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. $\{A(x_i)\}_{i:t_i=k}$ denotes the distribution of adjustment factor representation $A(X)$ with respect to the treatment arm $t = k$. Function $disc(\cdot)$ denotes the discrepancy of adjustment factor distribution between different treatment arms. Many integral probability metrics (IPMs) [23, 32], such as Maximum Mean Discrepancy (MMD) [9] 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 the instrumental factor I and the confounding factor C would not be embedded into $A(X)$, since I and C are associated with the treatment variable. Moreover, vice versa with maximizing the predictive power of $A(X)$ on Y, we can ensure all the information of adjustment factor would embed to $A(X)$, hence would not be embedded into other factors. Hence, the regularizer can help to decompose the adjustment factor.

4.2 Decomposing I and Balancing C

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

Firstly, we introduce the loss function of confounder balancing in our model. Most previous work [4, 11, 29] 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 the confounding factor C to the treatment variable T, that is, to make $C(X)$ become independent of T. Assuming that we have the decomposed representation of confounding factor $C(X)$, we propose to achieve confounder balancing[1] by directly learning sample weight ω with minimizing the following objective function:

$$\mathcal{L}_{C\_B} = disc\left(\{\omega_i \cdot C(x_i)\}_{i:t_i=0}, \{\omega_i \cdot C(x_i)\}_{i:t_i=1}\right) \quad (4)$$

where $\{\omega_i \cdot C(x_i)\}_{i:t_i=0}$ 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 sample reweighting with the learned ω.

Based on the property of the sample weight ω (i.e., $C(X) \perp T|\omega$), we can decompose the instrumental factor by conditional independence $I(X) \perp Y \mid T, \omega$. 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:

$$\begin{aligned}\mathcal{L}_I = &\textstyle\sum_{k=\{0,1\}} disc\left(\{\omega_i \cdot I(x_i)\}_{i:y_i=0}, \{\omega_i \cdot I(x_i)\}_{i:y_i=1}\right)_{i:t_i=k} \\ &+ \textstyle\sum_i l[t_i, g_I(I(x_i))]\end{aligned} \quad (5)$$

where $disc(\{\omega_i \cdot I(x_i)\}_{i:y_i=0}, \{\omega_i \cdot I(x_i)\}_{i:y_i=1})_{i:t_i=k}$ constrains the learned representation of instrumental factor I to be independent of 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 outcome, we will discuss in Section 5.

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. Moreover, vice versa with maximizing the predictive power of $I(X)$ on T, we can ensure all instrumental factor information would be embedded into $I(X)$, hence would not be embedded into other factors. Hence, this regularizer can help to decompose the instrumental factor accurately.

4.3 Deep Orthogonal Regularizer

Although the representation learning based on the proposed propositions mainly contributes to the decomposition of the feature information of instrumental variables I, confounding variables C and

[1] Recently, [16, 36] proposed alternatives for IPM (e.g., counterfactual variance) as a measure of imbalance, arguing that distributional distances are unnecessarily substantial. Therefore, there is still room for further improvement on confounder balancing.

adjustment variables A, data-driven neural networks tend to overfit the training data and lead to unclean disentanglement (like DR-CFR). Inspired by the orthogonal regularizer in [19] for variable decomposition, in this paper, we employ a deep orthogonal regularizer among the three representation networks for decomposing the factors $\{I, C, A\}$. We take 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 obtain $\bar{W}_I \in \mathbb{R}^m$, denoting the average contribution of each variable in X on the representation $I(X)$. Similarly, we 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 the vectors that have the same dimensions. Then, we propose to achieve hard decomposition by constraining orthogonality on each pair of them. The loss is as follow:

$$\mathcal{L}_O = \bar{W}_I^T \cdot \bar{W}_C + \bar{W}_C^T \cdot \bar{W}_A + \bar{W}_A^T \cdot \bar{W}_I \tag{6}$$

To guarantee the information flows of the representation networks, we softly constrain the total contribution of each $\bar{W}_I$, $\bar{W}_C$ and $\bar{W}_A$ to approximately 1, that can be found in tht regularization term *Reg* (Section 4.5). The orthogonal regularizer ensures each variable's information in X is either discarded or can only flow into one representation network for a hard decomposition. It can also reduce the influence of irrelevant variables on the prediction and prevent each representation network from overfitting.

4.4 Outcome Regression

With the decomposed representations, we propose to learn the outcome regression model for estimating the treatment effect. Similar to [12, 15, 30], 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 train these regression models only based on the decomposed representations of $C(X)$ and $A(X)$.

$$\mathcal{L}_R = \sum_i \omega_i \cdot l\left[y_i, h^{t_i}\left(C\left(x_i\right), A\left(x_i\right)\right)\right] \tag{7}$$

where the sample weight ω is learned from confounder balancing with Eq. 4.

