MUTUAL INFORMATION MINIMIZATION BASED DISENTANGLED LEARNING FRAMEWORK FOR CAUSAL EFFECT ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning treatment effect from observational data is a fundamental problem in causal inference. Recently, disentangled representation learning methods, such as DR-CFR and DeR-CFR, have witnessed great success in treatment effect estimation, which aim to decompose covariates into three disjoint factors. However, we argue that these methods cannot identify underlying factors well, as they cannot obtain independent disentangled factors. Inspired by the success of mutual information minimization in disentangled representation learning, we propose a novel method called *MimCE* in this paper: **M**utual **I**nformation **M**inimization based Disentangled Learning Framework for **C**ausal Effect **E**stimation. *MimCE* mainly focuses on obtaining independent disentangled factors for treatment effect estimation and numerous experiments demonstrate that it performs better than the state-of-the-art methods both on the predictive performance and model stability.

1 INTRODUCTION

Treatment effect estimation is of the upmost importance across in many domains, such as policy making (Lalonde, 1984; Athey & Imbens, 2016), medicine prediction (Shalit et al., 2017) and advertisement (Bottou et al., 2013; Sun et al., 2015). The causal inference often needs to answer counterfactual problems (Rubin, 1974; Pearl, 2009) like *"Would this patient have low blood sugar had she received a medication?"* or *"Would the customer buy the product had he got a 70% discount"*.

One golden standard approach to learn causal effect is to carry Randomized Controlled Trail (Pearl, 2009), where the treatment assignment mechanism does not depend on the covariates and is assigned to individuals randomly. However, this method is sometimes expensive, unethical or even infeasible, thus we often focus on estimating treatment effect from observational data. But, in such dataset, the treatment depends on some attributes of individual $\boldsymbol{x}$ as $p(t|\boldsymbol{x}) \neq p(\neg t|\boldsymbol{x})$ and will cause **selection bias** (Imbens & Rubin, 2015). For example, rich customers are more willing to watch the ads and buy the expensive goods relative to the poor ones. Therefore, it is vital to find all the confounders and control them to make precise predictions, which means *unconfoundedness* assumption often needs to be satisfied in observational study to make the treatment effect identifiable (Pearl, 2009).

Even though we already have all the confounders in our variables, we face a difficult problem that we cannot easily identify them from the numerous variables and then adjust them to balance through the *backdoor criterion* (Pearl, 2009). Existing methods achieve balance either by propensity score weighting methods (Austin, 2011) or representation learning methods of reducing the discrepancy between the treated and control group (e.g., BNN (Johansson et al., 2016) and CFR-net (Shalit et al., 2017)) while ignoring identification of the other latent factors. Recently, disentangled representation learning methods, DR-CFR (Hassanpour & Greiner, 2020) and DeR-CFR (Wu et al., 2020), have been proposed to learn three independent factors $\{\Gamma, \Upsilon, \Delta\}$, which respectively represents the factor that partially affects treatment, partially affects outcome and affects both treatment and outcome. Obviously, disentangled representation learning methods can achieve explicitly identification of the latent factors. We follow this path and expect to propose more effective and robust disentangled methods, as we find that existing methods sometimes cannot disentangle well due to the inefficient design of the disentangled tasks. For example, DR-CFR cannot effectively distinguish the difference between the Δ and $\{\Gamma, \Upsilon\}$ and therefore it has not achieved competitive performance. DeR-CFR designs very complicated tasks and needs effective parameter tuning to achieve completely variable

decomposition. Besides, these two methods cannot obtain **independent** disentangled representations, which is a sufficient condition for identifying the treatment effect. Also, generative models like CEVAE (Louizos et al., 2017) and GANITE (Yoon et al., 2018) also face the problem of model complexity although they have good explanation to the data generation process.

From these perspectives, we argue that an easy-handling and well-identifying model needs to be proposed to deal with the problems mentioned above, and we also argue that a good causal inference algorithm should perform well across different model settings and different datasets without much parameter tuning. Therefore, we propose to use an efficient share-bottom encoder to extract the features and three task-specific decoders to disentangle the factors (replace the original three representations networks in DR-CFR and DeR-CFR), then we put forward a mutual information minimization framework to obtain ideally independent disentangled factors through CLUB estimator proposed by (Cheng et al., 2020) and use the independent factors to estimate treatment effect. Theoretical analysis also proves the effectiveness of our proposed method. Then, we summarize our main contributions as followings:

- We introduce a share-bottom-encoder and task-specific-decoder architecture for counterfactual inference instead of three separate representation networks.
- We extend state-of-the-art disentangled methods for counterfactual inference using mutual information minimization method to learn ideally independent disentangled representations and conduct theoretical analysis to prove the effectiveness.
- We perform sufficient experiments that show our method makes great progress in inferring individual treatment effect across several challenging datasets and robustness analysis demonstrates that *MimCE* can enhance model stability.

The rest of the paper is organized as follows: we discuss related work in Section 2, The details of our *MimCE* are presented in Section 3. In Section 4, we will talk about the details of our experiments. Finally, we show the conclusion and future work in Section 5.

2 Related works

Estimating treatment effect from observational data is a widespread concerned problem in many fields. During past decades, several kinds of methods have been proposed to solve this problem, such as propensity score based matching (Rosenbaum & Rubin, 1983; Dehejia & Wahba, 2002) and weighting (Austin, 2011) methods, some tree-based estimators like BART (Chipman et al., 2010) and Causal Forests (Wager & Athey, 2015), deep learning based methods, such as TARNET (Shalit et al., 2017), BNN (Johansson et al., 2016) and CFR (Shalit et al., 2017), which deal with the selection bias through reducing the discrepancy of the hidden embedding between the treated and control group. D^2VD (Kuang et al., 2017) proposes a kind of data-driven variables decomposition algorithm for treatment effect estimation. DR-CFR (Hassanpour & Greiner, 2020) and DeR-CFR (Wu et al., 2020) extends data-driven variables decomposition algorithm to individual treatment effect estimation (ITE) by explicitly learning three latent factors $\{\Gamma, \Upsilon, \Delta\}$ and have achieved state-of-the-art performance.

Besides, due to increasing attention of generative models in causal inference, methods like CEVAE (Louizos et al., 2017) and GANITE (Yoon et al., 2018) have also been proposed for ITE estimation, while the generative models had not achieved competitive performance until TEDVAE (Zhang et al., 2021) was proposed, which utilizes variational auto-encoder to learn three disentangled factors. As can be seen from above works, disentangled representation learning methods have played an important role in causal inference due to its effective variable decomposition and identification.

Mutual information (MI) is often utilized as a regularizer in loss functions and has widely used in many machine learning tasks, such as mutual information maximization in representation learning (Hjelm et al., 2019; Kim & Mnih, 2018) and generative models (Chen et al., 2016). Recently, mutual information minimization has also gained increasing attention in disentangled representation learning (Chen et al., 2018) and (Cheng et al., 2020) proposes a MI upper bound called CLUB to deal with the MI minimization task and various experiments have demonstrated the effectiveness of this method.

