ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240818

On the Effects of Irrelevant Variables in Treatment Effect Estimation with Deep Disentanglement

Ahmad Saeed Khan[a], Erik Schaffernicht[a] and Johannes Andreas Stork[a]

[a]Örebro University, Örebro, Sweden

Abstract. Estimating treatment effects from observational data is paramount in healthcare, education, and economics, but current deep disentanglement-based methods to address selection bias are insufficiently handling irrelevant variables. We demonstrate in experiments that this leads to prediction errors. We disentangle pre-treatment variables with a deep embedding method and explicitly identify and represent irrelevant variables, additionally to instrumental, confounding and adjustment latent factors. To this end, we introduce a reconstruction objective and create an embedding space for irrelevant variables using an attached autoencoder. Instead of relying on serendipitous suppression of irrelevant variables as in previous deep disentanglement approaches, we explicitly force irrelevant variables into this embedding space and employ orthogonalization to prevent irrelevant information from leaking into the latent space representations of the other factors. Our experiments with synthetic and real-world benchmark datasets show that we can better identify irrelevant variables and more precisely predict treatment effects than previous methods, while prediction quality degrades less when additional irrelevant variables are introduced.

1 Introduction

Treatment effect estimation from observational data is challenging because the uncontrolled mode of data collection can lead to selection bias. Selection bias causes a distributional difference between observed pre-treatment variables for different treatment groups, leading to biased counterfactual predictions. Managing this imbalance between treatment groups is therefore an important objective for improving treatment effect estimation [4, 15].

Deep disentanglement approaches [11, 3] use representation learning to identify the underlying factors as instrumental, confounding, or adjustment. This allows them to balance factors individually for improving treatment effect estimation [11]. However, this assumes that all pre-treatment variables are pre-screened for relevance, which is impractical in increasingly prevalent data-driven and big data settings.

Our empirical analysis shows that ignoring the presence of irrelevant variables in the data critically degrades predictions with a significant drop in the precision in estimation of heterogeneous effects (PEHE) for established benchmark datasets (see Fig. 1). Relying on serendipitous suppression of irrelevant variables as state-of-the-art deep disentanglement approaches do, is insufficient as it does not reliably prevent irrelevant information from leaking into other factors. Instead, it is necessary to actively disentangle irrelevant variables from other covariates used for prediction. This is supported by theoretical results emphasizing harmful consequences of unprincipled covariate inclusion [24].

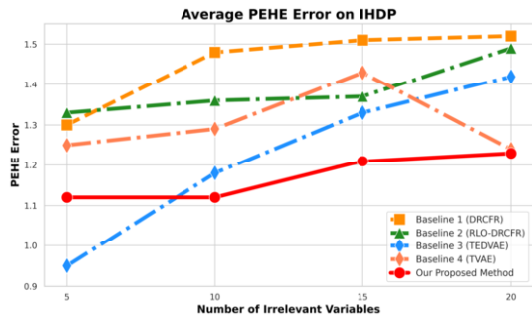

Figure 1. Average PEHE error on IHDP dataset against number of irrelevant variable dimensions (smaller the better). PEHE generally degrades with more irrelevant factors but our method is less affected.

In this paper, we address the issue of unidentified irrelevant pre-treatment variables with a novel deep disentanglement approach for estimating treatment effects which explicitly identifies and represents irrelevant factors, additionally to instrumental, confounding and adjustment factors. We achieve disentanglement of irrelevant factors by introducing an additional embedding space for irrelevant factors using covariate reconstruction and orthogonality objectives.

We empirically evaluate our approach and compare it to state-of-the-art deep disentanglement baselines using the infant health and development program (IHDP), jobs and a synthetic dataset with varying number of irrelevant variables. We find that our model is better than baselines at identifying and disentangling the latent factors, including irrelevant factors, according to perturbation importance analysis [1, 7, 32] and analysis of weights of the representation networks [33]. We also observe better performance on PEHE and policy risk evaluation criteria with increased number of irrelevant variables as compared to the baselines. Our approach is practicable and in principle compatible with previous deep disentanglement-based works as the additional channel and reconstruction objective leave the other representation networks and objectives unaltered.

Our core contributions are:

- Investigating the impact of irrelevant variables on estimation of treatment effects for state-of-the-art disentangled representational learning methods.
- The proposal of an autoencoder-based approach to disentangle irrelevant factors explicitly.
- A thorough evaluation of our approach, showing that it outperforms the baseline methods in estimating individual treatment effects, especially in the presence of irrelevant variables.

2 Related Work

Selection bias in observational data is a well known problem in treatment effect estimation, which is classically countered by balancing confounders with matching, stratification, and re-weighting approaches [26, 25, 22]. However, assuming unfoundedness or relying on prior knowledge of the causal structure is unsuitable in real-world, high-dimensional settings, leading to underperformance [19].

Deep representation learning has been proposed for balancing variables in higher-dimensional settings [15, 35, 28, 12]. The idea behind these approaches is to make the embedded data look like a Randomized Controlled Trial (RCT) by minimizing the discrepancy between treatment groups. Maximum mean discrepancy [8] and Wasserstein distance [30] are integral probability measures that are both used as discrepancy losses in these methods. We also employ the latter to balance our embedding spaces. The appealing simplicity of balancing all covariates in a shared single embedding space, however, overlooks the fact that usually not all covariates contribute to both, treatment and effect.

Disentanglement approaches, in contrast, account for the underlying causal structure by creating separate representations for instrumental, confounding and adjustment factors [11, 18, 20, 33, 3]. Disentangled representations give insights and can be used to reduce, as well as account for, the negative impact of selection bias [11]. A major challenge for these approaches is imperfect decomposition with pre-treatment variables information leaking into unrelated factors, which can degrade performance on the downstream prediction task. To ensure better separation, several techniques have been proposed such as different orthogonalization [18, 20, 33] and mutual information [3] objectives. While Kuang et al. [18, 20] only consider linear embeddings, Wu et al. [33] present a deep orthogonal regularizer for deep representation networks, which we also use in this work. Zhang et al. [36] propose to use variational autoencoders (VAE) for separating the three factors (TEDVAE). VAEs have been used previously in Louizos et al. [23] to estimate the confounding factors only, without the use of any discrepancy loss. Wu and Fukumizu [34] present a prognostic score-based VAE approach to estimate causal effects for data with limited overlap between treatment groups.

