Learning Decomposed Representations for Treatment Effect Estimation

Anpeng Wu*, Junkun Yuan*, Kun Kuang†, Bo Li, Runze Wu, Qiang Zhu, Yueting Zhuang, Fei Wu, *Senior Member, IEEE*

Abstract—In observational studies, confounder separation and balancing are the fundamental problems of treatment effect estimation. Most of the previous methods focused on addressing the problem of confounder balancing by treating all observed pre-treatment variables as confounders, ignoring confounder separation. 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 (i.e., instrumental variables) and some only contribute to the outcome (i.e., adjustment variables). 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 variables and outcome variables, we propose a synergistic learning framework to i) separate confounders by learning decomposed representations of both confounders and non-confounders, ii) balance confounder with sample re-weighting technique, and simultaneously iii) 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.

Index Terms—Treatment Effect, Decomposed Representation, Confounder Separation 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 [1], [2], [3], [4], [5], [6] and one of the important components of interpretable artificial intelligence [7], [8], [9], [10], [11]. 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 [12], unethical or even infeasible [13]. Hence, it is incredibly imperative and highly demanding to develop automatic statistical approaches to infer treatment effect in observational studies.

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 variables $\{I, C, A\}$ under an unknown joint distribution $Pr(X) = Pr(I, C, A)$, where I denotes the instrumental variables that only affect the treatment, C refers to the confounding variables (confounders) that are the common cause of the treatment and the outcome, and adjustment variables A only determine the outcome. Taking the medical scenario as an example, we might collect lots of historical data from patients, 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, etc.). 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 variables (confounders) 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 variables I; gene and environment belong to the set of adjustment variables 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, it depends on some or all attributes of unit X (i.e., the variables I and C in Figure 1). This change could result in confounding bias, i.e., $Pr(T|X) \neq Pr(T)$. To eliminate the bias, previous methods, such as propensity score-based methods [14], [15], [16] and variables balancing methods [17], [18], [19], simply treated all observed pre-treatment variables as confounding variables for balancing. However, back-door criteria [20], [21] demonstrated that controlling the confounding variables is sufficient for removing that bias. In contrast, controlling the instrumental variables invariably leads to increased confusion bias, if it exists. Moreover, [22], [23]

*Equal contribution.
†Corresponding author.

- A. Wu, J. Yuan, K. Kuang, Q. Zhu, Y. Zhuang and F. Wu are with the College of Computer Science and Technology, Zhejiang University, China. (E-mail: anpwu@zju.edu.cn; yuanjk@zju.edu.cn; kunkuang@zju.edu.cn; zhuq@zju.edu.cn; yzhuang@cs.zju.edu.cn; wufei@cs.zju.edu.cn).
- B. Li is with the School of Economics and Managemen, Tsinghua University, China. (E-mail: libo@sem.tsinghua.edu.cn).
- R. Wu is with the Fuxi AI Lab, NetEase Inc, Hangzhou, China. (E-mail: wurunze1@corp.netease.com).

*Manuscript received ***** **, ****; revised ***** **, ****.*

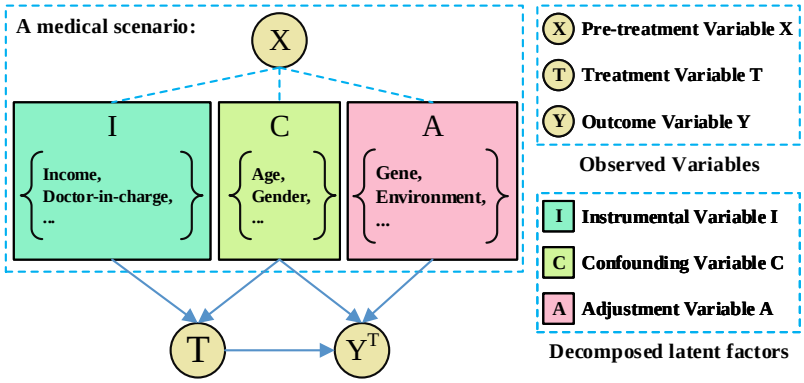


Fig. 1. The intuitive illustration of our proposed causal framework w.r.t a medical scenario. Here, the historical data includes patients' pre-treatment variables X, the treatment T and the final outcome Y. Among these historical data, age and gender would simultaneously affect the treatment (doctors'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

demonstrated that separating confounding variables and adjustment variables would reduce the estimated treatment effect variance. Overall, balancing the variables that mixed with non-confounders (i.e., instrumental and adjustment variables in Figure 1) would increase the bias and variance of treatment effect estimation [24], [25], [26]. Therefore, it is indispensable to decompose the three kinds of latent variables for reducing the bias and variance of treatment effect estimation.

Recently, [22], [23] 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 variables, which led to entanglement between instrumental and confounding variables. Moreover, it only focused on the settings with linear assumptions. [27] proposed to roughly separate the pre-treatment variables into three sets $\{I, C, A\}$ with a disentangled representation learning framework (similar to Figure 1). However, it could not guarantee the separation between the instrumental and the confounding variables (discussed in detail in the following section), leading to the entanglement among those three latent factors $\{I, C, A\}$. Hence, how to precisely decomposing the instrumental, confounding and adjustment variables for confounder balancing and treatment effect estimation is still an open problem in observational studies.

In this paper, we are interested in the case of a binary treatment (i.e., $T \in \{1, 0\}$). With considering the causal relationships among pre-treatment variables $X = \{I, C, A\}$, treatment T and outcome Y, we propose the following preliminary propositions for decomposing latent variables $\{I, C, A\}$ from pre-treatment variables X as shown in Figure 1: (i) **Decomposing A from X**: (i.a) the adjustment variables 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 variables's information (e.g., I and C) not be embedded into A, while (i.b) restrains A from embedding into other variables. (ii) **Decomposing I from X**: (ii.a) By learning sample weights ω [28], [29], we can well balance the confounding variables C, that is, one can break the dependency between C and T with sample weights ω (i.e., $C \perp T \mid \omega$). After that we can achieve the conditional independence between instrumental variables I and outcome variable Y given the treatment variable T. Formally, if $C \perp T \mid \omega$ then $I \perp Y \mid T$; and (ii.b) I should also predict T as accurately as possible. Condition (ii.a) constraints other variables not be embedded into I, while (ii.b) restrains I from embedding into other variables. (iii) **Balancing Confounders** C: re-weighting certain data instances to balance the representations of confounders could reduce confounding bias [27], [28], [29], [30], [31]. Instead of relying on the explicit propensity score, we directly optimize the global sample weight ω for each unit to balance confounder distributions between the treated and control populations, i.e., $C \perp T \mid \omega$. (iv) **Predicting factual and counterfactual outcomes** $\{Y^T, Y^{1-T}\}$: the decomposed representations of confounding variables C and adjustment variables A help to predict both factual Y^T and counterfactual outcome Y^{1-T}.