Our work is based on DR-CFR and DeR-CFR, instead of using the complicated disentangled tasks that are proposed in DeR-CFR, we borrow some ideas about hard variables decomposition from

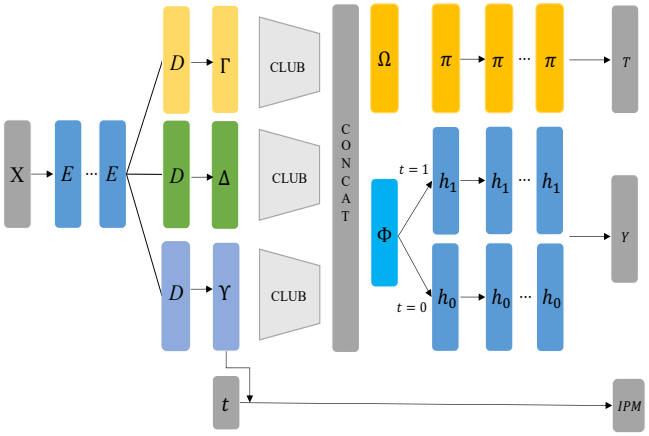


Figure 1: The proposed model architecture of *MimCE*, X refers to input features, E refers to share-bottom encoder, D refers to task-specific decoder, $\{\Gamma, \Delta, \Upsilon\}$ are disentangled factors, CLUB is the upper bound of the mutual information. $\Omega = \text{CONCAT}(\Gamma, \Delta)$, $\Phi = \text{CONCAT}(\Upsilon, \Delta)$, Υ and t are used to calculate the discrepancy IPM.

it. Then we propose a novel method called *MimCE*, which uses a share-bottom-encoder and task-specific-decoder architecture to obtain three factors and then make them independent from MI minimization task using the CLUB estimator. Experiments show that our method can obtain great improvement in several benchmarks and achieve effective variables decomposition under different settings.

3 METHODOLOGY

3.1 PRELIMINARY

We first present some notations. Given the observational dataset $\mathcal{D} = \left\{\left(\boldsymbol{x}_i, t_i, y_i^{t_i}\right)\right\}_{i=1}^{N}$, N is the number of the data samples, $\left(x_i, t_i, y_i^{t_i}\right)$ is the labeled data pairs, $\boldsymbol{x}_i \in \mathcal{X}$ is the input features referring context information, $y_i \in \mathcal{Y}$ is observed factual outcome and $t_i \in \mathcal{T}$ refers to potential interventions (e.g., for binary treatment $t \in \{0, 1\}$). For example, for decision on medicine usage, $\boldsymbol{x}$ can be age, gender, economic status and current situation of patients, y can be recovery time or whether to recover, t can be whether to take the medicine. Mathematically, we define our goal in this paper is to learn a function $\mathcal{F} : \mathcal{X} \times \mathcal{T} \to \mathcal{Y}$ to predict the potential outcomes and then estimate the *individual treatment effect (ITE)* [1] and the *average treatment effect (ATE):*

Definition 1. *The individual treatment effect (ITE) is formulated as*:

$$ITE = \mathbb{E}[y = 1|\boldsymbol{x}, do(t = 1)] - \mathbb{E}[y = 1|\boldsymbol{x}, do(t = 0)] \tag{1}$$

Definition 2. *The average treatment effect (ATE) is formulated as*:

$$ATE = \mathbb{E}[y = 1|do(t = 1)] - \mathbb{E}[y = 1|do(t = 0)] \tag{2}$$

Where $do(t)$ refers to remove all incoming edges of t (Pearl, 2009).

Besides, we assume that the following fundamental assumptions for treatment effect estimation are satisfied in this paper (Rosenbaum & Rubin, 1983):

Assumption 1. (***SUTVA***) *The Stable Unit Treatment Value Assumption requires that the response of a unit depends only on the treatment to which he himself was assigned and not affected by others.*

Assumption 2. (***Unconfoundedness***) *The treatment assignment mechanism is independent of the potential outcome when conditioning on the observed variables, Formally*: $Y_0, Y_1 \perp\!\!\!\perp t \mid \boldsymbol{x}$.

[1] The individual treatment effect (ITE) *aka* called conditional average treatment effect (CATE).

Assumption 3. (***Positivity***) *Each unit has a non-zero probability to be assigned to each treatment when given the observed contexts, i.e.,* $0 < P(t = 1|\boldsymbol{x}) < 1$.

3.2 PROPOSED METHOD

The model architecture of *MimCE* is shown in Figure 1. It is consist of three components, which are share-bottom-encoder and task-specific-decoder part, CLUB mutual information minimization part and predictions task part (i.e., outcome prediction task, treatment prediction task and imbalance prediction task respectively). We introduce the details of our method as followings.

3.2.1 DISENTANGLED REPRESENTATIONS FOR COUNTERFACTUAL INFERENCE

Without loss of generality, for the dataset $\mathcal{D} = \left\{\left(x_i, t_i, y_i^{t_i}\right)\right\}_{i=1}^{N}$, similar to DR-CFR (Hassanpour & Greiner, 2020), we argue that it is generated from three underlying factors $\{\Gamma, \Delta, \Upsilon\}$. Then we aim to recover the underlying factors from the input variables by defining three prediction tasks as follows:

- Predict outcomes from $\Phi = \text{CONCAT}(\Upsilon, \Delta)$ and define the loss as: $\mathcal{L}_{\text{pred}} = \mathcal{L}[y_i, h^{t_i}(\Phi(x_i))]$.
- Predict treatment from $\Omega = \text{CONCAT}(\Gamma, \Delta)$ and define the loss as: $\mathcal{L}_{\text{treat}} = \mathcal{L}[t_i, \pi(\Omega(x_i))]$.
- Calculate discrepancy loss by $\mathcal{L}_{\text{disc}} = \text{IPM}\left(\{\Upsilon(x_i)\}_{i:t_i=0}, \{\Upsilon(x_i)\}_{i:t_i=1}\right)$.

Then we summarize the loss function $\mathcal{L}_{\text{MAIN}}$ based on the three base tasks:

$$\mathcal{L}_{\text{MAIN}} = \mathcal{L}_{\text{pred}} + \alpha \cdot \mathcal{L}_{\text{treat}} + \beta \cdot \mathcal{L}_{\text{disc}} \tag{3}$$

We also find that the sample weight ω_i used in DR-CFR is often harm to treatment effect estimation, the reason may due to the flowing of some information of Υ into Δ, which will cause biased estimator of treatment assignment mechanism. Therefore, we exclude the weight as we cannot guarantee perfectly disentanglement between Δ and Υ, and the weight may hurt the robustness of our model in unknown real environment (i.e., we can utilize the weight if and only if we disentangle Υ from Δ thoroughly). Besides, α and β are weights for each task, and we use *Wasserstein distance* as our integral probability metric in this paper.

3.2.2 SHARE-BOTTOM ENCODER AND TASK-SPECIFIC DECODER

As we discussed above, efficient treatment effect estimation models should effectively disentangle the factors first. Most of existing disentangled methods use three isolated representation learning networks to extract underlying factors[2], we argue that this method is lack of efficiency because of redundant parameters and little information exchange between tasks. Instead, we adopt a share-bottom-encoder and task-specific-decoder architecture to extract the latent factors, which can be formulated as:

$$\begin{gathered} \boldsymbol{z}_h = E(\boldsymbol{x}) \\ \boldsymbol{z}_\Gamma = D_\Gamma(\boldsymbol{z}_h), \boldsymbol{z}_\Upsilon = D_\Upsilon(\boldsymbol{z}_h), \boldsymbol{z}_\Delta = D_\Delta(\boldsymbol{z}_h) \end{gathered} \tag{4}$$

$\boldsymbol{x}$ is input features and $\boldsymbol{x} \subseteq \mathbb{R}^d$, d refers to dimension of input features. $\boldsymbol{z}_h$ is hidden embedding and $\boldsymbol{z}_h \subseteq \mathbb{R}^k$, k refers to dimension of the hidden units. $\boldsymbol{z}_\Gamma, \boldsymbol{z}_\Delta, \boldsymbol{z}_\Upsilon$ are representations of the underlying factors respectively and $\boldsymbol{z}_\Gamma, \boldsymbol{z}_\Delta, \boldsymbol{z}_\Upsilon \subseteq \mathbb{R}^d$ (i.e., consistent with input dimension). The E refers to a share-bottom-encoder with multi-layer MLPs and the D is a task-specific-decoder with an one-layer MLP (we use ReLU activation in decoder to obtain non-negative factors).