4.5 Objective Function

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

$$\mathcal{L} = \mathcal{L}_R + \alpha \cdot \mathcal{L}_A + \beta \cdot \mathcal{L}_I + \gamma \cdot \mathcal{L}_{C\_B} + \mu \cdot \mathcal{L}_O + \lambda \cdot Reg \tag{8}$$

where *Reg* refers to the regularization term on the DeR-CFR parameters:

$$Reg = \mathcal{R}_{\mathcal{W}} + \mathcal{R}_{C\_B} + \mathcal{R}_O \tag{9}$$

where $\mathcal{R}_{\mathcal{W}}$ is the l_2 regularization on the parameters of subnetworks $\{I, C, A, h^0, h^1, g_I, g_A\}$. $\mathcal{R}_{C\_B}$ restricts the sample weight ω not to be all *zero*. To guarantee the information flows of the representation networks, we use $\mathcal{R}_O$ to softly constrain the sum of each $\bar{W}_I$, $\bar{W}_C$, and $\bar{W}_A$ to approximately 1. The details of each regularization are introduced in the appendix for saving space.

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

$$\mathcal{L}_{-\omega} = \mathcal{L}_R + \alpha \cdot \mathcal{L}_A + \beta \cdot \mathcal{L}_I + \mu \cdot \mathcal{L}_O + \lambda \cdot Reg \tag{10}$$

$$\mathcal{L}_{\omega} = \mathcal{L}_R + \gamma \cdot \mathcal{L}_{C\_B} + \lambda \cdot Reg \tag{11}$$

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 appendix.

5 DISCUSSION ON CONTINUOUS SCENES

For continuous or multi-valued treatment and outcome, we can approximately achieve the three propositions by making treatment and outcome binary during the process of decomposing or utilizing the mutual information between $\{I, C, A\}$ with T and Y (CLUB [8]).

5.1 Binarize the treatment and outcome

For the case of the continuous treatments T and continuous outcomes Y: In the prediction phase, we can use the regression network to predict the continuous treatment and outcome, but in the representation decomposition stage, we need to binarize the continuous treatment and outcome separately to approximately achieve the group division and some independence constraints. Without loss of generality, we use the median to divide the dataset based on the treatments T and the outcomes Y:

$$t_i^* := \begin{cases} 0, & t_i < median(\{t_i\}) \\ 1, & t_i \geq median(\{t_i\}) \end{cases} \tag{12}$$

$$y_i^* := \begin{cases} 0, & y_i < median(\{y_i\}) \\ 1, & y_i \geq median(\{y_i\}) \end{cases} \tag{13}$$

where $median(\{t_i\})$ refers to the median of factual outcome $\{t_i\}$ and $median(\{y_i\})$ refers to the median of factual outcome $\{y_i\}$. Similarly, other conditional division methods are also applicable.

While decomposing A and I and balancing C, we can use t_i^* and y_i^* to achieve the group division and minimize the discrepancy of I, C and A in different groups by the function $disc(\cdot)$. When the treatment is continuous, we will take $C(X)$, $A(X)$ and T to regress Y to replace Eq. 7 as follows:

$$\mathcal{L}_R = \textstyle\sum_i \omega_i \cdot l\left[y_i, h\left(C\left(x_i\right), A\left(x_i\right), t_i\right)\right] \tag{14}$$

Then we can use the objective function and training strategy (Section 4.5) to optimize the representation for confounder identification and conducting counterfactual inference. In this context, we binarize the continuous treatment and outcome to extend DeR-CFR on continuous scenes.

5.2 Utilize the mutual information

On the case of continuous treatment and outcome, the proposition 1 can be implemented by minimizing the mutual information between $A(X)$ and T (CLUB [8]):

$$\mathcal{L}_A = \textstyle\sum_i l[y_i, g_A(A(x_i))] + MI\left(A\left(x\right), t\right) \tag{15}$$

where $MI(a, b)$ refers to the mutual information of distribution a and b.

Besides, we can approximately achieve the conditional independence $I(X) \perp Y \mid T, \omega$ by minimizing the mutual information between $I(X)$ and Y (CLUB [8]):

$$\mathcal{L}_I = \sum_i l[t_i, g_I(I(x_i))] + \sum_{k=\{0,1\}} MI\left(I\left(x_i\right), y_i\right)_{i:t_i^*=k} \tag{16}$$

Based on mutual information, DeR-CFR can be applied to continuous outcome scenarios but not applicable to continuous treatment (it can not balance C or decompose I directly).

6 EXPERIMENTS

6.1 Baselines

We compare the proposed algorithm (**DeR-CFR**) with the following baselines. (1) **CFR-MMD** and **CFR-WASS** [15, 30]: CounterFactual Regression with MMD and Wasserstein metrics; (2) **CFR-ISW** [11]: CounterFactual Regression with Importance Sampling Weights; (3) **SITE** [35]: local Similarity preserved Individual Treatment Effect estimator; and (4) **DR-CFR** [12]: Disentangled Representations for CounterFactual Regression. For continuous scenes, we binarize the continuous treatment to run these baselines and utilize the learned representation to regress the continuous outcomes.

6.2 Experiments on Real Dataset

6.2.1 Dataset. In order to evaluate the proposed method, we conduct the experiment on three real-world datasets that are adopted in [35]: IHDP, Jobs and Twins-28. IHDP aims to evaluate the effect of specialist home visits on premature infants' future cognitive test scores and Jobs aims to estimate the effect of job training programs on employment status.