In real-world observational studies unidentified, irrelevant variables are inevitable and our empirical results show that irrelevant variables can degrade prediction results. Removing irrelevant variables has been studied in feature selection field [10] and many representation learning approaches are implicitly considered to remove irrelevant information. Yet there are indications [9, 17] that this is not sufficient, especially for tabular data. However, applying classic feature selection in a disentanglement task for treatment effect estimation is not straight forward due to proxy variables.

Some of the disentanglement-based works discussed before consider irrelevant variables: Hassanpour and Greiner [11] includes one single Gaussian noise factor in their synthetic dataset for evaluations; Wu et al. [33] claim that orthogonal regularization reduces the influence of irrelevant variables on the prediction; and Kuang et al. [18] mention that they eliminate irrelevant variables in their linear embedding with L_1 penalties. While the latter analyze separation of confounders and adjustment factors, they do not report on the identification of irrelevant factors. In contrast to our work, none of the approaches above explicitly represents irrelevant factors to achieve disentanglement. All of the approaches above rely on serendipitous suppression of irrelevant factors which is in some cases encouraged by regularization.

Targeted VAE (TVAE) is the most similar work to our approach.

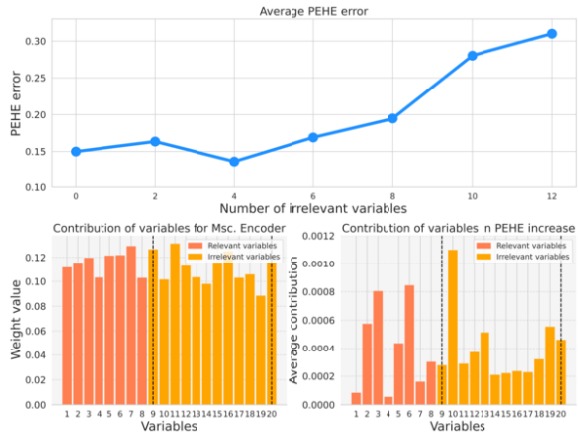

Figure 2. (Top): Illustrates the rise in PEHE as the number of irrelevant variables grows based on a baseline approach (TVAE). (Left): visualizes the individual contributions of variables towards learning the encoder for miscellaneous factors. Notably, the contribution of irrelevant variables mirrors that of relevant ones, underscoring the limitations of TVAE in disentangling irrelevant variables. (Right): shows average contribution of each variable in PEHE increase by using permutation of variables. Irrelevant variables are significantly participating in PEHE increase.

It employs an encoder to manage miscellaneous factors [31], but a crucial distinction arises. While TVAE aims to disentangle irrelevance when it's intertwined with relevant variables, it falls short in identifying and disentangling irrelevant variables existing in separate dimensions in pre-treatment variables. Our findings, depicted in Figure 2, reveal a notable increase in PEHE error as irrelevant variables increase based on TVAE. Furthermore, the encoder designed for miscellaneous factors demonstrates an inability to disentangle these variables within the data. Additionally, as depicted in Figure 2, it becomes evident that TVAE faces challenges in mitigating the influence of irrelevant variables, which contribute to the increase in PEHE. These observations are based on the same synthetic data utilized in the original study by Vowels et al. [31].

3 Formalization and Assumptions

In this section, we first give the notations and assumptions for treatment effect estimation in observational data. Moreover, we also define underlying latent factors of pre-treatment variables.

Formally, observational studies have a dataset: $\mathcal{D} = \{x_i, t_i, y_i\}_{i=1}^{N}$, where the i^{th} instance has some contextual information $x_i \in \mathcal{X} \subseteq \mathbb{R}^K$ (often called pre-treatment variables: e.g., gender and age), t_i is the observed treatment from the set of treatments $\mathcal{T}$ (e.g., 0: medication, 1: surgery) and $y_i \in \mathcal{Y}$ (e.g., recovery time; $\mathcal{Y} \subseteq \mathbb{R}^+$) is the respective outcome as the result of particular treatment t_i. In data $\mathcal{D}$, we only observe one outcome against the used treatment (known as factual outcome y_i^t) but alternative output (counterfactual outcome $y_i^{\neg t}$) is never observed in the data. In such datasets, $\mathcal{X}$ influences treatment assignment policy which causes selection bias in the data, where the condition $P(\mathcal{T}|\mathcal{X}) = P(\mathcal{T})$ does not hold and it lacks RCT properties[14, 11].

Hassanpour and Greiner [11] assume, without loss of generality, that $\mathcal{X}$ is generated by unknown joint distribution $P(\mathcal{X} \mid \Gamma, \Delta, \Upsilon)$, where Γ, Δ, Υ are latent factors; we are keeping the previously established notations for the these factors. Γ (instrumental factors) only influence treatment selection, Δ (confounding factors) affect both treatment selection and outcome, while Υ (adjustment factors) impact outcome only. We assume that there is another underlying irrel-

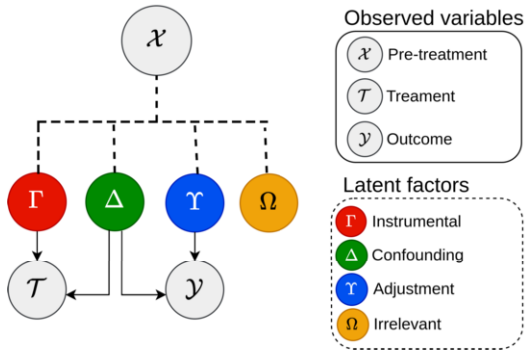

Figure 3. Underlying factors of $\mathcal{X}$. Observe that Ω has no associated downstream task with any observed variable.