The intuition of the architecture actually derives from multi-task learning (Ma et al., 2018). We can treat the disentangling process as three single prediction tasks, then we can get more time-efficient and memory-efficient model through parameter sharing on the bottom layers, and sharing information among different tasks can further enhance feature representation. Hence, instead of disentangling the factors directly from input features, we disentangle them from a common hidden feature space. Experiments show that the proposed encoder-decoder architecture performs better than three isolated representation learning networks significantly.

[2]DR-CFR describes its representations of the underlying factors as *"Three representation learning networks; one for each underlying factor:* $\Gamma(x), \Delta(x), \Upsilon(x)$" and DeR-CFR describes it as *"Three decomposed representation networks for learning latent factors, one for each underlying factor: I(X), C(X) and A(X)*".

3.2.3 MUTUAL INFORMATION MINIMIZATION FRAMEWORK

Through disentangled representation learning, we aim to encode the hidden features into several separate embedding parts. In order to learn independent disentangled representations in ITE estimation, DeR-CFR uses orthogonal regularizer between product of weight matrices to achieve completely variables decomposition. However, our experiment shows that this method may become less effective when representation layer goes deeper[3]. Besides, it can only acquire uncorrelated factors instead of independent factors, as orthogonal regularizer can be approximately treated as calculating *Pearson's Correlation Coefficient* ρ[4]. In order to obtain independent disentangled factors, we apply recent work on mutual information minimization to our ITE estimation method and propose to minimize the mutual information among the three underlying factors to ensure independence. Mutual information is a fundamental measure of the dependence between two random variables. Mathematically, the definition of MI between variables $\boldsymbol{x}$ and $\boldsymbol{y}$ is:

$$\mathrm{I}(\boldsymbol{x}, \boldsymbol{y}) = \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}\left[\log \frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})}\right] \tag{5}$$

Following (Cheng et al., 2020), we introduce to use **C**ontrastive **L**og-ratio **U**pper **B**ound (CLUB) to accomplish mutual information minimization among underlying factors, and CLUB is defined as $\mathrm{I}_{\mathrm{CLUB}}(\boldsymbol{x}, \boldsymbol{y}) = \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}\left[\log p(\boldsymbol{y}|\boldsymbol{x})\right] - \mathbb{E}_{p(\boldsymbol{x})}\mathbb{E}_{p(\boldsymbol{y})}\left[\log p(\boldsymbol{y}|\boldsymbol{x})\right]$ when the conditional distribution $p(\boldsymbol{y}|\boldsymbol{x})$ is known. We describe the properties of CLUB in the following theorem:

Theorem 1. *For two random variables $\boldsymbol{x}$ and $\boldsymbol{y}$, $I_{\mathrm{CLUB}}(\boldsymbol{x}, \boldsymbol{y})$ is an upper bound of $I(\boldsymbol{x}, \boldsymbol{y})$:*

$$\mathrm{I}(\boldsymbol{x}, \boldsymbol{y}) \leq \mathrm{I}_{\mathrm{CLUB}}(\boldsymbol{x}, \boldsymbol{y}) \tag{6}$$

Equality is achieved if and only if $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent.

Unfortunately, the conditional relation between variables is unavailable in our task, then we use a variational distribution $q_\theta(\boldsymbol{y}|\boldsymbol{x})$ to approximate $p(\boldsymbol{y}|\boldsymbol{x})$ to further extend the CLUB estimator into our scenario. vCLUB is defined as $\mathrm{I}_{\mathrm{vCLUB}}(\boldsymbol{x}, \boldsymbol{y}) = \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}\left[\log q_\theta(\boldsymbol{y}|\boldsymbol{x})\right] - \mathbb{E}_{p(\boldsymbol{x})}\mathbb{E}_{p(\boldsymbol{y})}\left[\log q_\theta(\boldsymbol{y}|\boldsymbol{x})\right]$ and has the following properties:

Theorem 2. *The variational CLUB term $I_{\mathrm{vCLUB}}(\boldsymbol{x}, \boldsymbol{y})$ remains a MI upper bound if the variational joint distribution $q_\theta(\boldsymbol{x}, \boldsymbol{y}) = q_\theta(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{x})$ satisfy the following inequality:*

$$\mathrm{KL}(p(\boldsymbol{x}, \boldsymbol{y})||q_\theta(\boldsymbol{x}, \boldsymbol{y})) \leq \mathrm{KL}(p(\boldsymbol{x})p(\boldsymbol{y})||q_\theta(\boldsymbol{x}, \boldsymbol{y})) \tag{7}$$

Which means that vCLUB can still hold a MI upper bound when we have good variational approximation $q_\theta(\boldsymbol{y}|\boldsymbol{x})$. The proofs of Theorem 1 and 2 are available in Appendix A1.

Since we aim to obtain independent disentangled representations, we simply minimize the following objective function:

$$\mathcal{L}_{\mathrm{CLUB}} = \mathrm{I}_{\mathrm{vCLUB}}(\boldsymbol{z}_\Gamma, \boldsymbol{z}_\Delta) + \mathrm{I}_{\mathrm{vCLUB}}(\boldsymbol{z}_\Delta, \boldsymbol{z}_\Upsilon) + \mathrm{I}_{\mathrm{vCLUB}}(\boldsymbol{z}_\Upsilon, \boldsymbol{z}_\Gamma) \tag{8}$$

We conclude that we can acquire pairwise independent disentangled factors through minimizing $\mathcal{L}_{\mathrm{CLUB}}$, as shown in the following theorem:

Theorem 3. *If the objective function $\mathcal{L}_{\mathrm{CLUB}}$ is minimized to zero, then we have*:

$$p(\boldsymbol{z}_\Gamma, \boldsymbol{z}_\Delta) = p(\boldsymbol{z}_\Gamma)p(\boldsymbol{z}_\Delta), \quad p(\boldsymbol{z}_\Delta, \boldsymbol{z}_\Upsilon) = p(\boldsymbol{z}_\Delta)p(\boldsymbol{z}_\Upsilon), \quad p(\boldsymbol{z}_\Upsilon, \boldsymbol{z}_\Gamma) = p(\boldsymbol{z}_\Upsilon)p(\boldsymbol{z}_\Gamma) \tag{9}$$

That is, $\boldsymbol{z}_\Delta$, $\boldsymbol{z}_\Gamma$ and $\boldsymbol{z}_\Upsilon$ are pairwise independent.

Proof. Due to the $\mathrm{I}(\boldsymbol{x}, \boldsymbol{y})$ is positive semi-definite, then minimizing $\mathcal{L}_{\mathrm{CLUB}}$ is equal to minimize each part of it to zero. Without loss of generality, we take $\mathrm{I}_{\mathrm{vCLUB}}(\boldsymbol{z}_\Gamma, \boldsymbol{z}_\Delta)$ as an example and according to the Theorem 1 and 2:

$$0 \leq \mathrm{I}(\boldsymbol{z}_\Gamma, \boldsymbol{z}_\Delta) \leq \mathrm{I}_{\mathrm{vCLUB}}(\boldsymbol{z}_\Gamma, \boldsymbol{z}_\Delta) \tag{10}$$

Ideally, if the MI upper bound between each two variables is minimized to zero, we have:

$$\mathrm{I}(\boldsymbol{z}_\Gamma, \boldsymbol{z}_\Delta) = \sum_{\boldsymbol{z}_\Gamma}\sum_{\boldsymbol{z}_\Delta} p(\boldsymbol{z}_\Gamma, \boldsymbol{z}_\Delta) \log \frac{p(\boldsymbol{z}_\Gamma, \boldsymbol{z}_\Delta)}{p(\boldsymbol{z}_\Gamma)p(\boldsymbol{z}_\Delta)} = 0 \Leftrightarrow p(\boldsymbol{z}_\Gamma, \boldsymbol{z}_\Delta) = p(\boldsymbol{z}_\Gamma)p(\boldsymbol{z}_\Delta) \tag{11}$$

[3] Figure 4 shows the evidence that the performance will drop for IHDP-A when encoder goes deeper.