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 $\{I, C, A\}$ for feature decomposition, 2) optimize sample weights ω for confounder balancing, and 3) learn a counterfactual regression model to predict the counterfactual outcome Y^{1-T} (or the potential outcome $\{Y^0, Y^1\}$ on out-of-distribution data) for treatment effect estimation in observational studies. Our DeR-CFR algorithm is based on the standard assumptions [32] 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 propose a novel DeR-CFR algorithm to jointly decompose instrumental, confounding, and adjustment variables 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 almost

all decomposed features correspond to their real semantics, which any other method can not achieve.
- Extensive experiments show our approach achieves a better performance of treatment effect estimation in observational studies with both synthetic and real-world datasets, the error metric PEHE is reduced by 10% on average compared with the best baseline and the ATE bias is reduced by up to 30 %. Especially in IHDP dataset, the error metric PEHE is reduced by 26% and the ATE bias is reduced by 32.5%.

The rest of the paper is organized as follows. Section 2 reviews the related work. Section 3 gives the notations and propositions for confounder separation and balancing. The details of our proposed algorithm for decomposed representations for counterfactual regression are introduced in Section 4. Experimental results and analyses are reported in Section 5. Finally, Section 6 concludes the paper.

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 [16], [33], [34], [35], [36], [37]; or optimize sample weight by entropy balancing, residual balancing and stable balancing [17], [18], [38], [39]. Those existing methods focus on confounder balancing alone, while ignoring the importance of confounder separation. Recently, [25], [26] pointed out the necessity of confounder separation and selection for causal inference, due to the fact that the control of some non-confounders (e.g., variables related to the instrumental variables) would generate additional bias and amplify the variance. Besides, many methods [4], [20], [40] have been proposed for confounder selection, but most assume the causal structure is known prior.

[28], [29] proposed a representation learning method for confounder balancing by minimizing the distribution difference between different treatment arms in embedding space. Based on these works, [41] proposed to optimize a context-aware importance sampling weight with representation learning jointly. Rather than taking the ITE estimators to balance distribution globally, [30] proposed a local similarity preserving approach for representation learning. In this paper, we propose a decomposed representation learning approach for confounder separation along with a model-free weight schema for confounder balancing.

Our work is related to (N-)D^{2}VD [22], [23] and DR-CFR [27]. (N-)D^{2}VD [22], [23] proposed a data-driven variables decomposition algorithm to automatically separate confounder and adjustment variables for treatment effect estimation. The main limitation is that they ignored the differentiation between instrumental variables and confounders, leading to imprecise confounder separation and failing to provide a precise estimation of ITE. Aiming at disentangling the three latent factors $\{I, C, A\}$ from the pre-treatment variables X, DR-CFR [27] 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 confounders, which may introduce additional bias. Moreover, DR-CFR [27] relied on the correct model specification (propensity score) on treatment for confounder balancing with the importance sampling weights. Our proposed algorithm is different from these methods in two ways: (i) Confounder Separation: we propose a series of decomposition regularizers to guarantee the explicit fine-grained decomposition among the instrumental, confounder, and adjustment variables; (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 studies, then propose a series of propositions to decompose instrumental, confounding and adjustment variables 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., {0:not recovery, 1:recovery }) $y_i^{t_i} \in \mathcal{Y}$ as a result of choosing treatment t_i.

In our context, we focus on the case of the binary treatment, and 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 potential outcomes $\{y_i^0, y_i^1\}$to inference the counterfactual outcome $y_i^{1-t_i}$.

Our analysis in this paper relies on the following standard assumptions [32] for treatment effect estimation.

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 and the causal relationships among pre-treatment variables $X = \{I, C, A\}$, treatment T and outcome Y, we further generate the following preliminary propositions to support decomposition and representation learning of these three latent factors.

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

Proposition 2: Under the unconfounderness assumption, confounder balancing with a global sample weight ω [28], [29] can help to break the dependence between the confounding variables and the treatment variable. Formally, $C \perp T \mid \omega$.

Proposition 3: After confounder balancing with sample weight ω, the instrumental variables would become conditional independent of the outcome, given the treatment variable. That is, if $C \perp T \mid \omega$, we have $I \perp Y \mid T$.

Proposition 1 can be easily understood by the definition of adjustment variables. We can denote the path between adjustment variables and treatment variable as the collider structure at Y: $A \rightarrow Y \leftarrow T$, hence $A \perp T$. Under the unconfounderness assumption, C is a sufficient set to block all information from T to Y except $T \rightarrow Y$ after decomposing A and I. Obviously, proposition 2 can be guaranteed by the back-door criterion [20]. By balancing the confounders, the path between instrumental variables and outcome can be denoted as $I \rightarrow T \rightarrow Y$, hence $I \perp Y \mid T, \omega$ in proposition 3. Note that these three propositions are in no particular order, we will simultaneously optimize the above objectives.

Decomposing A: Proposition 1 can only constrain that the information of other variables (i.e., I and C) would not be embedded into A, but A might be embedded into other variables, 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 variables A.

Decomposing I: Similarly, proposition 3 only constrain that other variables (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 variables. In our context, we propose to jointly maximize the predictive power of I on treatment T for the precise decomposition of instrumental variables I.

Balancing C: By decomposing I and A from X, we separate and balance confounder C, i.e., $C \perp T \mid \omega$.