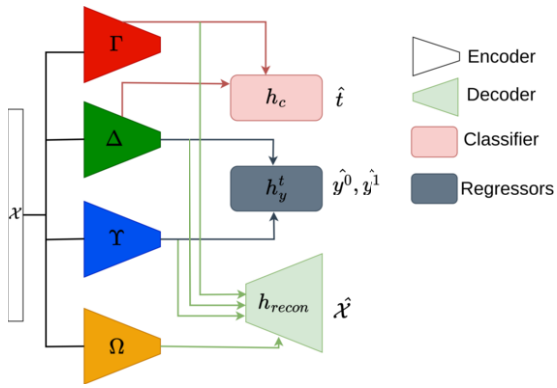

Figure 4. High level architecture of DRI-ITE.

evant latent factor (Ω) behind the generation of $\mathcal{X}$, depicted in Figure 3. Moreover, we also assume that latent factors are associated with separate dimensions of $\mathcal{X}$ as stated in Kuang et al. [18]. Learning the representation of Ω helps to match the true data generation process without harming the identifiability of causal effects [31].

The objective of this paper is to estimate Individual Treatment Effect (ITE) for each x_i: $ite_i = y_i^1 - y_i^0$, by learning a function $f: \mathcal{X} \times \mathcal{T} \rightarrow \mathcal{Y}$. However, it is not straightforward to learn such function f because $\mathcal{D}$ contains selection bias and irrelevant factors (Ω). It is essential to disentangle Ω from other latent factors to efficiently mitigate selection bias and to have reliable estimate of ITE by avoiding the overfitting of regression function f [18]. Empirically, we have observed a decline in the performance of recent disentangled representation learning methods for the ITE estimation with the increasing presence of Ω, as shown in the Figure 1.

Our work, like other methods in this domain, also relies on three assumptions as presented in Rubin [27].

1. **Stable Unit Treatment Value:** The treatment assignment to one unit does not affect the distribution of potential outcomes of the other unit.
2. **Unconfoundedness:** There is no unmeasured confounding. All confounding effect on $\mathcal{Y}$ and $\mathcal{T}$ has been measured, formally, $\mathcal{Y} \perp\!\!\!\perp \mathcal{T} \mid \mathcal{X}$.
3. **Overlap:** assumption states that the probability of assigning any treatment to x is higher than zero. Formally, $P(t \mid x) > 0 \, \forall t \in \mathcal{T}, \forall x \in \mathcal{X}$.

The assumptions of unconfoundedness and overlap are jointly known as strong ignorability.

4 Methods

Considering the likelihood of Ω being present in observational data, it becomes imperative to devise an approach that disentangles Ω and estimates ITE robustly.

Thereto, we propose Disentangled Representation with Irrelevant Factor for Individual Treatment Effect Estimation (DRI-ITE) for the binary treatment case, which learns disentangled representation with four latent factors (Γ, Δ, Υ, Ω), accounting for selection bias and simultaneously learns to predict counterfactual outcome for the final estimate of treatment effect. We achieve disentanglement of Ω by introducing an additional embedding space for irrelevant factors using $\mathcal{X}$ reconstruction and orthogonality objectives.

Figure 4 shows the DRI-ITE architecture, which contains four representational networks (encoders). Each network learns one specific latent factor. Two regression networks (one for each treatment group) learn to predict factual and counterfactual outcomes, and help two representational networks to disentangle Δ and Υ using $\mathcal{L}_{reg}$. One classification network learns to predict the treatment and helps in disentangling Γ and Δ using $\mathcal{L}_{class}$.

Finally, estimating Ω directly is difficult since there is no associated downstream task as shown in Fig.3. Instead, we employ a decoder that reconstructs $\mathcal{X}$. The core idea is that reconstructing the input data $\mathcal{X}$ in this autoencoder fashion requires to capture all latent factors including those not relevant to the ITE estimation. Intuitively, this allows us to use orthogonality objectives to separate the irrelevant factors into their own embedding space as $\Omega = \mathcal{X} \setminus \{\Gamma, \Delta, \Upsilon\}$.

From a computational point of view, DRI-ITE is moderately more expensive than comparable approaches due to the extra embedding space to learn the irrelevant factors.

The formal algorithm is provided in 1 and we will discuss details of each loss function in the following.

Algorithm 1 Disentangled Representation with Irrelevant Factor for Individual Treatment Effect Estimation (DRI-ITE)

Input: $\mathcal{D} = \{x_1, t_1, y_1\}, ..., \{x_N, t_N, y_N\}$
Output: $\hat{y}^1, \hat{y}^0$
Loss function: $\mathcal{L}_{main}$
Components: Four representation networks $\{\Gamma(.), \Delta(.), \Upsilon(.), \Omega(.)\}$, two regression networks $\{ h_y^0(.), h_y^1(.)\}$, one decoder $h_{recon}(.)$ and one classification network h_c

for i = 1 **to** N
$\{x_i, t_i, y_i\}_{i=1}^N \rightarrow \{\Gamma(x_i), \Delta(x_i), \Upsilon(x_i), \Omega(x_i)\}$
$h_c(\Gamma(x_i), \Delta(x_i)) \rightarrow \hat{t}_i$
$h_y^0(\Delta(x_i), \Upsilon(x_i)), h_y^1(\Delta(x_i), \Upsilon(x_i)) \rightarrow \hat{y}^1, \hat{y}^0$
$h_{recon}(\Gamma(x_i), \Delta(x_i), \Upsilon(x_i), \Omega(x_i)) \rightarrow x_i$
$w \leftarrow \text{Adam}\{\mathcal{L}_{main}\}$
end for
return $\hat{y}^1, \hat{y}^0$

The main objective function to be minimized is as follows:

$$\begin{aligned} \mathcal{L}_{main} = & \mathcal{L}_{reg} + \alpha \cdot \mathcal{L}_{class} + \beta \cdot \mathcal{L}_{disc} \\ & + \gamma \cdot \mathcal{L}_{recons} + \lambda \cdot \mathcal{L}_{orth} \\ & + \mu \cdot Reg(h_y^1, h_y^0, h_c, h_{recon}). \end{aligned} \tag{1}$$

Reg is a regularization term for the respective functions and $\alpha, \beta, \gamma, \lambda, \mu$ are weighting parameters.