[4] Orthogonal regularizer refers to $E(\boldsymbol{x}\boldsymbol{y}) = 0$, if we simply assume that $E(\boldsymbol{x}) = E(\boldsymbol{y}) = 0$, then orthogonal regularizer is equal to $\rho = 0$, i.e., uncorrelated.

*Then we conclude that $\boldsymbol{z}_\Gamma$ and $\boldsymbol{z}_\Delta$ are **independent**. In the same way, $\boldsymbol{z}_\Delta$, $\boldsymbol{z}_\Gamma$ and $\boldsymbol{z}_\Upsilon$ are **pairwise independent**.*

Based on the results that we obtain independent disentangled factors from theorem 3, we demonstrate that individual treatment effect can be identified through following theorem:

Theorem 4. *The individual treatment effect is identifiable if we obtain the independent disentangled representations $\boldsymbol{z}_\Gamma$, $\boldsymbol{z}_\Delta$ and $\boldsymbol{z}_\Upsilon$ from $\boldsymbol{x}$.*

Proof. Firstly, we define some notations and rules described in (Pearl, 2009):

Let $\mathcal{G}$ be the directed acyclic graph, $p(\cdot)$ stand for the probability distribution. We denote $\mathcal{G}_{\overline{t}}$ by the graph obtained by deleting from $\mathcal{G}$ all edges pointing into t. Likewise, we denote $\mathcal{G}_{\underline{t}}$ by the graph obtained by deleting from $\mathcal{G}$ all edges emerging from t. For any disjoint subsets of variables t, y, $\boldsymbol{x}$ and $\boldsymbol{z}$, we have the following rules.

Rule 1. (Insertion/deletion of observations)

$$p(y|do(t), \boldsymbol{x}, \boldsymbol{z}) = p(y|do(t), \boldsymbol{x}) \qquad if \quad (y \perp\!\!\!\perp \boldsymbol{z} \mid t, \boldsymbol{x})_{\mathcal{G}_{\overline{t}}} \tag{12}$$

Rule 2. (Action/observation exchange)

$$p(y|do(t), \boldsymbol{x}, \boldsymbol{z}) = p(y|t, \boldsymbol{x}, \boldsymbol{z}) \qquad if \quad (y \perp\!\!\!\perp t \mid \boldsymbol{x}, \boldsymbol{z})_{\mathcal{G}_{\underline{t}}} \tag{13}$$

Then, we denote $\hat{\tau}$ as an estimator of the ITE and use independetn factors $\boldsymbol{z}_\Gamma, \boldsymbol{z}_\Delta, \boldsymbol{z}_\Upsilon$ to replace $\boldsymbol{x}$,

$$\hat{\tau} = \hat{p}(y = 1|\boldsymbol{z}_\Gamma, \boldsymbol{z}_\Delta, \boldsymbol{z}_\Upsilon, do(t = 1)) - \hat{p}(y = 1|\boldsymbol{z}_\Gamma, \boldsymbol{z}_\Delta, \boldsymbol{z}_\Upsilon, do(t = 0)) \tag{14}$$

Following Rule 1, we can remove z_Γ in equation (14) knowing that $(y \perp\!\!\!\perp z_\Gamma \mid t, z_\Delta, z_\Upsilon)_{\mathcal{G}_{\overline{t}}}$,

$$\hat{\tau} = \hat{p}(y = 1|\boldsymbol{z}_\Delta, \boldsymbol{z}_\Upsilon, do(t = 1)) - \hat{p}(y = 1|\boldsymbol{z}_\Delta, \boldsymbol{z}_\Upsilon, do(t = 0)) \tag{15}$$

Besides, we have $(y \perp\!\!\!\perp t \mid z_\Upsilon, z_\Delta)_{\mathcal{G}_{\underline{t}}}$ and using Rule 2,

$$\hat{\tau} = \hat{p}(y = 1|\boldsymbol{z}_\Delta, \boldsymbol{z}_\Upsilon, t = 1) - \hat{p}(y = 1|\boldsymbol{z}_\Delta, \boldsymbol{z}_\Upsilon, t = 0) \tag{16}$$

Therefore, the individual treatment effect is identifiable when we condition on $\boldsymbol{z}_\Delta$ and $\boldsymbol{z}_\Upsilon$.

While in some scenarios, people assume that it is sufficient to estimate ITE under *unconfoundedness* assumption and control all variables in model. However, (Zhang et al., 2021; Pearl, 2009; Abadie & Imbens, 2004; Hahn, 1998) illustrate that conditioning on redundant variables that are uncorrelated to outcome may bring **biased** as well as **high-variance** estimator. Hence, it is significant to estimate ITE only through the disentangled factors z_Δ and z_Υ.

Besides, we also compare it to other disentangled methods to prove the effectiveness of MI minimization in ITE estimation, one is from the idea of matrix orthogonal regularity used in DeR-CFR (Wu et al., 2020), called **W**eight **M**atrix **O**rthogonality (WMO), and the other one is to directly restrict the inner product of the three factors to zero, which is abbreviated as **I**nner **P**roduct **O**rthogonality (IPO). WMO and IPO are defined as:

$$\begin{cases} \mathcal{L}_{\text{WMO}} = \bar{\boldsymbol{W}}_\Gamma^T \cdot \bar{\boldsymbol{W}}_\Delta + \bar{\boldsymbol{W}}_\Delta^T \cdot \bar{\boldsymbol{W}}_\Upsilon + \bar{\boldsymbol{W}}_\Upsilon^T \cdot \bar{\boldsymbol{W}}_\Gamma \\ \mathcal{L}_{\text{IPO}} = \boldsymbol{z}_\Gamma^T \cdot \boldsymbol{z}_\Delta + \boldsymbol{z}_\Delta^T \cdot \boldsymbol{z}_\Upsilon + \boldsymbol{z}_\Upsilon^T \cdot \boldsymbol{z}_\Gamma \end{cases} \tag{17}$$

$\boldsymbol{W} \subseteq \mathbb{R}^{d\times d}$ refers to products of the encoder and decoder layers, then we use the average of the $\boldsymbol{W}$ and get $\bar{\boldsymbol{W}} \subseteq \mathbb{R}^{d\times 1}$ to represent the contribution of input variables on disentangled factors. To ensure we get non-zero disentangled factors, we also restrict the following objectives:

$$\mathcal{L}_{\text{WREG}} = \theta \cdot \sum_i (1 - \sum_j \omega_{ij})^2 + (1 - \theta) \cdot \sum_i (1 - \sum_j \upsilon_{ij})^2 \tag{18}$$

Where $i \in \{\Gamma, \Delta, \Upsilon\}$ and ω_{ij} and υ_{ij} represent each dimension of $\bar{\boldsymbol{W}}$ and $\boldsymbol{z}$ respectively, θ is a parameter to balance these two parts.[5] Then we summarize the total objective function $\mathcal{L}_{MimCE}$ as:

$$\mathcal{L}_{MimCE} = \underbrace{\mathcal{L}_{\text{pred}} + \alpha \cdot \mathcal{L}_{\text{treat}} + \beta \cdot \mathcal{L}_{\text{disc}}}_{\mathcal{L}_{\text{MAIN}}} + \underbrace{\gamma \cdot \mathcal{L}_{\text{CLUB}} + \eta \cdot \mathcal{L}_{\text{WREG}}}_{\mathcal{L}_{\text{MIM}}} + \lambda \cdot \mathcal{L}_{\text{REG}} \tag{19}$$

We combine $\mathcal{L}_{\text{CLUB}}$ and $\mathcal{L}_{\text{WREG}}$ as total MI minimization objective function $\mathcal{L}_{\text{MIM}}$. $\mathcal{L}_{\text{REG}}$ penalizes the model complexity and α, β, γ, η and λ are weights for these objectives.