The twins dataset is derived from the all twins born in the USA between the year of 1989 and 1991 [1]. When a unit is the heavier one in the twins, the treatment is $t_i = 1$, and the lighter one is $t_i = 0$. Besides, we obtained 28 variables related to parents, pregnancy, and birth. The outcome is the children's mortality after one year. We focus on same-sex twins weighing less than 2000g and without missing features. The final dataset contains 5271 records. To develop the instrument variables, we generate 38-dimension variables for each unit: $X = \{X_1, X_2, ..., X_{38}\}$, where $X_1, X_2, \cdots, X_{10} \sim \mathcal{B}(5, 0.5)$ and $\{X_{11}, X_{12}, ..., X_{38}\}$ comes from the original data. The treatment assignment strategy is: $t_i|x_i \sim \text{Bern}\left(\text{sigmoid}\left(w^T X_{AB} + n\right)\right)$, where $w^T \sim U\left((-0.1, 0.1)^{44\times 1}\right)$ and $n \sim N(0, 0.1)$. We conduct our experiments on the 10 realizations of Twins with a 63/27/10 proportion of train/validation/test splits.

6.2.2 Metrics. On IHDP and Twins, we adopt the Precision in Estimation of Heterogeneous Effect (PEHE) [12, 13] 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 ATE prediction $\epsilon_{\text{ATE}} = |ATE - \widehat{ATE}|$ to evaluate performance, where ATE $= \mathbb{E}(y^1) - \mathbb{E}(y^0)$.

On Jobs dataset, there is no ground truth for counterfactual outcomes, so the policy risk [30] is adopted, which is defined as: $\mathcal{R}_{pol} = 1 - \mathbb{E}\left[y^1|\pi_f(x) = 1, t = 1\right]\mathcal{P}\left(\pi_f(x) = 1\right) - \mathbb{E}\left[y^0|\pi_f(x) = 0, t = 0\right]\mathcal{P}\left(\pi_f(x) = 0\right)$, 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.

6.2.3 Results. We report the results, including the mean and standard deviation (std) of treatment effect over 100 replications on IHDP, 10 replications on Jobs and Twins-28 datasets in Table 1. The results show that in comparison with state-of-the-art methods, DeR-CFR outperforms all baselines and achieves a significant improvement on PEHE and ϵ_{ATE} measures on the IHDP dataset. On Jobs and Twins, DeR-CFR has comparable performance to the state-of-art in estimating treatment effects. Our algorithm does not achieve such significant improvement on Jobs and Twins-28 than IHDP data; the main reason we analyzed is that (i) on Jobs, most of the manually selected variables may be confounding variables, DeR-CFR would be not prominent compared with other baseline in this case; (ii) on Twins, all variables are discrete and most units have similar data, which leads to the low improvement of our DeR-CFR algorithm.

Table 2 investigates the effects of each module of the DeR-CFR by conducting ablation experiments on IHDP. From Tabel 1 and Table 2, we can draw the following conclusions: (i) With explicitly learning the decomposed representations, 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 confuse the decomposed representations learning and damage the performance of ITE estimation on IHDP dataset.

6.3 Experiments on Synthetic Dataset

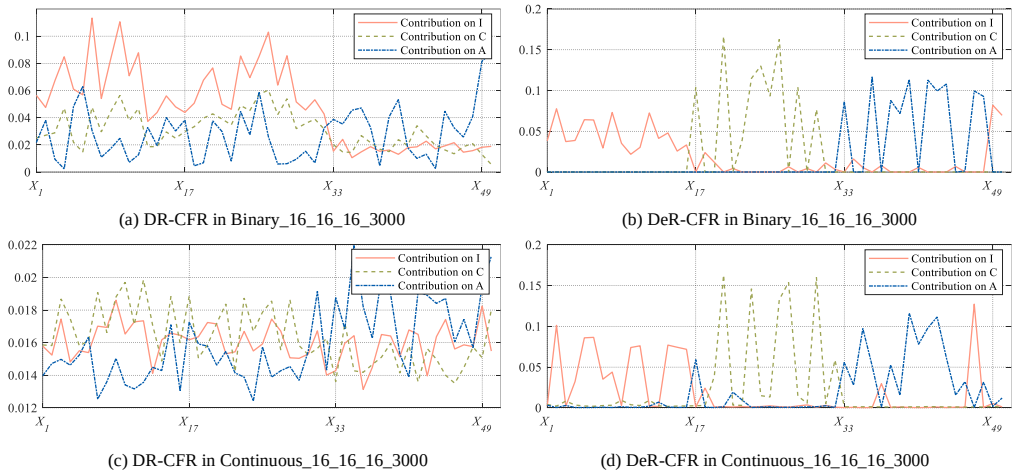
6.3.1 Dataset. To generate synthetic datasets, we design two different sample sizes $n = \{3000, 10000\}$ and two dimensional settings $\{m_I, m_C, m_A\}$=$\{8, 8, 8\}$ or $\{16, 16, 16\}$, where m_I, m_C, and m_A denote the dimensions of instrumental variables, confounding variables and adjustment variables, respectively. Thus, the total dimension of pre-treatment 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 $X_1, X_2, \cdots, X_m \sim \mathcal{N}(0, 1)$.