We define $\mathcal{L}_{reg}$ as:

$$\mathcal{L}_{reg} = \mathcal{L}[y_i, h_y^{ti}(\Delta(x_i), \Upsilon(x_i))]. \tag{2}$$

$\mathcal{L}_{reg}$ is the regression loss (Mean squared error: MSE). We train two regression networks as used in Shalit et al. [28] and Hassanpour and Greiner [11] to predict observed outcome based on respective treatment. It is noteworthy that these regressors are learning on the concatenation of the Δ and Υ factors. Minimizing $\mathcal{L}_{reg}$ ensures that information regarding the outcome y is retained in these two latent factors and both representational networks learn its respective factors.

We define $\mathcal{L}_{class}$ as:

$$\mathcal{L}_{class} = \mathcal{L}[t_i, h_c(\Gamma(x_i), \Delta(x_i))]. \tag{3}$$

$\mathcal{L}_{class}$ is the classification loss (Binary cross-entropy: BCE). Classifier h_c learns to predict the treatment using the concatenation of Δ and Γ.

We define $\mathcal{L}_{disc}$ as follows:

$$\mathcal{L}_{disc} = disc[\Upsilon(x_i)t_{i=0}, \Upsilon(x_i)t_{i=1}]. \tag{4}$$

By minimizing $\mathcal{L}_{disc}$, we ensure that Υ contains no influence from Γ. In other words, $\mathcal{L}_{disc}$ helps to mitigate selection bias caused by Γ to have unbiased predictions for the downstream task. We use the Wasserstein distance as discrepancy loss as proposed by Cheng et al. [3].

The definition of $\mathcal{L}_{recons}$ is as follows:

$$\mathcal{L}_{recons} = \mathcal{L}[x_i, h_{recon}(\Gamma(x_i), \Delta(x_i), \Upsilon(x_i), \Omega(x_i))]. \tag{5}$$

$\mathcal{L}_{recons}$ is the reconstruction loss (MSE) used by the autoencoder to reconstruct $\mathcal{X}$ based on all four embedding spaces.

$\mathcal{L}_{orth}$ is deep orthogonal regularizer to ensure distinction among latent factors. Its idea is originally inspired by Kuang et al. [18]. We used the loss $\mathcal{L}_{orth}$ in the same way as it was used by Wu et al. [33]. However, instead of constraining orthogonality on pairs of average weight vectors of just three representational networks, we constrain orthogonality for the three more pairs to keep Ω separate from all other three basic factors. We define $\mathcal{L}_{orth}$ as follows:

$$\begin{aligned}\mathcal{L}_{orth} =& \bar{W}_\Gamma^T \cdot \bar{W}_\Delta + \bar{W}_\Delta^T \cdot \bar{W}_\Upsilon + \bar{W}_\Upsilon^T \cdot \bar{W}_\Gamma \\ &+ \bar{W}_\Omega^T \cdot \bar{W}_\Gamma + \bar{W}_\Omega^T \cdot \bar{W}_\Delta + \bar{W}_\Omega^T \cdot \bar{W}_\Upsilon,\end{aligned} \tag{6}$$

where $W \subseteq \mathbb{R}^{d \times d}$ is the product of weight matrices across all layers within a representational network, $\bar{W} \subseteq \mathbb{R}^{d \times 1}$ represents the row-wise average vector of the absolute values of W for each network. The vector $\bar{W}$ provides insight into the average contribution of each feature within that specific representational network. When $\mathcal{L}_{orth}$ is minimized, the dot products between the weight vectors become small or close to zero indicating orthogonality. Orthogonality between representations encourages each representational network to focus on capturing unique patterns and features relevant to its specific task. It prevents the networks from redundantly learning similar information. Alternatively, concepts from information theory i.e. total correlation or mutual information can also be utilized to separate information between the representational networks, but given the computational constraints of these methods it is common to employ deep orthogonal regularizers.

5 Experiments

Before discussing the results of the proposed method, we will briefly discuss the used datasets, evaluation criteria and experiment details. Our code is available [16].

5.1 Datasets

We use both synthetic and real-world datasets to evaluate the performance of the proposed method. A synthetic dataset allows to control all the latent factor that make up $\mathcal{X}$. To this purpose, we are augmenting the existing dataset proposed by Hassanpour and Greiner [11] with additional irrelevant variables. Additionally, performance on the commonly used IHDP and jobs dataset will be analyzed.

5.1.1 Synthetic Dataset

The dataset comprises a sample size of N, with dimensions $[m_\Gamma, m_\Delta, m_\Upsilon]$, along with mean and covariance matrices $(\mu_L, \sum_L)$ for each latent factor $L \in [\Gamma, \Delta, \Upsilon]$. A multivariate normal distribution is employed for data generation, and the covariates matrix is constructed as $N \times (m_\Gamma + m_\Delta + m_\Upsilon)$. The synthetic dataset is generated using the same settings and approach as presented by Hassanpour and Greiner [11]. To generate Ω, we follow the feature selection community by adding artificial contrasts. Each irrelevant variable is a permutation of a randomly selected feature generated for the other factors as simply using Gaussian or uniform distributions may not be sufficient [29].