[5]Empirically, we usually set $\theta = 0$ when encoder layers are deep versus $\theta = 1$ when encoder layers are shallow, as the products of weight matrices may omit some information when layer goes deeper.

4 EXPERIMENT

4.1 BENCHMARK DATASET

A fundamental problem in causal inference is that we cannot observe factual outcome and counterfactual outcome simultaneously, an often used solution is to synthesize datasets where the outcomes of all possible treatments are available or synthesize outcomes from real-worlds covariates. We use Infant Health and Development Program (IHDP) (Hill, 2011) and our synthetic dataset in this paper.

IHDP Benchmark. Similar to (Shalit et al., 2017; Hassanpour & Greiner, 2019; Zhang et al., 2021), we use a semi-synthetic dataset based on the Infant Health and Development Program (IHDP) as our benchmark which was first introduced by (Hill, 2011). The covariates come from a randomized experiment studying the effects of home visits by specialist on future cognitive test scores. The selection bias has been made by removing a biased subset of the treated population and it comprises 747 instances (139 treated, 608 control) with 25 covariates measuring different attributes of children and their mothers. The simulated outcomes are implemented as both setting "A" and setting "B" in the NPCI package and follow *linear* relationship and *nonlinear* relationship respectively.

Synthetic Benchmark. Given the observational dataset $\mathcal{D} = \left\{\left(\boldsymbol{x}_i, t_i, y_i^{t_i}\right)\right\}_{i=1}^{N}$, N is the number of the data samples. We assume $\boldsymbol{x}$ is generated from underlying factors $\{\Gamma, \Delta, \Upsilon\}$, k_Γ, k_Δ and k_Υ refer to dimension of these factors. Then we generate each datapoint from following 3 steps: 1). For each $\boldsymbol{x}_i \in \{\Gamma, \Delta, \Upsilon\}$, we generate samples from independent normal distributions $\boldsymbol{x}_i \sim \mathcal{N}(0, 1)$. 2). We create selection bias by defining $\pi(t = 1|\boldsymbol{z}) = 1/(1 + e^{-\boldsymbol{\theta}_t \cdot \Omega})$, where $\Omega = \text{CONCAT}(\Gamma, \Delta) + \varepsilon$, $\boldsymbol{\theta}_t \sim \mathcal{U}((2, 4)^{k_\Gamma + k_\Delta})$ and $\varepsilon \sim \mathcal{N}(0, 1)$. Then we sample t_i from $Bern(\pi(t = 1|\boldsymbol{z}_i))$ for each unit i. 3). The outcomes $Y_0 = \boldsymbol{\theta}_{y_0} \cdot \Phi^3/(k_\Gamma + k_\Delta) + \varepsilon$ and $Y_1 = \boldsymbol{\theta}_{y_1} \cdot \Phi^2/(k_\Upsilon + k_\Delta) + \varepsilon$, where $\Phi = \text{CONCAT}(\Upsilon, \Delta) + \varepsilon$, $\boldsymbol{\theta}_{y_0}, \boldsymbol{\theta}_{y_1} \sim \mathcal{U}((2, 4)^{k_\Upsilon + k_\Delta})$ and $\varepsilon \sim \mathcal{N}(0, 1)$. Finally, we repeat the procedure N times to generate train datasets.

4.2 EXPERIMENTAL RESULTS OF TREATMENT EFFECTS

Performance Metrics. Given a synthetic dataset that includes both factual and counterfactual outcomes, we can evaluate treatment effect estimation methods with two performance measures. The individual-based performance metric is $\epsilon_{PEHE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{\tau}_i - \tau_i)^2$, where $\hat{\tau}_i = \hat{y}_i^1 - \hat{y}_i^0$ is the predicted individual treatment effect and $\tau_i = y_i^1 - y_i^0$ is the actual effect. The population-based performance measure is $\epsilon_{ATE} = |\text{ATE} - \widehat{\text{ATE}}|$. The $\text{ATE} = \frac{1}{N}\sum_{i=1}^{N}(y_i^1 - y_i^0)$ and the $\widehat{\text{ATE}}$ is calculated from the estimated outcomes.

Baselines Methods. We will compare performances of the following treatment effect estimation methods in this paper, which can be divided into the following categories. *Baseline models*: **TARNET** (Shalit et al., 2017), **CFR-WASS** (Shalit et al., 2017), **CFR-ISW** (Hassanpour & Greiner, 2019), **SITE** (Yao et al., 2018). *Generative models*: **CEVAE** (Louizos et al., 2017), **GANITE** (Yoon et al., 2018), **TEDVAE** (Zhang et al., 2021). *Disentangled models*: **DR-CFR** (Hassanpour & Greiner, 2020), **DeR-CFR** (Wu et al., 2020), **MimCE** and its two variants **IpoCE** and **WmoCE** (i.e., we obtain *IpoCE* and *WmoCE* by replacing the $\mathcal{L}_{\text{CLUB}}$ with $\mathcal{L}_{\text{IPO}}$ and $\mathcal{L}_{\text{WMO}}$). The details of parameter settings are available in Appendix A.2.

Ablation Study. We also conduct an ablation study to examine the contributions of different components in *MimCE*. ***w/o* ED**: remove the share-bottom-encoder and task-specific-decoder. ***w/o* MIM**: remove MI minimization task. ***w/o* MIM+ED**: remove MIM and ED both.

In Table 1, we report the average results of the $\sqrt{\epsilon_{PEHE}}$ and ϵ_{ATE} metrics on IHDP-A and IHDP-B benchmarks (100 realizations with 63/27/10 proportion of train/validation/test splits), results show that *MimCE* achieves the best performance among the compared methods and its two variants, which demonstrate that *MimCE* is currently the most effective disentangled method in ITE estimation. The bottom part of Table 1 summarizes the results of the ablation study, from which we observe that all *MimCE* variants with some components removed witness clear performance drops when comparing to the full model on the $\sqrt{\epsilon_{PEHE}}$ metric, suggesting that each of the designed components contributes to the success of *MimCE*. Table 2 shows the average results of 10 replications on synthetic benchmarks with $k_\Gamma = k_\Delta = k_\Upsilon = 6$ and $N = 10000$.

Table 1: Results of different treatment effect estimation methods and ablation study of *MimCE*