Binary Setting: In this paper, we focus on the setting with binary treatment and binary outcome. We first generate binary treatment $t = binomial(1, 1/(1 + e^{-z}))$, where $z = \frac{1}{10}\theta_t \times X_{IC} + \varepsilon$, X_{IC} denotes the variables in X that belongs to I and C. Then, generate binary outcomes corresponding to different treatment arms as $y^0 = \text{sign}\left(\max\left(0, z^0 - \bar{z}^0\right)\right)$ 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 Binary_$m_I$_$m_C$_$m_A$_$n$ 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}.

Continuous Setting: Our algorithm can be also applied for continuous treatment and outcome as we discussed in Section 5. Here, we also generate continuous treatment and outcome as $t = 1/(1 + e^{-z})$ and $y = z^0 + t * z^1 + \epsilon_y$, where $\epsilon_y \sim \mathcal{N}(0, 0.1)$. We use Continuous_$m_I$_$m_C$_$m_A$_$n$ to denote continuous settings.

Table 1: The results (mean±std) of treatment effect estimation on real-world data.

Within-sample						
Datasets	**IHDP(Mean ± Std)**		**Jobs(Mean ± Std)**		**Twins-28(Mean ± Std)**	
Methods	PEHE	ϵ_{ATE}	$\mathcal{R}_{pol}(\pi)$	ϵ_{ATT}	PEHE	ϵ_{ATE}
CFR-MMD	0.702 ± 0.037	0.284 ± 0.036	0.194 ± 0.004	**0.041 ± 0.015**	0.279 ± 0.001	0.010 ± 0.004
CFR-WASS	0.702 ± 0.034	0.306 ± 0.040	0.194 ± 0.004	0.041 ± 0.016	0.277 ± 0.001	0.021 ± 0.001
CFR-ISW	0.598 ± 0.028	0.210 ± 0.028	0.189 ± 0.006	0.041 ± 0.017	0.279 ± 0.001	0.036 ± 0.002
SITE	0.609 ± 0.061	0.259 ± 0.091	0.224 ± 0.005	0.064 ± 0.022	0.279 ± 0.001	0.037 ± 0.003
DR-CFR	0.657 ± 0.028	0.240 ± 0.032	0.199 ± 0.006	0.064 ± 0.026	0.276 ± 0.001	**0.006 ± 0.002**
DeR-CFR	**0.444 ± 0.020**	**0.130 ± 0.020**	**0.187 ± 0.037**	0.053 ± 0.084	**0.276 ± 0.001**	0.008 ± 0.003
Out-of-sample						
Datasets	**IHDP(Mean ± Std)**		**Jobs(Mean ± Std)**		**Twins-28(Mean ± Std)**	
Methods	PEHE	ϵ_{ATE}	$\mathcal{R}_{pol}(\pi)$	ϵ_{ATT}	PEHE	ϵ_{ATE}
CFR-MMD	0.795 ± 0.078	0.309 ± 0.039	0.222 ± 0.019	0.084 ± 0.028	0.284 ± 0.005	0.010 ± 0.004
CFR-WASS	0.798 ± 0.088	0.325 ± 0.045	0.225 ± 0.023	0.102 ± 0.047	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	**0.074 ± 0.028**	0.283 ± 0.006	0.040 ± 0.004
DR-CFR	0.789 ± 0.091	0.261 ± 0.036	0.235 ± 0.015	0.119 ± 0.045	0.280 ± 0.005	0.009 ± 0.003
DeR-CFR	**0.529 ± 0.068**	**0.147 ± 0.022**	**0.208 ± 0.062**	0.093 ± 0.032	**0.279 ± 0.005**	**0.008 ± 0.004**



(a) DR-CFR in Binary_16_16_16_3000

(b) DeR-CFR in Binary_16_16_16_3000

(c) DR-CFR in Continuous_16_16_16_3000

(d) DeR-CFR in Continuous_16_16_16_3000

Figure 2: Visualization of the contribution of each variable in X on the decomposed representations of I, C and A under the settings with Binary_16_16_16_3000 (sub-figures a,b) and Continuous_16_16_16_3000 (sub-figures c,d), 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.

Table 2: Results (mean±std) of ablation studies on IHDP dataset (✓ refers to keeping the component in DeR-CFR).

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

6.3.2 Results of treatment effect estimation. In binary setting, we compare our DeR-CFR with the contending baselines under different 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 the explicit decomposition of instrumental, confounding and adjustment factors during representations learning, the performance of DeR-CFR is much better than DR-CFR.

In continuous setting with Continuous_16_16_16_3000, we report the mean square error (MSE) of counterfactual outcome prediction (detailed definition is introduce in the appendix) with 10 independent replications in Table 4. From the result, we can conclude that considering the decomposed representation of confounders and non-confounders, our DeR-CFR can achieve the best performance than baselines on counterfactual regression.

6.3.3 Results on Decomposed Representation. To evaluate the performance of decomposed 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 the previous section. Figure 2 reports the results under settings of binary_16_16_16_3000 (Figure 2(a,b)) and continuous_16_16_16_3000 (Figure 2(c,d)). It is evident in Figure 2 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

Table 3: Results (mean ± std) on synthetic data under different settings (Binary_m_I_m_C_m_A_n).