5.1.2 Infant Health and Development Program (IHDP)

IHDP is a binary treatment dataset based on experiment conducted by Brooks-Gunn et al. [2]. Hill [13] introduced selection bias in original RCT data to make it an observational dataset. It contains 25 covariates that describe different aspects of the child and mother, such as birth weight, neonatal health index, mother's age, drug status, etc. The data has 747 instances in total, 139 belong to the treated group and 608 belong to the control group. The purpose of the study/data was to check the effect of treatment (specialist home visits) on the cognitive health of children. IHDP does not contain irrelevant variables, therefore we augment it with artificial contrasts for the evaluation purpose.

5.1.3 Jobs

Jobs is an observational dataset collected under the Lalonde experiment [21]. It contains eight pre-treatment covariates; age, educ, black, hisp, married, nodegr, re74, re75. The treatment data is binary and shows whether a person received job training or not. At the same time, the outcome variable indicates the earnings of a person in 1978. The data has 614 instances in total, 185 belong to the treated group, and 429 belong to the control group [13, 5]. We use artificial contrasts for the evaluation purpose.

5.2 Evaluation criteria

The well-established criterion for treatment effect estimation is Precision in Estimation of Heterogeneous Effect (PEHE), which is defined as follows:

$$PEHE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{e}_i - e_i)^2} \tag{7}$$

where $\hat{e}_i = \hat{y}_i^1 - \hat{y}_i^0$ and $e_i = y_i^1 - y_i^0$ are predict and true effects respectively.

Secondly, we use another well-known criterion, Policy Risk ($\mathcal{R}_{pol}$), which is defined as under:

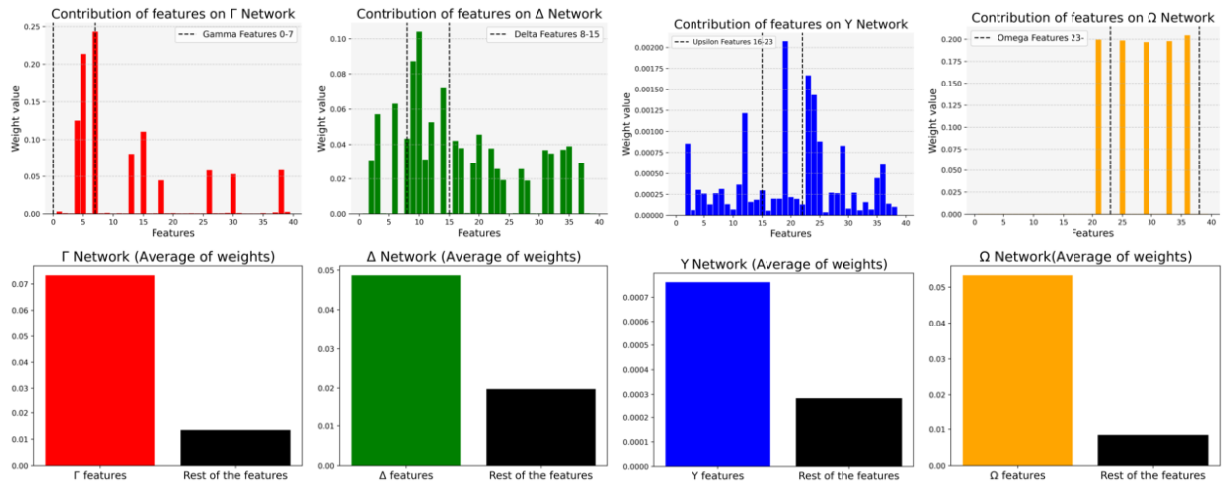

Figure 5. The visualization of feature contributions on each latent factor representational network is conducted for the dataset with dimensions 8, 8, 8, 15 (Γ,Δ,Υ,Ω) utilizing the $\bar{W}$ criterion based on DRI-ITE (ours). The top row visualizes all individual features, where high values are expected for the features between dotted lines, the bottom row represents the average over all features that are supposed to be represented by that particular network compared to the average weight of wrongly represented features.

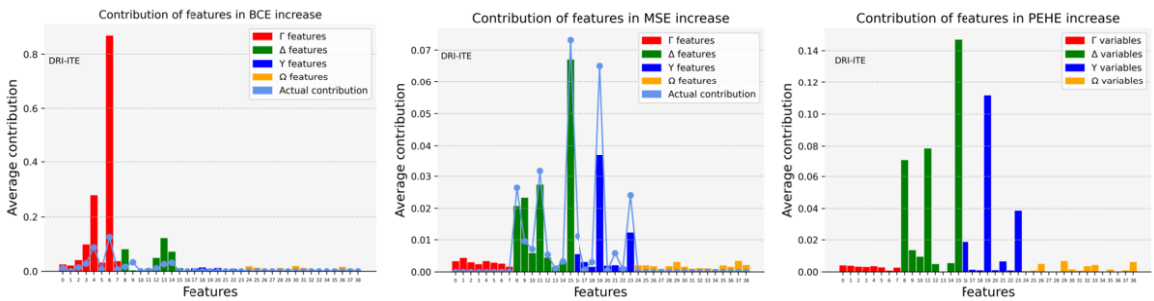

Figure 6. The average contribution of each feature in increasing BCE, MSE and PEHE loss is assessed by permuting features based on DRI-ITE (ours). Figure (a) shows the average contribution of each feature in increasing BCE (contribution of Γ and Δ features should have higher increases in BCE as compared to rest of the features, if Γ and Δ factors are identified correctly). Figure (b) shows the average contribution of each feature in increasing MSE (contribution of Δ and Υ features should have higher increases in MSE, if Δ and Υ factors are identified correctly). Figure(c) shows the average increase in PEHE (0.0017) by irrelevant variables using DRI-ITE (ours). A lower impact of irrelevant variables in increasing PEHE indicates accurate disentanglement of Ω and reliable ITE estimation.