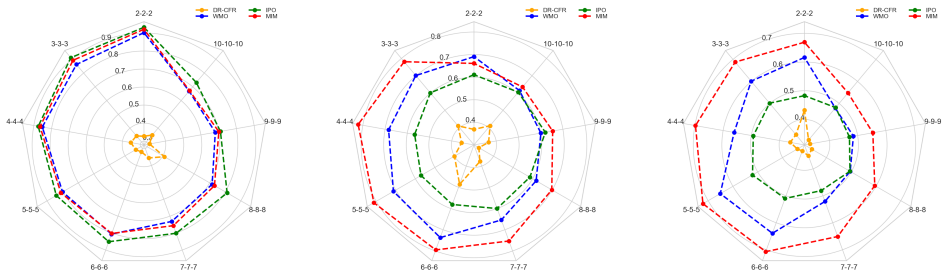
DATASET	IHDP-A				IHDP-B			
METHOD	In.S		Out.S		In.S		Out.S	
	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}
TARNET	0.886	0.231	0.951	0.275	2.512	0.238	3.156	0.354
CFR-WASS	0.729	0.263	0.742	0.296	2.379	0.247	2.518	0.342
CFR-ISW	0.682	0.224	0.691	0.230	2.350	0.291	2.558	0.401
SITE	0.675	0.211	0.683	0.241	2.265	0.261	2.341	0.367
CEVAE	1.831	0.343	2.052	0.391	2.883	0.358	3.214	0.426
GANITE	2.150	0.391	2.429	0.452	3.512	0.435	3.971	0.538
TEDVAE	0.562	0.143	0.586	0.151	2.071	0.228	2.242	0.316
DR-CFR	0.632	0.157	0.643	0.201	2.178	0.279	2.330	0.384
DeR-CFR	0.468	0.136	0.488	0.155	2.091	0.247	2.208	0.329
WmoCE	0.528	0.123	0.583	0.142	2.023	0.216	2.160	0.319
IpoCE	0.446	0.095	0.469	0.103	2.037	0.221	2.191	0.340
MimCE	**0.369***	**0.089***	**0.378***	**0.092***	**1.985***	**0.210***	**2.118***	**0.314***
w/o ED	0.518	0.122	0.546	0.137	2.136	0.241	2.280	0.339
w/o MIM	0.489	0.128	0.505	0.135	2.131	0.235	2.298	0.321
w/o ED+MIM	0.611	0.145	0.650	0.176	2.164	0.267	2.312	0.370

[1] The **bolded** values mean the best performance and * means significantly different from DR-CFR (t-test, $\alpha = 0.05$). In.S means train/validation dataset and Out.S means test dataset.

4.3 IDENTIFICATION OF THE DISENTANGLED FACTORS

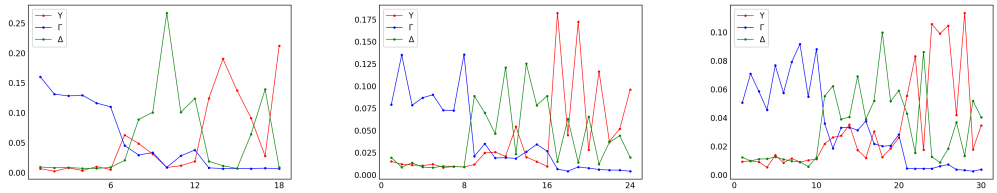
As we mentioned above, we use the products of the weight matrices of the encoder layers and decoder layers to evaluate if our method can achieve effective variables decomposition. Then we calculate the average of the weight matrix and get $\bar{\boldsymbol{W}} \subseteq \mathbb{R}^{d\times 1}$ and each dimension of the $\bar{\boldsymbol{W}}$ can approximately represent the contribution of each dimension of input variables on disentangled factors. We take Δ as an example and partition $\bar{\boldsymbol{W}}$ into two parts $\bar{\boldsymbol{W}}_{\Delta}$ and $\bar{\boldsymbol{W}}_{\neg\Delta}$. Then we sum the absolute values of weights in $\bar{\boldsymbol{W}}_{\Delta}$ and $\bar{\boldsymbol{W}}_{\neg\Delta}$, and calculate the percentage $M\% = \sum \bar{\boldsymbol{W}}_{\Delta} / (\sum \bar{\boldsymbol{W}}_{\Delta} + \bar{\boldsymbol{W}}_{\neg\Delta})$ to denote *"How much effective information is embedded in the disentangled factors, much is better"*.

However, $\bar{\boldsymbol{W}}$ is not the most suitable method to evaluate *WmoCE*, *IpoCE* and *MimCE* simultaneously (as only *WmoCE* directly optimizes the $\bar{\boldsymbol{W}}$). To make the results more explainable and comparable, we add $\mathcal{L}_{\text{WMO}}$ in *IpoCE* and *MimCE* (*aka* called *IpoCE* (+*wmo*) and *MimCE* (+*wmo*) in Table 3). If *IpoCE* and *MimCE* are more efficient, then we expect that the percentage $M\%$ of *IpoCE* (+*wmo*) and *MimCE* (+*wmo*)) should be higher than that of *WmoCE*.



(a) Identification of Γ (b) Identification of Υ (c) Identification of Δ

Figure 2: Radar charts that visualize the capability of DR-CFR, *WmoCE*, *IpoCE* (+*wmo*) and *MimCE* (+*wmo*) in identifying the factors $\{\Gamma, \Upsilon, \Delta\}$, The polygons' radii denotes $M\%$ and each vertex on the polygons refers to factors' dimension (k_{Δ}, k_{Γ} and k_{Υ}) of synthetic dataset.



(a) 6-6-6 (b) 8-8-8 (c) 10-10-10

Figure 3: Examples of *MimCE* (+*wmo*) in identifying the factors $\{\Gamma, \Upsilon, \Delta\}$. The x-axis and y-axis denote variable dimension k_i and percentage of each dimension on $\hat{\boldsymbol{W}}$ respectively. k-k-k refers to factors' dimension, $1 \sim k$ refers to Γ, $k+1 \sim 2k$ refers to Δ and $2k+1 \sim 3k$ refers to Υ.

The average results of 9×50 replications on synthetic benchmarks are shown in Figure 2 and Table 3, from which we can observe that *MimCE* (+*wmo*) can achieve a good identification on $\{\Gamma, \Upsilon, \Delta\}$, especially on $\{\Upsilon, \Delta\}$. Figure 3 shows examples of *MimCE* (+*wmo*) in identifying the factors.

Table 2: Out.s $\sqrt{\epsilon_{PEHE}}$ and ϵ_{ATE} results on synthetic benchmarks, represented in the form of "mean (standard deviation)"

METHOD	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}
DR-CFR	1.62 (0.08)	0.15 (0.05)
WmoCE	1.17 (0.06)	0.11 (0.04)
IpoCE	1.22 (0.06)	0.10 (0.02)
MimCE	**1.02 (0.05)**	**0.08 (0.02)**

Table 3: Identification of the underlying factors on synthetic benchmarks, represented in the form of "$M\%$"

METHOD	Γ	Υ	Δ
DR-CFR	34.4%	38.9%	35.3%
WmoCE	77.9%	66.8%	56.1%
IpoCE (+*wmo*)	**82.9%**	49.2%	44.2%
MimCE (+*wmo*)	79.5%	**73.4%**	**64.8%**

4.4 ROBUSTNESS ANALYSIS

It is often a difficult thing to evaluate the treatment effect estimation in real-world scenarios, so a good treatment effect estimation algorithm should perform well across different model settings. Then, we attempt to conduct robustness analysis based on changing the size of the encoder layer.

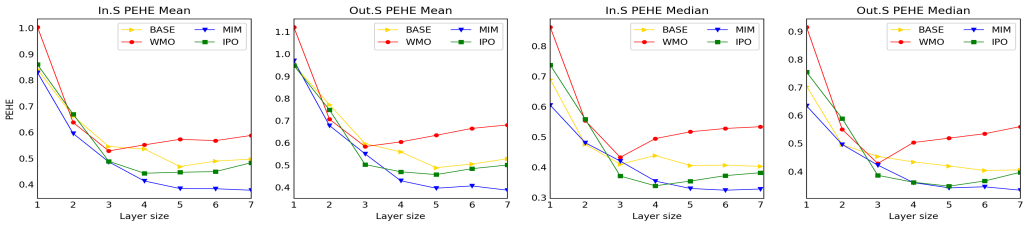


Figure 4: Robustness analysis of the influence of layer size on the $\sqrt{\epsilon_{PEHE}}$ metric for IHDP-A, from which we observe that *MimCE* performs more robust than other methods as the blue line shows.