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 variables for confounder separation and balancing, and simultaneously learn a counterfactual regression model for treatment effect estimation. The overall

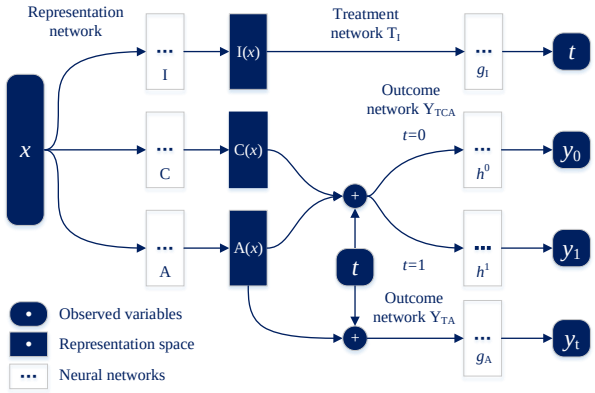


Fig. 2. Overview of DeR-CFR Architecture.

architecture (Figure 2) of our model consists of the following components:

- **Three decomposed representation networks** for learning latent factors, one for each underlying factor: $I(\cdot), C(\cdot), A(\cdot)$. We use $\hat{I} = I(X)$, $\hat{C} = C(X)$ and $\hat{A} = A(X)$ to denote the learned representations of instrumental variables I, confounders C, and adjustment variables A. **Two regression networks** maximize the predictive power of $I(X)$ on T and $A(X)$ on Y: $g_I(I(X))$ and $g_A(A(X))$.
- **Three decomposition and balancing regularizers** for confounder separation and balancing: the first is for decomposing adjustment variables A from X with considering $A(X) \perp T$ and $A(X)$ should predict Y as precisely as possible; the second is for decomposing instrumental variables I from X via constraining $I(X) \perp Y \mid T, \omega$, and $I(X)$ should be predictive to T; the last is designed for simultaneously balancing confounder $C(X)$ in different treatment arms, i.e., $C(X) \perp T \mid \omega$.
- **Two regression networks** for potential outcomes 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 from preliminary propositions, which help the representation networks learn the decomposed representations $\{I(X), C(X), A(X)\}$ for confounder separation, and also to improve the precision of regression networks via accurate confounder balancing with identified $C(X)$. The decomposition and balancing regularizers are the keys to bridge the representation networks and regression networks for treatment 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 variables 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 variables $A(X)$ by constraining the discrepancy of its distribution between treatment arms $T = 1$ and $T = 0$. Moreover, to

prevent the information of adjustment variables from being embedded into other variables, 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 variables as:

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

where $l[y_i, g_A(A(x_i))]$ would be an l_2-loss for continuous outcomes or a log-loss for binary outcomes. $\{A(x_i)\}_{i:t_i=k}$ denotes the distribution of adjustment variables representation $A(X)$ with respect to the treatment arm $t = k$. Function $disc(\cdot)$ denotes the discrepancy of adjustment variables distribution between different treatment arms. Many integral probability metrics (IPMs) [42], [43], such as Maximum Mean Discrepancy (MMD) [44] and Wasserstein distance [45], can be used to measure the discrepancy of distributions. In this paper, we use the MMD to calculate $disc(\cdot)$.

Data Flow: For representation network A and regression network g_A, we input the observed variables X and use the latent representation $A(X)$ to predict the outcome, i.e. $\mathbb{E}[Y|X] = g_A(A(X))$. By minimizing the above term $\mathcal{L}_A$, our model can ensure the information of the instrumental variables I and the confounding variables 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 variables would embed to $A(X)$, hence would not be embedded into other representations. Hence, the regularizer can help to decompose the adjustment variables, $A = A(X)$.

4.2 Decomposing I and Balancing C

From preliminary propositions, we know that if one can balance confounders with a global sample weights ω (i.e., $C(X) \perp T \mid \omega$), the instrumental variables would be conditional independent of the outcome variable given the treatment variable.

Firstly, we introduce the loss function of confounder balancing in our model. Most previous work [14], [16], [41] 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 causal link from the confounding variables C to the treatment variable T, that is, to make $C(X)$ become independent of T. Assuming that we have got the decomposed representation of confounding variables $C(X)$ after joint-training, we propose to achieve confounder balancing[1] by directly learning sample weight ω with minimizing the following objective function:

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

where $\{\omega_i \cdot C(x_i)\}_{i:t_i=0}$ refers to the weighted distribution of $C(X)$ on the samples with $t = 0$.

1. Recently, [46], [47] 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.

Data Flow: For representation network C, we input the observed variables X and get the latent representation $C(X)$. To avoid all the sample weights to be $zero$ or model only focuses on some samples and ignores others, we constrain the sample weight $\sum_{i:t_i=0} \omega_i = \sum_{j:t_j=1} \omega_j = 1, \omega_i > 0, \omega_j > 0$. Under the overlap assumption, if $\mathcal{L}_{C\_B}$ can be minimized to be $zero$, it means that the distribution of $C(X)$ between different treatment arms is consistent, on the case of the binary treatment $T = \{0, 1\}$. Then, we can achieve the independence between $C(X)$ and T by sample reweighting with the learned weight, $C(X) \perp T \mid \omega$.

Based on the property of the sample weight ω (i.e., $C(X) \perp T \mid \omega$), we can decompose the instrumental variables by conditional independence $I(X) \perp Y \mid T, \omega$. Moreover, to prevent the information of instrumental variables from being embedded into other variables, 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 variables is:

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

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 variables 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 or multi-valued outcome, we can approximately achieve the conditional independence $I(X) \perp Y \mid T, \omega$ by minimizing the mutual information between $I(X)$ and Y [49]:

$$\begin{aligned}\mathcal{L}_I &= \sum_i l[t_i, g_I(I(x_i))] + \sum_{k=\{0,1\}} MI(I(x_i), y_i)_{i:t_i=k} \\ &+ \sum_i l[t_i, g_I(I(x_i))]\end{aligned} \quad (6)$$

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

Data Flow: For representation network I and regression network g_I, we input the observed variables X and use the latent representation $I(X)$ to predict the treatment, i.e. $\mathbb{E}[T|X] = g_I(I(X))$. By minimizing the term $\mathcal{L}_I$, our model can ensure the information of confounding variables C and adjustment variables 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 variables information would be embedded into $I(X)$, hence would not be embedded into other representations. Hence, this regularizer helps to decompose the instrumental variables 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 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 [22], [23], [48] for variable decomposition, in this paper, we employ a

deep orthogonal regularizer among the three representation networks $\{I(X), C(X), A(X)\}$ for decomposing the variables $\{I, C, A\}$. We take the representation network for instrumental variables $I(X)$ 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:

$$\mathcal{L}_O = \bar{W}_I^T \cdot \bar{W}_C + \bar{W}_C^T \cdot \bar{W}_A + \bar{W}_A^T \cdot \bar{W}_I \tag{7}$$

Data Flow: For representation network $\{I, C, A\}$, we input the observed variables X and get the latent representation $\{I(X), C(X), A(X)\}$. 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 explicit 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 [27], [28], [29], we also train two regression networks for each treatment arm, $h^0(\cdot)$ and $h^1(\cdot)$, 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 $A(X)$.

$$\mathcal{L}_R = \sum_i \omega_i \cdot l\left[y_i, h^{t_i}\left(C\left(x_i\right), A\left(x_i\right)\right)\right] \tag{8}$$

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

Data Flow: For potential outcomes prediction network h^0 and h^1, we input the observed variables X to obtain the latent representation $C(X)$ and $A(X)$, and then use them to predict the potential outcomes, i.e. $\mathbb{E}[Y^0|X] = h^0(C(X), A(X))$ and $\mathbb{E}[Y^1|X] = h^1(C(X), A(X))$.