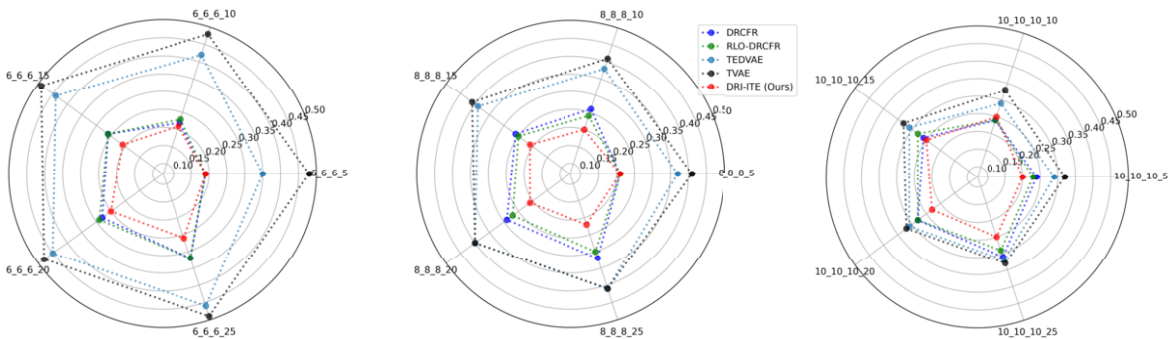

Figure 7. Radar charts visualizing the PEHE (mean values) results on the synthetic dataset. Each vertex represents the dimensions of latent factors (Γ_Δ_Υ_Ω). PEHE values closer to the center are better. Dashed red lines show our proposed method (DRI-ITE).

Table 1. Results of ablation study of loss function(bold numbers indicate smallest/best results).

Loss	8_8_8_5	8_8_8_10	8_8_8_15	8_8_8_20	8_8_8_25
$\mathcal{L}_{reg} + \mathcal{L}_{class} + \mathcal{L}_{disc}$	**0.21**(0.009)	0.25(0.01)	0.26(0.01)	0.28(0.01)	0.32(0.006)
$\mathcal{L}_{reg} + \mathcal{L}_{class} + \mathcal{L}_{disc} + \mathcal{L}_{orth}$	0.21(0.01)	0.25(0.02)	0.25(0.01)	0.28(0.01)	0.31(0.005)
$\mathcal{L}_{reg} + \mathcal{L}_{class} + \mathcal{L}_{disc} + \mathcal{L}_{orth} + \mathcal{L}_{recons}$	0.22(0.01)	**0.20**(0.01)	**0.20**(0.02)	**0.21**(0.01)	**0.21**(0.01)

$$\begin{aligned}\mathcal{R}_{pol}(\pi_f) =& 1 - (\mathbb{E}[Y^1 \mid \pi_f(x) = 1] \cdot p(\pi_f = 1) \\ &+ \mathbb{E}[Y^0 \mid \pi_f(x) = 0] \cdot p(\pi_f = 0))\end{aligned} \tag{8}$$

The policy risk is a measure of the average loss in value when following a specific treatment policy. The treatment policy ($\pi_f(x)$) is a set of rules based on the predictions of a model f. Specifically, if the difference in the model's predictions for treatment (1) and no treatment (0) is greater than a threshold (λ), then treat ($\pi_f(x) = 1$), otherwise do not treat ($\pi_f(x) = 0$). This formula involves the expected outcomes when following the treatment policy, weighted by the probabilities of applying the policy [28].

5.3 Experiment details

Table 2. Hyper-parameters and Ranges.

Hyper-parameter	Range
Latent dimensions	{5, 10, 15, 100}
Hidden dimensions	{50, 100, 200}
Layers	{2, 3, 4}
Batch size	{32, 64, 128, 256}
Learning rate	{$1e{-}2$, $1e{-}3$, $1e{-}4$, $1e{-}5$}
α, β, γ, λ, μ	{0.01, 0.1, 1, 5, 10, 100}

We employed three layers for the representational network of each latent factor (Γ, Δ, Υ, Ω). The hidden and output layer for Γ, Δ, Υ, Ω consisted of $10, 15$ neurons across multiple experiments. We utilized Adam as the optimizer, and ELU served as the activation function. The batch size was set to 256, the number of epochs to 5000 maximum, and the learning rate to $1e{-}5$. Following the approach outlined in [28], we employed $PEHE_{nn}$ on the validation set to save the best model. The data split between training and testing mirrored that used in [28, 11], with 20% of the training data reserved for the validation set. We used the same settings for jobs and synthetic datasets but only employed 100 dimensional representational networks to assign enough capacity for fair comparisons.

To select hyper-parameters, we employed grid search across different ranges (see Table 2). These parameter ranges were inspired by various baseline methods.

5.4 Results

We evaluate our method on two evaluation criteria: how accurately it identifies all disentangled latent factors (Subsection 5.4.1) and secondly how effectively it estimates the treatment effect using PEHE and policy risk ($\mathcal{R}_{pol}$) criterion (Subsection 5.4.2).

5.4.1 Identification of Disentangled Latent Factors

To quantify, how precisely our proposed method identifies all latent factors in the disentangled embedding spaces we use the following metrics:

- **Calculation of average weight vector:** We compute W, defined as the product of weight matrices across all layers within a representational network, and $\bar{W}$, which represents the row-wise average vector of the absolute values of W for each representational network. The vector $\bar{W}$ provides insight into the average contribution of each feature within that specific representational network. In the case of synthetic data, where the assignment of features to latent factors is known, we generated post-training plots of $\bar{W}$ for each network. The rationale behind this analysis [33] lies in the expectation that the average weights corresponding to features associated with a particular latent factor should exhibit higher values compared to other features.
- **Permutation feature importance analysis:** Secondly, we employed permutation feature importance theory [6] to validate the precise disentanglement of latent factors achieved by our representational networks. The underlying principle is straightforward: if shuffling a feature leads to an increase in model error after training, the feature is deemed important; otherwise, it can be considered unimportant. To the best of our knowledge, this is the first attempt to apply permutation feature importance theory to the domain of treatment effect estimation

In Figure 5, the top half shows the $\bar{W}$ bar plots for the Γ, Δ, Υ, and Ω representational networks using synthetic data. Notably, only relevant features exhibit high weights compared to the remaining ones. The figure confirms that each network accurately identifies its corresponding latent factors while effectively avoiding information leakage among them. The bottom half of Figure 5 presents the average weights of the respective features (between vertical lines) and the remaining features for each representational network. This visualization emphasizes that our approach selectively focuses on relevant information for each network, leading to accurate identification of latent factors.