5 CONCLUSION

We mainly focus on disentangled representation learning methods for ITE estimation in this paper, compared to existing disentangled methods, like DR-CFR, DeR-CFR and TEDVAE, we propose a MI minimization disentangled framework for causal effect estimation called *MimCE*, which not only has excellent predictive performance and disentangled performance but also owns good model stability. And many experiments demonstrate that *MimCE* is a state-of-the-art method.

For future work, we aim to extend our method to multi-treatment scenarios as existing models mainly focus on binary treatment. It is also a promising direction to utilize the MI minimization method in generative models as increasing attention have been paid to this field.

REFERENCES

Alberto Abadie and G. Imbens. Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74:235–267, 2004.

S. Athey and G. Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113:7353 – 7360, 2016.

P. Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46:399 – 424, 2011.

L. Bottou, J. Peters, J. Q. Candela, D. Charles, D. M. Chickering, Elon Portugaly, Dipankar Ray, P. Simard, and Edward Snelson. Counterfactual reasoning and learning systems: the example of computational advertising. *J. Mach. Learn. Res.*, 14:3207–3260, 2013.

T. Chen, Xuechen Li, Roger B. Grosse, and D. Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *NeurIPS*, 2018.

Xi Chen, Yan Duan, Rein Houthooft, J. Schulman, Ilya Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *NIPS*, 2016.

Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and L. Carin. Club: A contrastive log-ratio upper bound of mutual information. In *ICML*, 2020.

H. Chipman, E. George, and R. McCulloch. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4:266–298, 2010.

Rajeev H. Dehejia and S. Wahba. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics*, 84:151–161, 2002.

J. Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66:315–332, 1998.

N. Hassanpour and R. Greiner. Counterfactual regression with importance sampling weights. In *IJCAI*, 2019.

N. Hassanpour and R. Greiner. Learning disentangled representations for counterfactual regression. In *ICLR*, 2020.

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

R. Devon Hjelm, A. Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *ArXiv*, abs/1808.06670, 2019.

G. Imbens and D. Rubin. Causal inference for statistics, social, and biomedical sciences: An introduction. 2015.

Fredrik D. Johansson, U. Shalit, and D. Sontag. Learning representations for counterfactual inference. *ArXiv*, abs/1605.03661, 2016.

Hyunjik Kim and A. Mnih. Disentangling by factorising. In *ICML*, 2018.

Kun Kuang, Peng Cui, B. Li, Meng Jiang, Shiqiang Yang, and Fei Wang. Treatment effect estimation with data-driven variable decomposition. In *AAAI*, 2017.

R. Lalonde. Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 76:604–620, 1984.

Christos Louizos, U. Shalit, J. Mooij, D. Sontag, R. Zemel, and M. Welling. Causal effect inference with deep latent-variable models. *ArXiv*, abs/1705.08821, 2017.

Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, L. Hong, and Ed H. Chi. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018.

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

P. Rosenbaum and D. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

D. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.

U. Shalit, Fredrik D. Johansson, and D. Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *ICML*, 2017.

W. Sun, Pengyuan Wang, Dawei Yin, Jian Yang, and Yi Chang. Causal inference via sparse additive models with application to online advertising. In *AAAI*, 2015.

Stefan Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113:1228 – 1242, 2015.

Anpeng Wu, Kun Kuang, Junkun Yuan, B. Li, Pan Zhou, Jianrong Tao, Qiang Zhu, Yueting Zhuang, and Fei Wu. Learning decomposed representation for counterfactual inference. *ArXiv*, abs/2006.07040, 2020.

Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and A. Zhang. Representation learning for treatment effect estimation from observational data. In *NeurIPS*, 2018.

Jinsung Yoon, James Jordon, and M. Schaar. Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *ICLR*, 2018.

Weijia Zhang, Lin Liu, and Jiuyong Li. Treatment effect estimation with disentangled latent factors. In *AAAI*, 2021.

A APPENDIX

A.1 THEOREMS AND PROOF

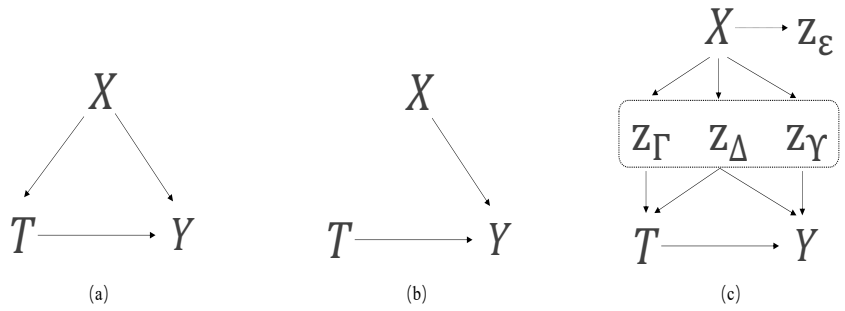


Figure 5: (a) refers to the *collider* in which X means confounding factors. (b) represents *do(t)* with removing all incoming edges of T. (c) represents our disentangled causal graph and z_ϵ are noise variables.

Theorem 1. *For two random variables $\boldsymbol{x}$ and $\boldsymbol{y}$,*

$$\mathrm{I}(\boldsymbol{x}, \boldsymbol{y}) \leq \mathrm{I}_{\mathrm{CLUB}}(\boldsymbol{x}, \boldsymbol{y}) \tag{20}$$

Equality is achieved if and only if $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent.

Proof. Similar to the proof in (Cheng et al., 2020), The Mutual Information is formulated as:

$$\mathrm{I}(\boldsymbol{x}, \boldsymbol{y}) = \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}\left[\log \frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})}\right] = \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}\left[\log p(\boldsymbol{y}|\boldsymbol{x}) - \log p(\boldsymbol{y})\right] \tag{21}$$

The Contrastive Log-ratio Upper Bound (CLUB) is defined as:

$$\mathrm{I}_{\mathrm{CLUB}}(\boldsymbol{x}, \boldsymbol{y}) = \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}\left[\log p(\boldsymbol{y}|\boldsymbol{x})\right] - \mathbb{E}_{p(\boldsymbol{x})}\mathbb{E}_{p(\boldsymbol{y})}\left[\log p(\boldsymbol{y}|\boldsymbol{x})\right] \tag{22}$$

We calculate the gap between $\mathrm{I}_{\mathrm{CLUB}}(\boldsymbol{x}, \boldsymbol{y})$ and $\mathrm{I}(\boldsymbol{x}, \boldsymbol{y})$:

$$\begin{aligned}
\Delta &= \mathrm{I}_{\mathrm{CLUB}}(\boldsymbol{x}, \boldsymbol{y}) - \mathrm{I}(\boldsymbol{x}, \boldsymbol{y}) \\
&= \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}\left[\log p(\boldsymbol{y}|\boldsymbol{x})\right] - \mathbb{E}_{p(\boldsymbol{x})}\mathbb{E}_{p(\boldsymbol{y})}\left[\log p(\boldsymbol{y}|\boldsymbol{x})\right] - \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}\left[\log p(\boldsymbol{y}|\boldsymbol{x}) - \log p(\boldsymbol{y})\right] \\
&= \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}\left[\log p(\boldsymbol{y})\right] - \mathbb{E}_{p(\boldsymbol{x})}\mathbb{E}_{p(\boldsymbol{y})}\left[\log p(\boldsymbol{y}|\boldsymbol{x})\right] \\
&= \mathbb{E}_{p(\boldsymbol{y})}\left[\log p(\boldsymbol{y}) - \mathbb{E}_{p(\boldsymbol{x})}\left[\log p(\boldsymbol{y}|\boldsymbol{x})\right]\right] \\
&= \mathbb{E}_{p(\boldsymbol{y})}\left[\log\left(\mathbb{E}_{p(\boldsymbol{x})}\left[p(\boldsymbol{y}|\boldsymbol{x})\right]\right) - \mathbb{E}_{p(\boldsymbol{x})}\left[\log p(\boldsymbol{y}|\boldsymbol{x})\right]\right]
\end{aligned} \tag{23}$$

Due to $\log(\cdot)$ is a concave function, by Jensen's Inequality, we have:

$$\log\left(\mathbb{E}_{p(\boldsymbol{x})}\left[p(\boldsymbol{y}|\boldsymbol{x})\right]\right) \geq \mathbb{E}_{p(\boldsymbol{x})}\left[\log p(\boldsymbol{y}|\boldsymbol{x})\right] \tag{24}$$

Therefore, the gap Δ is always non-negative, we have the following inequality:

$$\mathrm{I}(\boldsymbol{x}, \boldsymbol{y}) \leq \mathrm{I}_{\mathrm{CLUB}}(\boldsymbol{x}, \boldsymbol{y}) \tag{25}$$

We demonstrate that $\mathrm{I}_{\mathrm{CLUB}}(\boldsymbol{x}, \boldsymbol{y})$ is an upper bound of $\mathrm{I}(\boldsymbol{x}, \boldsymbol{y})$. The equality is satisfied when variables $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent.