4.5 The Regularization Term on DeR-CFR Parameters

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

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

Next, we describe each component of Reg in detail.

4.5.1 The regularization on the network parameters.

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

$$\mathcal{R}_{\mathcal{W}} = l_2\left(\mathcal{W}(I, C, A, h^0, h^1, g_I, g_A)\right) \tag{10}$$

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.

4.5.2 The regularization on the sample weight.

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

$$\mathcal{R}_{C\_B} = \left(\sum_{i:t_i=0} \omega_i - 1\right)^2 + \left(\sum_{j:t_j=1} \omega_j - 1\right)^2, w_i > 0, w_j > 0 \tag{11}$$

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

4.5.3 The regularization on the orthogonal regularizer.

While minimizing $\mathcal{L}_O$ (in Eq. 7), 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:

$$\mathcal{R}_O = \left(\sum_{k=1}^{m} \bar{W}_I^k - 1\right)^2 + \left(\sum_{k=1}^{m} \bar{W}_C^k - 1\right)^2 + \left(\sum_{k=1}^{m} \bar{W}_A^k - 1\right)^2 \tag{12}$$

4.6 Objective Function

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

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

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

$$Reg = \mathcal{R}_{\mathcal{W}} + \mathcal{R}_{C\_B} + \mathcal{R}_O \tag{14}$$

where $\mathcal{R}_{\mathcal{W}}$ is the l_2 regularization on the parameters of subnetworks $\{I(\cdot), C(\cdot), A(\cdot), h^0(\cdot), h^1(\cdot), g_I(\cdot), g_A(\cdot)\}$. $\mathcal{R}_{C\_B}$ restricts the sample weight ω not to be $zero$. 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.

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

$$\mathcal{L}_{-\omega} = \mathcal{L}_R + \alpha \cdot \mathcal{L}_A + \beta \cdot \mathcal{L}_I + \mu \cdot \mathcal{L}_O + \lambda \cdot Reg \tag{15}$$

$$\mathcal{L}_{\omega} = \mathcal{L}_R + \gamma \cdot \mathcal{L}_{C\_B} + \lambda \cdot Reg \tag{16}$$

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

Algorithm 1 Decomposed Representations for CounterFactual Regression

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

TABLE 1
Hyper-parameters and Ranges

Hyper-parameter	Range
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}

Algorithm 1 shows the details of the pseudo-code of DeR-CFR [2], where $\mathcal{W}$ is the the trainable parameter of $\{\{I(\cdot), C(\cdot), A(\cdot), h^0(\cdot), h^1(\cdot), g_I(\cdot), g_A(\cdot)\}\}$, ω is the trainable sample weights, and the maximum number of iterations is $\mathcal{I} = 3000$.

4.7 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 1 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. [51] 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 1). 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 2 lists all optimal hyper-parameters of DeR-CFR used for each dataset in the paper's experiments.

2. The code is available at: https://www.dropbox.com/sh/5m40z2vmthx0y10/AACXJFuOvgB24av1VqkrkmKRa?dl=0

5 EXPERIMENTS

5.1 Baselines

We compare the proposed algorithm (**DeR-CFR**) with the following baselines. (1)**D^{2}VD** and **N-D^{2}VD** [22], [23]: (Nonlinear) Data-Driven Variable Decomposition; (2) **CFR-MMD** and **CFR-WASS** [28], [29]: CounterFactual Regression with MMD and Wasserstein metrics; (3) **CFR-ISW** [41]: CounterFactual Regression with Importance Sampling Weights; (4) **SITE** [30]: local Similarity preserved Individual Treatment Effect estimator; and (5) **DR-CFR** [27]: Disentangled Representations for CounterFactual Regression.

5.2 Experiments on Real Dataset

5.2.1 Dataset.

In order to evaluate the proposed method, we conduct the experiment on three real-world datasets that are adopted in [30]: 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.

IHDP[3]**:** The original Randomized Controlled Trial (RCT) data of the Infant Health and Development Program (IHDP) aims at evaluating the effect of specialist home visits on the future cognitive test scores of premature infants. Hill [52] 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 [27], [28], [29].

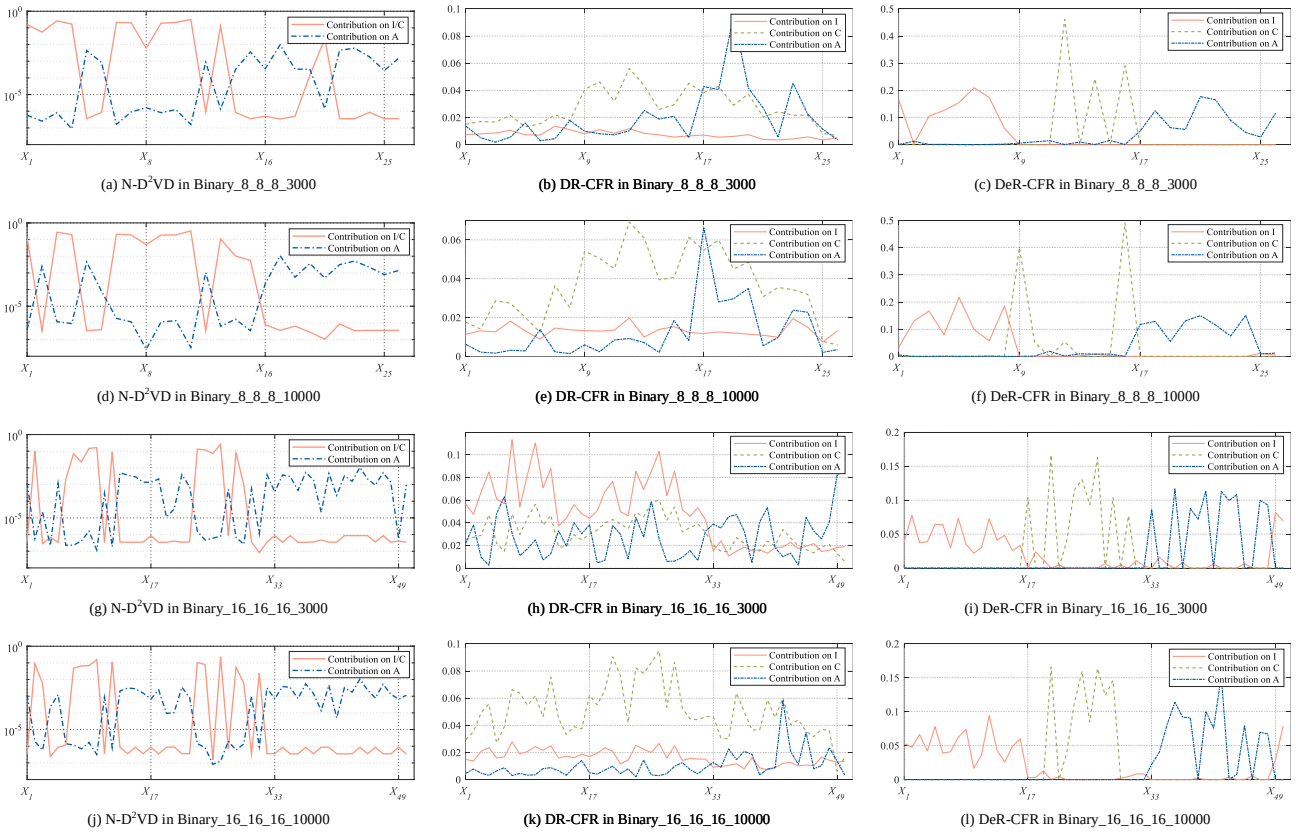
Jobs[4]**:** The Jobs dataset created by LaLonde [53] is a widely used benchmark in the causal inference community, based on the randomized controlled trials. The dataset aims to estimate the effect of job training programs on employment status. Jobs contains 17 variables, such as age, education level, etc. Following Smith and Todd [54], 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).