Figure 6 illustrates the identification of latent factors based on the second criterion of permutation feature importance theory [6]. Specifically, Figure 6 (a) vividly demonstrates that only Γ and Δ features actively contribute to increasing BCE loss. This observation supports the conclusion that our method accurately identifies Γ and Δ factors from the data. Likewise, Figure 6 (b) confirms the successful identification of Δ and Υ factors indicated by the increased MSE.

The illustration in Figure 6 (c) shows again that DRI-ITE accurately disentangles and identifies Ω, as permuting irrelevant features does not increase the PEHE, while permuting any relevant feature does. We conjecture that the baseline methods fail to capture the feature importance completely. If true, this should result in overall lower ITE estimation errors.

Investigating the synthetic dataset, it is evident from Figure 7 that DRI-ITE consistently outperforms the baseline methods on PEHE evaluation. As the dimensions of Ω increase, baseline methods experience a much stronger decline in performance. DRI-ITE demonstrates better performance, particularly in scenarios with high-dimensional Ω.

In Table 1, we perform an ablation study to analyze the impact

Table 3. PEHE (mean (std)) on IHDP with different dimensions of Ω and varied latent dimensions of representational networks (bold numbers indicate smallest/best results).

	Latent dimensions=10					Latent dimensions=15				
Data_Ω	DR-CFR	RLO-DRCFR	TEDVAE	TVAE	DRI-ITE(Ours)	DR-CFR	RLO-DRCFR	TEDVAE	TVAE	DRI-ITE(Ours)
IHDP_5	1.30(0.78)	1.33(0.81)	**0.95(0.62)**	1.25(0.38)	1.12(0.62)	1.19(0.62)	1.26(0.71)	**0.93(0.62)**	1.28(0.56)	1.06(0.60)
IHDP_10	1.48(0.94)	1.36(0.76)	1.18(0.80)	1.29(0.43)	**1.12(0.65)**	1.25(0.73)	1.34(0.72)	**1.15(0.82)**	1.31(0.46)	1.20(0.66)
IHDP_15	1.51(0.98)	1.37(0.78)	1.33(0.83)	1.43(0.58)	**1.21(0.69)**	1.29(0.73)	1.36(0.77)	1.35(0.86)	1.29(0.51)	**1.23(0.65)**
IHDP_20	1.52(1.01)	1.49(0.91)	1.42(0.94)	**1.23(0.48)**	**1.23(0.65)**	1.30(0.74)	1.30(0.70)	1.41(0.89)	1.23(0.48)	**1.15(0.61)**

of the components $\mathcal{L}_{orth}$ and $\mathcal{L}_{recons}$ on the PEHE compared to the basic loss (i.e., $\mathcal{L}_{reg} + \mathcal{L}_{class} + \mathcal{L}_{disc}$). The results show that the addition of the orthogonal loss results in a minor improvement in performance, while the addition of the reconstruction loss leads to a significant decrease in the PEHE when ten or more irrelevant variables are introduced to the original variable set.

5.4.2 Evaluation on Estimation of Treatment Effect

Successfully disentangling latent factors including Ω in itself is not enough, but ultimately we aim to have improved estimates of the ITE. We assessed the performance of DRI-ITE using PEHE on the IHDP benchmark dataset; and using policy risk ($\mathcal{R}_{pol}$) criterion on Jobs dataset.

We are comparing our results with four SOTA baseline disentanglement approaches.

- Disentangled Representations for Counterfactual Regression: **DR-CFR** [11].
- Learning Disentangled Representations for Counterfactual Regression via Mutual Information Minimization: **RLO-DRCFR** [3].
- Treatment Effect with Disentangled Autoencoder: **TEDVAE** [36].
- Targeted VAE: Variational and Targeted Learning for Causal Inference: **TVAE** [31]

Table 4. Policy risk (mean (std)) on Jobs with different dimensions of Ω (bold numbers indicate smallest/best results).

Data_Ω	DR-CFR	RLO-DRCFR	TEDVAE	TVAE	DRI-ITE(Ours)
Jobs_5	0.13(0.03)	0.13(0.03)	0.20(0.03)	0.14(0.01)	**0.11(0.02)**
Jobs_15	**0.12**(0.03)	**0.12**(0.03)	0.21(0.04)	0.15(0.01)	**0.12(0.04)**
Jobs_20	0.14(0.04)	0.13(0.04)	0.19(0.03)	0.22(0.08)	**0.11(0.02)**

We evaluated DRI-ITE on IHDP. Table 3 presents PEHE values on the widely used IHDP benchmark dataset. The PEHE values (mean$_{(std)}$) are calculated from the first 30 realizations of IHDP, incorporating different dimensions of Ω and varying latent dimensions of representational networks.

Again, the performance of SOTA methods tends to degrade strongly with increasing dimensions of Ω. As depicted in Table 3, DRI-ITE effectively maintains a low PEHE in comparison to baseline methods after the introduction of Ω. Particularly noteworthy is the struggle of baseline methods, in scenarios with low-dimensional representational networks, which supposedly suppress Ω through regularization. This results in the assimilation of information from Ω into other relevant factors, consequently leading to poor performance. In contrast, our method adeptly disentangles Ω and consistently outperforms baseline methods.

However, as the representational networks increase in dimensionality, the performance of baseline methods also improves. We observed that for the baselines, regularization becomes more effective in suppressing Ω in high-dimensional scenarios compared to low dimensional networks. Despite this, our approach continues to provide better results. However, TEDVAE shows good performance against small number of Ω but it fails to ignore Ω with higher dimensions.