Within-sample								
Setting	**Binary_8_8_8_3000**		**Binary_8_8_8_10000**		**Binary_16_16_16_3000**		**Binary_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	**0.325 ± 0.002**	**0.014 ± 0.006**	**0.234 ± 0.003**	**0.007 ± 0.002**	**0.404 ± 0.003**	**0.011 ± 0.004**	**0.307 ± 0.002**	**0.006 ± 0.002**
Out-of-sample								
Setting	**Binary_8_8_8_3000**		**Binary_8_8_8_10000**		**Binary_16_16_16_3000**		**Binary_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**	**0.376 ± 0.006**	**0.018 ± 0.005**

Table 4: Results (mean ± std) on continuous setting.

Methods	**MSE**	
	Within-sample	Out-of-sample
CFR-MMD	0.044 ± 0.008	0.048 ± 0.022
CFR-WASS	0.046 ± 0.006	0.054 ± 0.012
CFR-ISW	0.058 ± 0.014	0.064 ± 0.009
SITE	0.235 ± 0.033	0.216 ± 0.059
DR-CFR	0.098 ± 0.044	0.106 ± 0.032
DeR-CFR	**0.026 ± 0.002**	**0.028 ± 0.004**

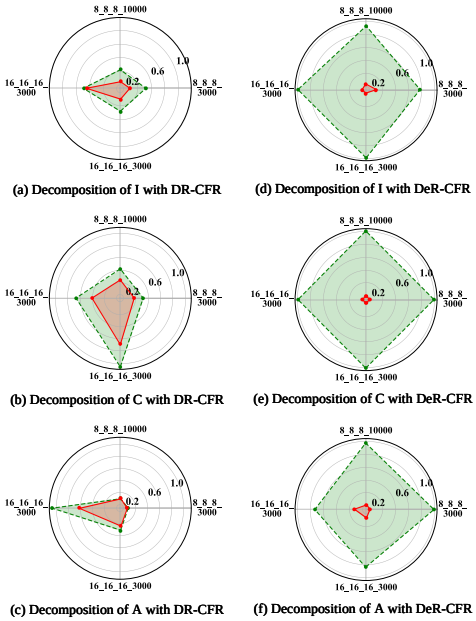


Figure 3: Radar charts that visualize the disentangled/decomposed representations of all three underlying factors from DR-CFR (sub-figures a,b,c) and DeR-CFR (sub-figures d,e,f) methods. Each vertex on the polygons denotes an experimental setting with form Binary_m_I_m_C_m_A_n. The green and red plots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

the motivation of the proposed DeR-CFR and is consistent with our analysis on the comparison of DeR-CFR and DR-CFR algorithms in the previous section.

Similar to the setting in DR-CFR [12], we also plot the radar charts on the representation of each factor for comparison in Figure 3. For example, in Figure 3(a), we calculate the average contribution of true variables of I in X, i.e., $X_I = \{X_1, \cdots, X_{16}\}$ on the representation of I (plotted with 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 decomposed representation, our DeR-CFR achieves much better decomposed/disentangled representations of all three underlying factors than DR-CFR. This is the key reason that our DeR-CFR can obtain significant improvement on treatment effect estimation than DR-CFR, as shown in Table 3.

6.4 Hyper-parameters Analysis

Given the complex multi-term objective function (Section 4.5) in DeR-CFR, we study the impact of each item on the accuracy of the potential outcomes under setting Binary_16_16_16_3000 by changing $\{\alpha, \beta, \gamma, \mu, \lambda\}$ in the scope $\{0, 0.01, 0.1, 1.0, 10, 100\}$. The result in Figure 4 demonstrates that the performance of DeR-CFR is mostly affected by changing in α and λ, reflecting the fact that decomposing adjustment factor A accurately will greatly contribute to the improvement of performance and limiting the complexity of the model is necessary. μ will guarantee the decomposition of three latent factors, which not only help each representation network to select information, but will also prevent the model from overfitting. β and γ may not affect the accuracy obviously, but they are an essential condition for confounder identification. With hyper-parameters analysis, we can choose the best hyper-parameters for experiments.

6.5 Mutual Information Interpretation.

We also demonstrate the mutual information with lower and upper bound (CLUB [8]) under setting Binary_16_16_16_3000. The results are summarized in Table 5, which demonstrates the learned I from DeR-CFR is weakly correlated with Y but highly correlated with T, and the learned A from DeR-CFR is weakly related to T but highly

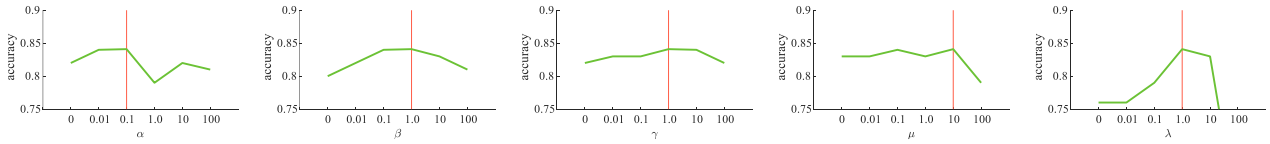


Figure 4: Hyper-parameter sensitivity analysis on $\{\alpha, \beta, \gamma, \mu, \lambda\}$. The green lines show the accuracy of the these parameters within the specified range $\{0, 0.01, 0.1, 1.0, 10, 100\}$. The red line indicates the best parameters for the setting.