Twins[5]**:** The original Twins dataset is derived from the all twins born in the USA between the year of 1989 and 1991 [55]. 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

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

TABLE 2
Optimal Hyper-parameters

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



(a) N-D^{2}VD in Binary_8_8_8_3000 **(b) DR-CFR in Binary_8_8_8_3000** (c) DeR-CFR in Binary_8_8_8_3000

(d) N-D^{2}VD in Binary_8_8_8_10000 **(e) DR-CFR in Binary_8_8_8_10000** (f) DeR-CFR in Binary_8_8_8_10000

(g) N-D^{2}VD in Binary_16_16_16_3000 **(h) DR-CFR in Binary_16_16_16_3000** (i) DeR-CFR in Binary_16_16_16_3000

(j) N-D^{2}VD in Binary_16_16_16_10000 **(k) DR-CFR in Binary_16_16_16_10000** (l) DeR-CFR in Binary_16_16_16_10000

Fig. 3. Visualization of the contribution of each variable in X on the decomposed representations of I, C and A under the different settings with Binary_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.

$X_1, X_2, \cdots, X_{10} \sim \mathcal{B}(5, 0.5)$ and $\{X_{11}, X_{12}, ..., X_{38}\}$ comes from the original data. The treatment assignment strategy is: $t_i|x_i \sim \text{Bern}\left(\text{sigmoid}\left(w^T X_{IC} + n\right)\right)$, 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.

5.2.2 Metrics.

On IHDP and Twins, we adopt the Precision in Estimation of Heterogeneous Effect (PEHE) [27], [52] 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 [29] 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.

5.2.3 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 3. 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 in the IHDP dataset: the error metric PEHE is reduced by 26% and the ATE bias is reduced by 32.5% compared with the best baseline. On Jobs and Twins, DeR-CFR has comparable performance to the state-of-the-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

TABLE 3
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}
D^{2}VD	11.41 ± 2.513	0.269 ± 0.181	-	0.125 ± 0.018	0.728 ± 0.014	0.006 ± 0.004
N-D^{2}VD	4.246 ± 0.818	1.726 ± 0.226	-	0.115 ± 0.024	0.703 ± 0.017	**0.003 ± 0.002**
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}
D^{2}VD	14.67 ± 9.797	1.429 ± 1.247	-	0.224 ± 0.081	0.726 ± 0.053	0.028 ± 0.019
N-D^{2}VD	299.7 ± 700.2	39.95 ± 81.98	-	0.138 ± 0.041	0.719 ± 0.091	0.025 ± 0.017
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 ± 0.058	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**

* (N-)D^{2}VD: The factual outcomes of selected samples are all 1, and almost all of them are $\left[\pi_f(x)=0, t=0\right]$, i.e. $\mathcal{R}_{pol} \approx 1-\mathbb{E}\left[y^0|\pi_f(x)=0, t=0\right]=0$. This is not an ideal phenomenon: $\hat{y}_1 \leq \hat{y}_0$. We use '-' to denote it.

TABLE 4
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**

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 in our DeR-CFR algorithm.

Table 4 reports the effects of each module of the DeR-CFR by conducting ablation experiments on IHDP. From Tabel 3 and Table 4, 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 $\{I, C, A\}$. (ii) Each component in our DeR-CFR is necessary, since missing any one of them would confuse the decomposed representation learning and damage the performance of ITE estimation on IHDP dataset.

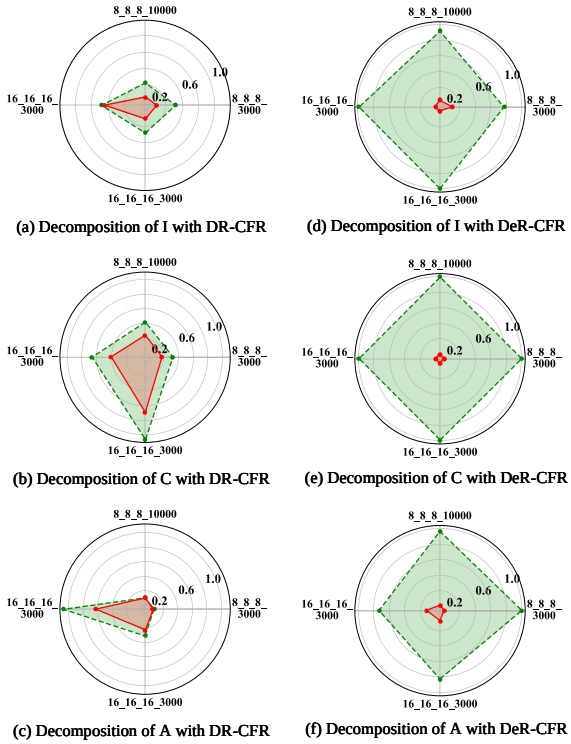


Fig. 4. Radar charts that visualize the disentangled/decomposed representations of all three underlying factors $\{I, C, A\}$ from DR-CFR (sub-figures a,b,c) and DeR-CFR (sub-figures d,e,f) methods. Each vertex on the polygons denotes an experimental setting with form Binary_m_I_m_C_m_A_n. The green and red plots denote the average contribution of true variables and other variables in X on the representation of each factor, respectively.