Theorem 2. *The variational CLUB term $\mathrm{I}_{\mathrm{vCLUB}}(\boldsymbol{x}, \boldsymbol{y})$ remains a MI upper bound if the variational joint distribution $q_\theta(\boldsymbol{x}, \boldsymbol{y}) = q_\theta(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{x})$ satisfy the following inequality:*

$$\mathrm{KL}(p(\boldsymbol{x}, \boldsymbol{y})||q_\theta(\boldsymbol{x}, \boldsymbol{y})) \leq \mathrm{KL}(p(\boldsymbol{x})p(\boldsymbol{y})||q_\theta(\boldsymbol{x}, \boldsymbol{y})) \tag{26}$$

Then $\mathrm{I}(\boldsymbol{x}, \boldsymbol{y}) \leq \mathrm{I}_{\mathrm{vCLUB}}(\boldsymbol{x}, \boldsymbol{y})$, the equality holds when $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent.

Proof. Similar to Theorem 1, we calculate the gap between $\mathrm{I}_{\mathrm{vCLUB}}(\boldsymbol{x}, \boldsymbol{y})$ and $\mathrm{I}(\boldsymbol{x}, \boldsymbol{y})$:

$$
\begin{aligned}
\Delta &= \mathrm{I}_{\mathrm{vCLUB}}(\boldsymbol{x}, \boldsymbol{y}) - \mathrm{I}(\boldsymbol{x}, \boldsymbol{y}) \\
&= \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}\left[\log q_\theta(\boldsymbol{y}|\boldsymbol{x})\right] - \mathbb{E}_{p(\boldsymbol{x})}\mathbb{E}_{p(\boldsymbol{y})}\left[\log q_\theta(\boldsymbol{y}|\boldsymbol{x})\right] - \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}\left[\log p(\boldsymbol{y}|\boldsymbol{x}) - \log p(\boldsymbol{y})\right] \\
&= \mathbb{E}_{p(\boldsymbol{y})}\left[\log p(\boldsymbol{y})\right] - \mathbb{E}_{p(\boldsymbol{x})p(\boldsymbol{y})}\left[\log q_\theta(\boldsymbol{y}|\boldsymbol{x})\right] - \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}\left[\log p(\boldsymbol{y}|\boldsymbol{x}) - \log q_\theta(\boldsymbol{y}|\boldsymbol{x})\right] \\
&= \mathbb{E}_{p(\boldsymbol{x})p(\boldsymbol{y})}\left[\log \frac{p(\boldsymbol{y})}{q_\theta(\boldsymbol{y}|\boldsymbol{x})}\right] - \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}\left[\log \frac{p(\boldsymbol{y}|\boldsymbol{x})}{q_\theta(\boldsymbol{y}|\boldsymbol{x})}\right] \\
&= \mathbb{E}_{p(\boldsymbol{x})p(\boldsymbol{y})}\left[\log \frac{p(\boldsymbol{y})p(\boldsymbol{x})}{q_\theta(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})}\right] - \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}\left[\log \frac{p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})}{q_\theta(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})}\right] \\
&= \mathbb{E}_{p(\boldsymbol{x})p(\boldsymbol{y})}\left[\log \frac{p(\boldsymbol{x})p(\boldsymbol{y})}{q_\theta(\boldsymbol{x},\boldsymbol{y})}\right] - \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}\left[\log \frac{p(\boldsymbol{x},\boldsymbol{y})}{q_\theta(\boldsymbol{x},\boldsymbol{y})}\right] \\
&= \mathrm{KL}(p(\boldsymbol{x})p(\boldsymbol{y})||q_\theta(\boldsymbol{x},\boldsymbol{y})) - \mathrm{KL}(p(\boldsymbol{x},\boldsymbol{y})||q_\theta(\boldsymbol{x},\boldsymbol{y}))
\end{aligned}
\tag{27}
$$

Therefore, we conclude that $\mathrm{I}_{\mathrm{vCLUB}}(\boldsymbol{x}, \boldsymbol{y})$ is the upper bound of $\mathrm{I}(\boldsymbol{x}, \boldsymbol{y})$ if and only if when $\mathrm{KL}(p(\boldsymbol{x},\boldsymbol{y})||q_\theta(\boldsymbol{x},\boldsymbol{y})) \leq \mathrm{KL}(p(\boldsymbol{x})p(\boldsymbol{y})||q_\theta(\boldsymbol{x};\boldsymbol{y}))$, and the equality holds when $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent. More supplementary material about the CLUB estimator and its properties are available in (Cheng et al., 2020).

A.2 Hyperparameters settings

The details of the model architecture and optimal hyperparameter are as follows: For IHDP-A, we use 7 hidden layers with 30 neurons for each layer to encode the input variables into feature space[6], and 1 layer MLP with ReLU activation to decode the features into disentangled factors; the Y_0 and Y_1 prediction part are both 3 hidden layers with 100 neurons and LeakyReLU activation; the T prediction part is a 1 linear layer, the weight θ is set to 1. For IHDP-B and synthetic dataset, 1 layer with 100 neurons and ELU activation is used to encode features, the other parts are same to setting A while the weight $\theta = 0$, the difference between the θ is that θ can be seen as a symbol of balance between weight matrix and hidden units, the products of weight matrices may omit some information for IHDP-A as we use 7 layers compared whit only 1 layer for IHDP-B. Table 4 shows the details on our hyper-parameter search space of *MimCE*.

Table 4: Hyperparameter settings

Hyperparameter	Range
Encoder layer dim	$\{1, 2, 3, 4, 5, 6, 7\}$
Decoder layer dim	$\{1, 2, 3\}$
Encoder hidden units	$\{30, 50, 70, 100\}$
Y_0, Y_1, T layer dim	$\{1, 2, 3\}$
Y_0, Y_1, T hidden units	$\{50, 100\}$
$\alpha, \beta, \gamma, \eta, \lambda$	$\{0, 0.01, 0.1, 1, 10\}$
θ	$\{0, 1\}$

The configuration of the objective function weight and training process are as follows: we use the Adam optimizer with 1e-2 learning rate and 1e-2 weight decay weight, and the loss weight for $\mathcal{L}_{\text{treat}}$, $\mathcal{L}_{\text{disc}}$, $\mathcal{L}_{\text{CLUB}}$ and $\mathcal{L}_{\text{WREG}}$ are all set to 1 (i.e., we do not search these parameter as we argue that a good causal inference algorithm should preform well on simple weight settings).

In addition to this, we evaluate on the validation set every 10 epochs and we stop the running process when the loss does not drop for 10 rounds on the validation set and use the best model to predict the test set.

[6]For IHDP-A encoder layers, we don't use activation function.