Table 5: Mutual Information interpretation for DeR-CFR.

MI	DR-CFR		DeR-CFR	
	T	Y	T	Y
I	0.0267 ~ 0.0472	0.0158 ~ 0.0150	0.1993 ~ 0.3874	0.0010 ~ 0.0823
C	0.0157 ~ 0.2115	0.0141 ~ 0.2004	0.3729 ~ 0.4561	0.3599 ~ 0.4439
A	0.0001 ~ 0.0004	0.0001 ~ 0.0004	0.0439 ~ 0.2113	0.2494 ~ 0.4151
X	0.4892 ~ 0.6485	0.3365 ~ 0.6605	0.4892 ~ 0.6485	0.3365 ~ 0.6605

correlated with Y. Consistent with the results in Figure 2, the mutual information between factors $\{I, C, A\}$ with treatment T and Y shows DeR-CFR does decompose instrumental variables I, confounding variables C and adjustment variables A. In addition, the results show that the representation network I in DR-CFR overfits the training data and the learned A from DR-CFR may be empty (i.e., $A = \emptyset$) without explicit decomposition constraints.

7 CONCLUSION

In this paper, we focus on the problem of estimating 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 factor and confounding factor. Hence, we propose a Decomposed representations learning algorithm for CounterFactual Regression (DeR-CFR) with explicit decomposition constraints for confounder identification and balancing, and simultaneously estimate the treatment effect via counterfactual inference. Empirical results demonstrate the advantages of the DeR-CFR algorithm compared with state-of-the-art methods.

REFERENCES

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

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein gan. *arXiv preprint arXiv:1701.07875* (2017).

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

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

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

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

[7] Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. 2013. Counterfactual reasoning and learning systems: The example of computational advertising. *The Journal of Machine Learning Research* 14, 1 (2013), 3207–3260.

[8] Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. 2020. CLUB: A Contrastive Log-ratio Upper Bound of Mutual Information. *arXiv preprint arXiv:2006.12013* (2020).

[9] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *Journal of Machine Learning Research* 13, Mar (2012), 723–773.

[10] Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. 2017. Deep IV: A flexible approach for counterfactual prediction. In *International Conference on Machine Learning*. PMLR, 1414–1423.

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

[12] Negar Hassanpour and Russell Greiner. 2020. Learning Disentangled Representations for CounterFactual Regression. In *International Conference on Learning Representations*.

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

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

[15] Fredrik Johansson, Uri Shalit, and David Sontag. 2016. Learning representations for counterfactual inference. In *International conference on machine learning*. 3020–3029.

[16] Fredrik D Johansson, David Sontag, and Rajesh Ranganath. 2019. Support and invertibility in domain-invariant representations. *arXiv preprint arXiv:1903.03448* (2019).

[17] Ron Kohavi and Roger Longbotham. 2011. Unexpected results in online controlled experiments. *ACM SIGKDD Explorations Newsletter* 12, 2 (2011), 31–35.

[18] Kun Kuang, Peng Cui, Bo Li, Meng Jiang, Yashen Wang, Fei Wu, and Shiqiang Yang. 2019. Treatment Effect Estimation via Differentiated Confounder Balancing and Regression. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 14, 1 (2019), 1–25.

[19] Kun Kuang, Peng Cui, Bo Li, Meng Jiang, Shiqiang Yang, and Fei Wang. 2017. Treatment effect estimation with data-driven variable decomposition. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[20] Robert J LaLonde. 1986. Evaluating the econometric evaluations of training programs with experimental data. *The American economic review* (1986), 604–620.

[21] Sheng Li, Nikos Vlassis, Jaya Kawale, and Yun Fu. 2016. Matching via Dimensionality Reduction for Estimation of Treatment Effects in Digital Marketing Campaigns.. In *Proceedings of the Twenty-fifth International Joint Conference on Artificial Intelligence, IJCAI-16*. 3768–3774.

[22] Yao Liuyi, Chu Zhixuan, Li Sheng, Li Yaliang, Gao Jing, and Zhang Aidong. 2020. A Survey on Causal Inference. *arXiv preprint arXiv:2002.02770* (2020).

[23] Alfred Müller. 1997. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability* (1997), 429–443.

[24] Jessica A Myers, Jeremy A Rassen, Joshua J Gagne, Krista F Huybrechts, Sebastian Schneeweiss, Kenneth J Rothman, Marshall M Joffe, and Robert J Glynn. 2011. Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American journal of epidemiology* 174, 11 (2011), 1213–1222.

[25] Judea Pearl. 2009. *Causality*. Cambridge university press.

[26] Judea Pearl. 2012. On a class of bias-amplifying variables that endanger effect estimates. *arXiv preprint arXiv:1203.3503* (2012).