5.3 Experiments on Synthetic Dataset

5.3.1 Dataset.

To generate synthetic datasets, we design two different sample sizes $n=\{3000, 10000\}$ and two settings of variable dimensions $\{m_I, m_C, m_A\}$=$\{8, 8, 8\}$ 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

TABLE 5
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}
D^{2}VD	3.641 ± 2.027	0.037 ± 0.017	1.334 ± 0.082	0.034 ± 0.003	0.679 ± 0.015	0.074 ± 0.003	0.667 ± 0.006	0.072 ± 0.001
N-D^{2}VD	1.893 ± 0.463	0.201 ± 0.064	1.788 ± 0.645	0.008 ± 0.007	0.905 ± 0.037	0.073 ± 0.006	1.306 ± 0.073	0.056 ± 0.007
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}
D^{2}VD	3.654 ± 2.134	0.049 ± 0.051	1.173 ± 0.554	0.043 ± 0.024	0.723 ± 0.123	0.065 ± 0.034	0.686 ± 0.044	0.061 ± 0.021
N-D^{2}VD	1.725 ± 0.244	0.195 ± 0.146	1.347 ± 0.675	0.039 ± 0.046	1.454 ± 0.277	0.088 ± 0.057	1.289 ± 0.358	0.045 ± 0.026
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**

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}.

5.3.2 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 5. We see that DeR-CFR outperforms other state-of-the-art methods in PEHE and ϵ_{ATE} in synthetic datasets, the error metric PEHE is reduced by 10% on average compared with the best baseline and the ATE bias is reduced by up to 20%. Moreover, with the explicit decomposition of instrumental, confounding and adjustment variables during representation learning, the performance of DeR-CFR is much better than DR-CFR. 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.

5.3.3 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 for DR-CFR and DeR-CFR, i.e., $\bar{W}_I, \bar{W}_C, \bar{W}_A \in \mathbb{R}^m$ as described in the previous section. In high-dimensional variables and non-linear settings, N-D^{2}VD extends D^{2}VD to a non-linear version. We choose the coefficient vector (feature selection layer) of N-D^{2}VD to evaluate the separation performance of D^{2}VD-based methods. Figure 3 reports the results under the different settings with Binary_m_I_m_C_m_A_n. It is evident in Figure 3 that our DeR-CFR algorithm can precisely separate the three underlying factors $\{I, C, A\}$ and almost all separated factors correspond to their real semantics, while the baseline DR-CFR fails to disentangle those factors. In addition, although N-D^{2}VD can successfully separate adjustment variables A from low dimensional observation variables X, it fails in complex high-dimensional datasets, and it ignores the separation between instrumental variables and confounders. This result validates the motivation of the proposed DeR-CFR and is consistent with our analysis on the comparison of DeR-CFR, DR-CFR and D^{2}VD algorithms in the previous section.

Similar to the setting in DR-CFR [27], we also plot the radar charts on the representation of each factor ($\{I, C, A\}$) in Figure 4 for further comparison between DR-CFR and DeR-CFR. For example, in Figure 4(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 $\{I, C, A\}$ 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 5.

5.4 Training Cost Analysis

The above algorithms are trained based on the network model, and different network structures and constraints will increase model complexity and training cost. In all synthetic and real-world datasets, we implement 10 replications to study the average training time(s) for the proposed model in a single execution and compare it to baselines. From the results (Table 6), we have the following observations: (1) Adopting Wass distance to measure the discrepancy of representation distributions will be more time-consuming than MMD; (2) Re-weighting techniques and three representation

TABLE 6
Training time(s) of various methods in a single execution on different datasets.

Time(s)	IHDP	Jobs	Twins	Binary_8_8_8_3000	Binary_8_8_8_10000	Binary_16_16_16_3000	Binary_16_16_16_10000
D^{2}VD	76.9	82.5	102.3	83.5	88.1	84.6	98.1
N-D^{2}VD	116.8	120.1	546.1	555.4	705.9	685.4	875.4
CFR-MMD	33.8	36.2	61.8	40.5	48.4	41.1	54.6
CFR-WASS	41.6	68.7	102.7	49.2	62.2	49.5	66.5
CFR-ISW	48.7	48.2	80.2	48.6	53.7	55.2	59.5
SITE	51.5	46.8	81.7	72.4	128.5	75.4	132.6
DR-CFR	72.6	135.6	159.8	61.7	69.1	61.8	78.7
DeR-CFR	72.3	147.2	219.9	104.7	240.1	108.4	244.7

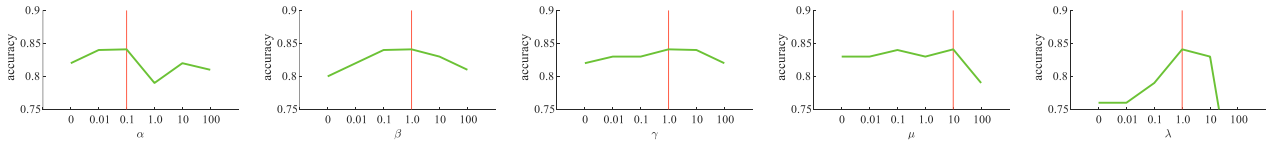


Fig. 5. Hyper-parameter sensitivity analysis on $\{\alpha, \beta, \gamma, \mu, \lambda\}$. The green lines show the accuracy of the these parameters within the specified range $\{0, 0.01, 0.1, 1.0, 10, 100\}$. The red line indicates the best parameters for the setting.

networks would increase model complexity and training cost by comparing with the results of CFR-MMD, CFR-ISW and DR-CFR; (3) As model complexity increases, model accuracy and separation performance are improved. DeR-CFR with orthogonality constraints, among representation-based algorithms, has the best estimation and separation performance but has the largest model complexity. Fortunately, its single execution time is less than 300 seconds, which is still within the acceptable range; (4) The convergence speed of N-D^{2}VD is slow, and a single convergence takes more than 600 seconds on large datasets.

Hardware configuration: 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 configuration: Python with TensorFlow 1.15.0, NumPy 1.17.4, and MatplotLib 3.1.1.

5.5 Hyper-parameters Analysis

Given the complex multi-term objective function (Section 4.6) 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 results in Figure 5 demonstrates that the performance of DeR-CFR is mostly affected by changing in α and λ, reflecting the fact that decomposing adjustment variables 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 $\{I, C, A\}$, 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 essential conditions for confounder separation. With hyper-parameters analysis, we can choose the best hyper-parameters for experiments.