Table 4 presents a comparison between baseline methods and DRI-ITE regarding policy risk ($\mathcal{R}_{pol}$) criteria. These results are estimates (mean(std)) derived from the initial 30 realizations of the jobs dataset. Notably, the table illustrates that the performance of DRI-ITE remains consistently better and unaffected by the inclusion of Ω. In contrast, baseline methods experience a decline in performance as the dimensionality of Ω increases.

These results substantiate our assertion that SOTA methods lack an explicit and reliable mechanism to disentangle or ignore Ω. Conversely, our approach consistently disentangles Ω factors and reliably estimates ITE across all scenarios in comparison to SOTA methods. Moreover, these results are statistically significant based on the t-test with $\alpha = 0.05$.

6 Conclusion

In this paper, we address the problem of learning disentangled representation for Individual Treatment Effect (ITE) estimation with observational data. While deep disentanglement-based methods have been widely employed, they face limitations in handling irrelevant factors, leading to prediction errors. In the era of data-driven and big data approaches, where pre-screening for relevance is impractical, our work seeks to provide a robust solution to the inevitable presence of irrelevant factors in observational studies. We present a novel approach that goes beyond traditional deep disentanglement methods by explicitly identifying and representing irrelevant factors, in addition to instrumental, confounding, and adjustment factors. Our method leverages a deep embedding technique, introducing a reconstruction objective to create a dedicated embedding space for irrelevant factors through an autoencoder. Our empirical experiments, conducted on synthetic and real-world benchmark datasets, demonstrate the efficacy of our method. We showcase an improved ability to identify irrelevant factors and achieve more precise predictions of treatment effects compared to previous approaches. While our approach primarily addresses the scenario with two treatment groups, in future we plan to work with multiple and continuous treatments.

Acknowledgements

This work has been supported by the Industrial Graduate School Collaborative AI & Robotics funded by the Swedish Knowledge Foundation Dnr:20190128, and the Knut and Alice Wallenberg Foundation through Wallenberg AI, Autonomous Systems and Software Program (WASP).

References

[1] L. Breiman. Random forests. *Machine Learning*, 2001.

[2] J. Brooks-Gunn, F. ruey Liaw, and P. K. Klebanov. Effects of early intervention on cognitive function of low birth weight preterm infants. *The Journal of Pediatrics*, 1992.

[3] M. Cheng, X. Liao, Q. Liu, B. Ma, J. Xu, and B. Zheng. Learning disentangled representations for counterfactual regression via mutual information minimization. In *ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022.

[4] W. G. Cochran and D. B. Rubin. Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A*, 1973.

[5] R. H. Dehejia and S. Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 1999.

[6] A. Fisher, C. Rudin, and F. Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *The Journal of Machine Learning Research*, 2018.

[7] A. Fisher, C. Rudin, and F. Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *The Journal of Machine Learning Research*, 2019.

[8] A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, B. Schölkopf, et al. Covariate shift by kernel mean matching. *Dataset Shift in Machine Learning*, 2009.

[9] L. Grinsztajn, E. Oyallon, and G. Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems*, 2022.

[10] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 2003.

[11] N. Hassanpour and R. Greiner. Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*, 2019.

[12] N. Hassanpour and R. Greiner. Counterfactual regression with importance sampling weights. *International Joint Conference on Artificial Intelligence*, 2019.

[13] J. L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 2011.

[14] G. W. Imbens and D. B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

[15] F. D. Johansson, U. Shalit, and D. Sontag. Learning representations for counterfactual inference. In *International Conference on Machine Learning*, 2016.

[16] A. S. Khan, E. Schaffernicht, and J. A. Stork. Code for on the effects of irrelevant variables in treatment effect estimation with deep disentanglement. *Github*, 2024. URL https://github.com/askhanatgithub/DRI_ITE.

[17] B. Kim, H. Kim, K. Kim, S. Kim, and J. Kim. Learning not to learn: Training deep neural networks with biased data. In *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019.

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

[19] K. Kuang, P. Cui, B. Li, M. Jiang, Y. Wang, F. Wang, and S. Yang. Treatment effect estimation via differentiated confounder balancing and regression. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, page 1–25, 2019.

[20] 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*, 2020.

[21] R. J. LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 1986.

[22] S. Li, N. Vlassis, J. Kawale, and Y. Fu. Matching via dimen-sionality reduction for estimation of treatment effects in digital marketing campaigns. In *International Joint Conference on Artificial Intelligence*, 2016.

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

[24] J. Pearl. On a class of bias-amplifying covariates that endanger effect estimates. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence, 2010*, 2010.

[25] P. R. Rosenbaum. Model-based direct adjustment. *Journal of the American Statistical Association*, 1987.

[26] P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 1983.

[27] D. B. Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 2005.

[28] U. Shalit, F. D. Johansson, and D. Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, 2017.

[29] E. Tuv, A. Borisov, G. Runger, and K. Torkkola. Feature selection with ensembles, artificial variables, and redundancy elimination. *The Journal of Machine Learning Research*, 2009.

[30] C. Villani et al. *Optimal transport: old and new*. Springer, 2009.

[31] M. J. Vowels, N. C. Camgoz, and R. Bowden. Targeted vae: Variational and targeted learning for causal inference. In *2021 IEEE International Conference on Smart Data Services (SMDS)*, 2021.

[32] P. Wei, Z. Lu, and J. Song. Variable importance analysis: A comprehensive review. *Reliability Engineering & System Safety*, 2015.

[33] A. Wu, J. Yuan, K. Kuang, B. Li, R. Wu, Q. Zhu, Y. Zhuang, and F. Wu. Learning decomposed representations for treatment effect estimation. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

[34] P. Wu and K. Fukumizu. Beta intact vae: Identifying and estimating causal effects under limited overlap. *arXiv preprint arXiv:2110.05225*, 2021.

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

[36] W. Zhang, L. Liu, and J. Li. Treatment effect estimation with disentangled latent factors. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.