[27] Judea Pearl et al. 2009. Causal inference in statistics: An overview. *Statistics surveys* 3 (2009), 96–146.

[28] Paul R Rosenbaum. 1987. Model-based direct adjustment. *J. Amer. Statist. Assoc.* 82, 398 (1987), 387–394.

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

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

[31] Jeffrey A Smith and Petra E Todd. 2005. Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of econometrics* 125, 1-2 (2005), 305–353.

[32] Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. 2009. On integral probability metrics,\phi-divergences and binary classification. *arXiv preprint arXiv:0901.2698* (2009).
[33] Tyler J VanderWeele. 2019. Principles of confounder selection. *European journal of epidemiology* 34, 3 (2019), 211–219.
[34] Tyler J VanderWeele and Ilya Shpitser. 2011. A new criterion for confounder selection. *Biometrics* 67, 4 (2011), 1406–1413.
[35] Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. 2018. Representation learning for treatment effect estimation from observational data. In *Advances in Neural Information Processing Systems*. 2633–2643.
[36] Yao Zhang, Alexis Bellot, and Mihaela van der Schaar. 2020. Learning overlapping representations for the estimation of individualized treatment effects. *arXiv preprint arXiv:2001.04754* (2020).
[37] José R Zubizarreta. 2015. Stable weights that balance covariates for estimation with incomplete outcome data. *J. Amer. Statist. Assoc.* 110, 511 (2015), 910–922.

A THE REGULARIZATION TERM ON DER-CFR PARAMETERS

In the DeR-CFR Algorithm, *Reg* refers to the regularization term on network parameters:

$$Reg = \mathcal{R}_{\mathcal{W}} + \mathcal{R}_{C\_B} + \mathcal{R}_O \tag{17}$$

Next, we describe each component of *Reg* in detail.

A.1 The regularization on the network parameters.

In the DeR-CFR Algorithm, we add l_2 regularization on the parameters of subnetworks $\{I, C, A, h^0, h^1, g_I, g_A\}$ to prevent over-fitting:

$$\mathcal{R}_{\mathcal{W}} = l_2\left(\mathcal{W}(I, C, A, h^0, h^1, g_I, g_A)\right) \tag{18}$$

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 model's parameter value is small enough. To prevent overfitting, we penalize the immense value in the network parameters $\mathcal{W}(I, C, A, h^0, h^1, g_I, g_A)$ by l_2 regularization.

A.2 The regularization on the sample weight.

$\mathcal{R}_{C\_B}$ restricts the sample weight ω not to be all *zero* and approximately 1:

$$\mathcal{R}_{C\_B} = \left(\sum_{i:t_i=0} \omega_i - 1\right)^2 + \left(\sum_{i:t_i=1} \omega_i - 1\right)^2 \tag{19}$$

To avoid all the sample weights to be *zero* and maintain original quantity allocation on each treatment arm, we constrain the sample weight $\sum_{i:t_i=0} \omega_i = \sum_{i:t_i=1} \omega_i = 1$.

A.3 The regularization on the orthogonal regularizer.

While minimizing $\mathcal{L}_O$ (in Eq. 6), the deep orthogonal regularizer may lead to the result $\bar{W}_I^k = \bar{W}_C^k = \bar{W}_A^k = 0$ for all dimension k. To guarantee the information flows of the representation networks, we softly constrain the sum of each $\bar{W}_I$, $\bar{W}_C$, and $\bar{W}_A$ to approximately 1:

$$\mathcal{R}_O = \left(\sum_{k=1}^{m} \bar{W}_I^k - 1\right)^2 + \left(\sum_{k=1}^{m} \bar{W}_C^k - 1\right)^2 + \left(\sum_{k=1}^{m} \bar{W}_A^k - 1\right)^2 \tag{20}$$

Table 6: Hyper-parameters Space of DeR-CFR

Hyper-parameters	Values
the number of the constrained layers l	{2, all}
batch norm 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}
$\{\alpha, \beta, \gamma, \mu, \lambda\}$	{1e-3, 1e-2, 1e-1 1, 5, 10, 100}

B MSE: THE PERFORMANCE METRIC IN CONTINUOUS SCENES

When the treatment and outcome are continuous, the goal of counterfactual outcome prediction is to get a counterfactual estimation function $h\left(C\left(x_i\right), A\left(x_i\right), t_i\right)$ that is close to true response function $f\left(x_i, t_i\right) = z^0\left(x_i\right) + t_i * z^1\left(x_i\right)$, typically measured by Mean Square Error (MSE) [10]:

$$MSE = \frac{1}{n}\sum_i \left(h\left(C\left(x_i\right), A\left(x_i\right), \frac{i}{n}\right) - \left(z^0\left(x_i\right) + \frac{i}{n} * z^1\left(x_i\right)\right)\right)^2 \tag{21}$$

where n denotes the number of units, and we use $^i/_n$ to replace t_i in the process of calculating MSE by simulating Monte Carlo sampling.

C PSEUDO-CODE OF DER-CFR

As mentioned in the DeR-CFR Algorithm, the overall architecture of the 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 $h^1(C(X), A(X))$.