5.6 Mutual Information Interpretation.

We also demonstrate the mutual information [49] with lower and upper bound under setting Binary_16_16_16_3000. The results are summarized in Table 7, 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 correlated with Y. Consistent with the results in Figure 3, the mutual information between variables $\{I, C, A\}$ 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.

TABLE 7
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

6 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 separation. Although some promising algorithms have been proposed for confounder separation/disentanglement, they cannot guarantee the decomposition of instrumental variables and confounding factor. In light of this, we propose a Decomposed representation learning algorithm for CounterFactual Regression (DeR-CFR) with explicit decomposition constraints for confounder separation 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.

Acknowledgment

This work was supported in part by National Key Research and Development Program of China (No.

2018AAA0101900), National Natural Science Foundation of China (No. 61625107, No. 62006207, No. 72171131), Key R & D Projects of the Ministry of Science and Technology (No. 2020YFC0832500), the Tsinghua University Initiative Scientific Research Grant (No. 2019THZWJC11), the Fundamental Research Funds for the Central Universities and Zhejiang Province Natural Science Foundation (No. LQ21F020020), Technology and Innovation Major Project of the Ministry of Science and Technology of China (No. 2020AAA0108400, No. 2020AAA01084020108403).

REFERENCES

[1] P. W. Holland, "Statistics and causal inference," *Journal of the American statistical Association*, vol. 81, no. 396, pp. 945–960, 1986.

[2] J. Pearl *et al.*, "Causal inference in statistics: An overview," *Statistics surveys*, vol. 3, pp. 96–146, 2009.

[3] W. Zhou, H. Xiong, L. Duan, K. Xiao, and R. Mee, "Paradoxical correlation pattern mining," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 8, pp. 1561–1574, 2018.

[4] L. Zhang, Y. Wu, and X. Wu, "Causal modeling-based discrimination discovery and removal: Criteria, bounds, and algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 11, pp. 2035–2050, 2019.

[5] S. Zhang, T. Jiang, T. Wang, K. Kuang, Z. Zhao, J. Zhu, J. Yu, H. Yang, and F. Wu, "Devlbert: Learning deconfounded visiolinguistic representations," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 4373–4382.

[6] S. Zhang, D. Yao, Z. Zhao, T. Chua, and F. Wu, "Causerec: Counterfactual user sequence synthesis for sequential recommendation," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 367–377.

[7] Y. Zhuang, M. Cai, X. Li, X. Luo, Q. Yang, and F. Wu, "The Next Breakthroughs of Artificial Intelligence: The Interdisciplinary Nature of AI," *Engineering*, vol. 6, no. 3, pp. 245–247, 2020.

[8] Y. Zhu, T. Gao, L. Fan, S. Huang, M. Edmonds, H. Liu, F. Gao, C. Zhang, S. Qi, Y. N. Wu, J. B.Tenenbaum, and S. Zhua, "Dark, Beyond Deep: A Paradigm Shift to Cognitive AI with Humanlike Common Sense," *Engineering*, vol. 6, no. 3, pp. 310–345, 2020.

[9] N. Lei, D. An, Y. Guo, K. Su, S. Liu, Z. Luo, S. Yau, and X. Gu, "A Geometric Understanding of Deep Learning," *Engineering*, vol. 6, no. 3, pp. 361–374, 2020.

[10] Y. Zhuang, F. Wu, C. Chen, and Y. Pan, "Challenges and opportunities from big data to Knowledge in AI2.0," *Frontiers of Information Technology & Electronic Engineering*, vol. 18, no. 1, pp. 3–14, 2017.

[11] X. Duan, S. Tang, S. Zhang, Y. Zhang, Z. Zhao, J. Xue, Y. Zhuang, and F. Wu, "Temporality-enhanced knowledgememory network for factoid question answering," *Frontiers of Information Technology & Electronic Engineering*, vol. 19, no. 1, pp. 104–115, 2018.

[12] R. Kohavi and R. Longbotham, "Unexpected results in online controlled experiments," *ACM SIGKDD Explorations Newsletter*, vol. 12, no. 2, pp. 31–35, 2011.

[13] L. Bottou, J. Peters, J. Quiñonero-Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson, "Counterfactual reasoning and learning systems: The example of computational advertising," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 3207–3260, 2013.

[14] P. C. Austin, "An introduction to propensity score methods for reducing the effects of confounding in observational studies," *Multivariate behavioral research*, vol. 46, no. 3, pp. 399–424, 2011.

[15] H. Bang and J. M. Robins, "Doubly robust estimation in missing data and causal inference models," *Biometrics*, vol. 61, no. 4, pp. 962–973, 2005.

[16] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.

[17] 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)*, vol. 80, no. 4, pp. 597–623, 2018.

[18] J. R. Zubizarreta, "Stable weights that balance covariates for estimation with incomplete outcome data," *Journal of the American Statistical Association*, vol. 110, no. 511, pp. 910–922, 2015.

[19] J. Hainmueller, "Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies," *Political Analysis*, vol. 20, no. 1, pp. 25–46, 2012.

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

[21] J. Pearl, "On a class of bias-amplifying variables that endanger effect estimates," *arXiv preprint arXiv:1203.3503*, 2012.

[22] 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.

[23] K. Kuang, P. Cui, H. Zou, B. Li, J. Tao, F. Wu, and S. Yang, "Data-driven variable decomposition for treatment effect estimation," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2020.

[24] J. A. Myers, J. A. Rassen, J. J. Gagne, K. F. Huybrechts, S. Schneeweiss, K. J. Rothman, M. M. Joffe, and R. J. Glynn, "Effects of adjusting for instrumental variables on bias and precision of effect estimates," *American journal of epidemiology*, vol. 174, no. 11, pp. 1213–1222, 2011.

[25] T. J. VanderWeele, "Principles of confounder selection," *European journal of epidemiology*, vol. 34, no. 3, pp. 211–219, 2019.

[26] K. Kuang, P. Cui, B. Li, M. Jiang, Y. Wang, F. Wu, and S. Yang, "Treatment effect estimation via differentiated confounder balancing and regression," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 14, no. 1, pp. 1–25, 2019.

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

[28] F. Johansson, U. Shalit, and D. Sontag, "Learning representations for counterfactual inference," in *International conference on machine learning*, 2016, pp. 3020–3029.

[29] 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*. JMLR. org, 2017, pp. 3076–3085.

[30] 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*, 2018, pp. 2633–2643.

[31] Y. Pan, "Multiple Knowledge Representation of Artificial Intelligence," *Engineering*, vol. 6, no. 3, pp. 216–217, 2020.

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

[33] P. R. Rosenbaum, "Model-based direct adjustment," *Journal of the American Statistical Association*, vol. 82, no. 398, pp. 387–394, 1987.

[34] S. Li, N. Vlassis, J. Kawale, and Y. Fu, "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*, 2016, pp. 3768–3774.

[35] Y. Liuyi, C. Zhixuan, L. Sheng, L. Yaliang, G. Jing, and Z. Aidong, "A survey on causal inference," *arXiv preprint arXiv:2002.02770*, 2020.

[36] X.-H. Li, C. C. Cao, Y. Shi, W. Bai, H. Gao, L. Qiu, C. Wang, Y. Gao, S. Zhang, X. Xue, and L. Chen, "A survey of data-driven and knowledge-aware explainable ai," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2020.

[37] K. Kuang, L. Li, Z. Geng, L. Xu, K. Zhang, B. Liao, H. Huang, P. Ding, W. Miao, and Z. Jiang, "Causal Inference," *Engineering*, vol. 6, no. 3, pp. 253–263, 2020.

[38] J. Li, S. Ma, T. Le, L. Liu, and J. Liu, "Causal decision trees," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 2, pp. 257–271, 2017.

[39] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[40] T. J. VanderWeele and I. Shpitser, "A new criterion for confounder selection," *Biometrics*, vol. 67, no. 4, pp. 1406–1413, 2011.

[41] 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*, 2019, pp. 5880–5887.

[42] A. Müller, "Integral probability metrics and their generating classes of functions," *Advances in Applied Probability*, pp. 429–443, 1997.

[43] 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.

[44] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *Journal of Machine Learning Research*, vol. 13, no. Mar, pp. 723–773, 2012.
[45] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," *arXiv preprint arXiv:1701.07875*, 2017.
[46] F. D. Johansson, D. Sontag, and R. Ranganath, "Support and invertibility in domain-invariant representations," *arXiv preprint arXiv:1903.03448*, 2019.
[47] Y. Zhang, A. Bellot, and M. van der Schaar, "Learning overlapping representations for the estimation of individualized treatment effects," *arXiv preprint arXiv:2001.04754*, 2020.
[48] X. Xu, X. Wu, F. Wei, W. Zhong, and F. Nie, "A general framework for feature selection under orthogonal regression with global redundancy minimization," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2021.
[49] P. Cheng, W. Hao, S. Dai, J. Liu, Z. Gan, and L. Carin, "Club: A contrastive log-ratio upper bound of mutual information," *arXiv preprint arXiv:2006.12013*, 2020.
[50] H. Jason, L. Greg, L. Kevin, and T Matt, "Deep IV: A flexible approach for counterfactual prediction," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2017, pp. 1414–1423.
[51] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of machine learning research*, vol. 13, no. Feb, pp. 281–305, 2012.
[52] J. L. Hill, "Bayesian nonparametric modeling for causal inference," *Journal of Computational and Graphical Statistics*, vol. 20, no. 1, pp. 217–240, 2011.
[53] R. J. LaLonde, "Evaluating the econometric evaluations of training programs with experimental data," *The American economic review*, pp. 604–620, 1986.
[54] J. A. Smith and P. E. Todd, "Does matching overcome lalonde's critique of nonexperimental estimators?" *Journal of econometrics*, vol. 125, no. 1-2, pp. 305–353, 2005.
[55] D. Almond, K. Y. Chay, and D. S. Lee, "The costs of low birth weight," *The Quarterly Journal of Economics*, vol. 120, no. 3, pp. 1031–1083, 2005.

Anpeng Wu received the B.S. degree in 2020 from the College of Science, Zhejiang University of Technology. He is a first-year Ph.D. candidate in the Department of Computer Science and Technology of Zhejiang University. His main research interests including causal inference, representation learning and reinforcement learning.

Junkun Yuan received the B.S. degree in 2019 from Zhejiang University of Technology. He is a second-year Ph.D. candidate in the Department of Computer Science and Technology of Zhejiang University. His main research interests including causal inference, domain adaptation and domain generalization.

Kun Kuang received his Ph.D. degree from Tsinghua University in 2019. He is now an Associate Professor in the College of Computer Science and Technology, Zhejiang University. He was a visiting scholar with Prof. Susan Athey's Group at Stanford University. His main research interests include Causal Inference, Artificial Intelligence, and Causally Regularized Machine Learning. He has published over 40 papers in major international journals and conferences, including SIGKDD, ICML, ACM MM, AAAI, IJCAI, TKDE, TKDD, Engineering, and ICDM, etc.

Bo Li received a Ph.D degree in Statistics from the University of California, Berkeley, and a bachelor's degree in Mathematics from Peking University. He is an Associate Professor at the School of Economics and Management, Tsinghua University. His research interests are statistical methods for high-dimensional data, statistical causal inference and data-driven decision making. He has published widely in academic journals across a range of fields including statistics, management science and economics.

Runze Wu received the B.S. degree in computer science and technology from University of Science and Technology of China, Hefei, Anhui, CN, in 2013 and the Ph.D. degree in computer science and technology from University of Science and Technology of China, Hefei, Anhui, CN, in 2018. He is currently working at NetEase Fuxi AI Lab, Hangzhou, China. His research interests include data mining, machine learning, network analysis and user profiling.

Qiang zhu received the Ph.D. degree in computer science and technology from University of California-Santa Barbara, USA, in 2007. His research interest is Computer Vision.

Yueting Zhuang received his B.Sc., M.Sc. and Ph.D. degrees in Computer Science from Zhejiang University, China, in 1986, 1989 and 1998 respectively. From February 1997 to August 1998, Yueting Zhuang was a visiting scholar at Prof. Thomas Huang's group, University of Illinois at Urbana-Champaign. Currently, He is a full professor and the Dean of the College of Computer Science, Zhejiang University. His research interests mainly include artificial intelligence, multimedia retrieval, computer animation and digital library.

Fei Wu received the B.S. degree from Lanzhou University, Lanzhou, Gansu, China, the M.S. degree from Macao University, Taipa, Macau, and the Ph.D. degree from Zhejiang University, Hangzhou, China. He is currently a full professor with the College of Computer Science and Technology, Zhejiang University. He was a Visiting Scholar with Prof. B. Yu's Group, University of California, Berkeley, from 2009 to 2010. His current research interests include multimedia retrieval, sparse representation, and machine learning.