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

$$\mathcal{L}_{-\omega} = \mathcal{L}_R + \alpha \cdot \mathcal{L}_A + \beta \cdot \mathcal{L}_I + \mu \cdot \mathcal{L}_O + \lambda \cdot Reg \tag{22}$$

$$\mathcal{L}_{\omega} = \mathcal{L}_R + \gamma \cdot \mathcal{L}_{C\_B} + \lambda \cdot Reg \tag{23}$$

We minimize $\mathcal{L}_{-\omega}$ using stochastic gradient descent to update the parameters of the representation and hypothesis network, and

Table 7: Optimal Hyper-parameters

Hyper-parameters	IHDP	Jobs	Twins	Binary	Continuous
l	2	2	all	all	all
batch norm	False	True	True	False	False
rep normalization	True	True	True	False	False
$\{d_R, d_y, d_t\}$	[7, 4, 1]	[5, 4, 1]	[7, 7, 3]	[2, 5, 5]	[5, 2, 3]
$\{h_R, h_y, h_t\}$	[32, 256, 256]	[32, 128, 128]	[64, 64, 64]	[256, 128, 128]	[256, 64, 64]
$\{\alpha, \beta, \gamma, \mu, \lambda\}$	[5, 100, 1, 10, 1e-2]	[1e-2, 1, 1e-2, 5, 1e-3]	[1e-2, 1e-3, 1e-3, 5, 5]	[1e-1, 1, 1, 10, 1]	[1e-1, 1, 1e-1, 1, 10]

minimize $\mathcal{L}_\omega$ to update ω. Algorithm 1 shows the details of the pseudo-code of DeR-CFR [2].

Algorithm 1 Decomposed Representations 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 the potential outcomes, two network g_I, g_A to enforce I, A to predict Treatment and Factual outcome as precisely as possible.
for $i = 0, 1, 2, \ldots$ **do**
  $\left\{x_i, t_i, y_i^F\right\}_{i=1}^N \rightarrow \{I(X), C(X), A(X)\}$
  $\{I(X)\} \rightarrow g_I(I(X)) \rightarrow \hat{t}$
  $\{A(X)\} \rightarrow g_A(A(X)) \rightarrow \hat{y}$
  $h^0(C(X), A(X)), h^1(C(X), A(X)) \rightarrow \hat{y}^0, \hat{y}^1$
  update $\mathcal{W} \leftarrow \text{Adam}\{\mathcal{L}_{-\omega}\}$
  update $\omega \leftarrow \text{Adam}\{\mathcal{L}_\omega\}$
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 sample weights, and the maximum number of iterations is $\mathcal{I} = 3000$.

Hardware used: Ubuntu 16.04.5 LTS operating system with 2 * Intel Xeon E5-2678 v3 CPU, 384GB of RAM, and 4 * GeForce GTX 1080Ti GPU with 44GB of VRAM.

Software used: Python with TensorFlow 1.15.0, NumPy 1.17.4, and MatplotLib 3.1.1.

D DETAILED DESCRIPTION OF REAL-WORLD DATA

D.1 Semi-synthetic Benchmark: IHDP

The original Randomized Controlled Trial (RCT) data of the Infant Health and Development Program (**IHDP** [3]) aims at evaluating the effect of specialist home visits on the future cognitive test scores of premature infants. Hill [13] 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 [12, 15, 30].

D.2 Real-world Data: Jobs

The **Jobs** [4] dataset created by LaLonde [20] is a widely used benchmark in the causal inference community, based on the randomized controlled trials. The dataset aims to estimate the effect of job training programs on employment status. Jobs contains 17 variables, such as age, education level, etc. Following Smith and Todd [31], we use 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 3212 samples into train/validation/test with a 56/24/20 ratio (10 realizations).

D.3 Real-world Data: Twins

The original **Twins** [5] dataset is derived from the all twins born in the USA between the year of 1989 and 1991 [1].

E HYPER-PARAMETER OPTIMIZATION

This algorithm selects ELU as the non-linear activation function and adopts Adam optimizer to minimize DeR-CFR's objective function with a learning rate of 1e-3. We assign an adaptive weight to each unit in the training process and regard all samples as one full-batch. The maximum number of iterations is 3000. Table 6 states the number and range of values tried per hyper-parameter during the paper's development. We return the best-evaluated iterate with early stopping and optimize the hyper-parameters in DeR-CFR by minimizing objective loss.

Bergstra et al. [6] demonstrated that trials on random search would be more efficient than grid search for optimizing hyper-parameter. In this paper, we randomly choose trails to determine the best Hyper-parameters for each Dataset within the Hyper-parameters space (Tabel 6). 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 $\alpha = \beta = \gamma = \mu = \lambda = 0, k =$all. And then, we proceed to the other Hyper-parameters search to optimize our model. Tabel 7 lists all optimal hyper-parameters of DeR-CFR used for each dataset in the paper's experiments.

[2]The code is available at the anonymous link: https://www.dropbox.com/sh/5m40z2vmthx0y10/AACXJFuOvgB24av1VqkrkmKRa?dl=0
[3]http://www.fredjo.com/
[4]http://www.fredjo.com/
[5]http://www.nber.org/data/linked-birth-infant-death-data-vital-statistics-data.html