Multi-Cause Effect Estimation with Disentangled Confounder Representation

Jing Ma[1] , **Ruocheng Guo**[2] , **Aidong Zhang**[1] and **Jundong Li**[1*]
[1]University of Virginia, Charlottesville, VA, USA 22904
[2]Arizona State University, Tempe, AZ, USA 85287
{jm3mr, aidong, jundong}@virginia.edu, rguo12@asu.edu

Abstract

One fundamental problem in causality learning is to estimate the causal effects of one or multiple treatments (e.g., medicines in the prescription) on an important outcome (e.g., cure of a disease). One major challenge of causal effect estimation is the existence of unobserved confounders – the unobserved variables that affect both the treatments and the outcome. Recent studies have shown that by modeling how instances are assigned with different treatments together, the patterns of unobserved confounders can be captured through their learned latent representations. However, the interpretability of the representations in these works is limited. In this paper, we focus on the multi-cause effect estimation problem from a new perspective by learning disentangled representations of confounders. The disentangled representations not only facilitate the treatment effect estimation but also strengthen the understanding of causality learning process. Experimental results on both synthetic and real-world datasets show the superiority of our proposed framework from different aspects.

1 Introduction

One major challenge of causal effect estimation is the existence of *unobserved confounders* [Rosenbaum and Rubin, 1983], i.e., the unobserved variables which influence both treatment assignment and outcomes [Shalit *et al.*, 2017]. Unobserved confounders can cause confounding bias to the estimated treatment effect. A line of recent studies infers unobserved confounders by modeling how instances are assigned with different treatments together. The problem is known as the *multiple treatment effect* (MTE) estimation [Wang and Blei, 2019]. A typical example is to estimate how intervening the cast of movies would change their potential revenues, e.g., "how much does the revenue (outcome) increase or decrease if Oprah Winfrey is in the movie?". Here, the genre of the movie is an unobserved confounder that affects which actors would star the movie as well as the revenue; and different actors in the cast (multiple treatment assignment) can provide complementary insights in revealing the movie genre.

Specifically, recent studies of MTE estimation capture the unobserved confounders by learning their latent representations [Wang and Blei, 2019; Saini *et al.*, 2019] through the interactions between instances and treatments. However, the interpretability of the learned representations is limited, which could be an unknown mixture of several latent confounders' representations. In the movie cast example, the movie genre can be an unobserved confounder, but establishing the connection between this unobserved confounder and a particular part of the learned representations is difficult. In fact, it can greatly facilitate the human understanding of the confounding bias by separating the distinct, informative factors of variations in the confounders' representations [Locatello *et al.*, 2019]. Motivated by the recent progress of disentangled representation learning [Higgins *et al.*, 2016; Tran *et al.*, 2017] which learns factorized representations of the independent data generative factors, we investigate the MTE estimation problem from a new perspective by learning disentangled representations for confounders to improve the interpretability of causality learning.

However, learning disentangled representations of confounders for MTE estimation remains nascent due to the following challenges: (1) Different latent confounders are not only mixed together but also can exhibit hierarchical patterns (e.g., high-level latent confounders such as "the movie is an animation movie" and low-level latent confounders like "the animation movie is from Disney"), which further increases the complexity of disentangled representation learning. (2) When estimating the treatment effects, most existing works take different treatments separately [Lopez *et al.*, 2017] (i.e., constructing a prediction model for each treatment), which cannot capture the inherent dependencies between different treatments (e.g., two treatments could be similar if many instances are assigned with them simultaneously).

To address these challenges, we propose *DIRECT* – a novel framework of **D**isentangled mult**I**ple t**R**eatment **E**ffe**CT** estimation with the following desiderata: (1) To capture the hierarchical patterns of mixed confounders, we propose to disentangle the representations of latent confounders at two different levels. We first assume that treatments can be grouped into different clusters, as observed in many real-world scenarios [Schnabel, 2016]. Then by separately inferring con-

[*]Corresponding Author.

founders from the interactions between instances and each cluster of treatments (e.g., comedy actors or action actors), the learned confounder representations will become disentangled at the macro-level. Meanwhile, at the micro-level, we force different dimensions of the learned confounder representations to capture isolated factors with a carefully designed variational autoencoder (VAE) framework. (2) To tackle the issue that different treatments are often processed separately, we jointly consider multiple treatments simultaneously by leveraging their inherent dependencies. Specifically, we learn a trainable function to obtain the representation for each treatment based on treatment assignments. One appealing byproduct is that the framework can be generalized to new treatments that are not in the training data. Our main contributions include: 1) **Problem**: We formulate a new problem of disentangled multiple treatment effect estimation; 2) **Framework:** We propose a novel framework DIRECT to address this problem by learning disentangled confounder representations at two granularity levels; 3) **Experiments:** We conduct extensive experiments to show the superiority of DIRECT w.r.t. MTE estimation and interpretability.

2 Problem Definition

We use $\{\boldsymbol{A}, \boldsymbol{Y}\}$ to denote the observational data, where $\boldsymbol{A} = \{\boldsymbol{a}_i\}_{i=1}^n$ denotes the treatment assignment, and $\boldsymbol{a}_i = \{a_{i,j}\}_{j=1}^m$ refers to the assignment of m different treatments on the i-th instance. Without loss of generality, we focus on treatments with binary values, i.e., $a_{i,j} \in \{0, 1\}$. The *observed outcome* is denoted by $\boldsymbol{Y} = \{y_i\}_{i=1}^n$, and $y_i \in \mathbb{R}$. We build our framework upon the potential outcome framework [Vemuri, 2015]. We represent the *potential outcome* in the multiple treatment setting by $\boldsymbol{Y}_a = \{y_i(\boldsymbol{a})\}_{i=1}^n$, where $y_i(\boldsymbol{a})$ is the value of the outcome that would be observed if the i-th instance receives the treatment assignment $\boldsymbol{a} \in \{0, 1\}^m$. Then the ITE for the i-th instance over $\boldsymbol{a}$ is defined as $\tau_{i,\boldsymbol{a}} = y_i(\boldsymbol{a}) - y_i(\boldsymbol{0})$, where $y_i(\boldsymbol{0})$ refers to the potential outcome when no treatment is assigned to the i-th instance.

Definition 1. *(Disentangled multiple treatment effect estimation) Given observational data $\{\boldsymbol{A}, \boldsymbol{Y}\}$, our goal is to: (1) learn disentangled representations for the latent confounders. (2) estimate the ITE $\tau_{i,\boldsymbol{a}}$ for each instance i under any treatment assignment $\boldsymbol{a}$.*

In our work, we relax the *strong ignorability* assumption [Rosenbaum and Rubin, 1983] (i.e., no unobserved confounders) by assuming that there may exist confounders $\boldsymbol{Z} = \{\boldsymbol{z}_1, ..., \boldsymbol{z}_n\}$ which (might be unobserved) can causally influence $\boldsymbol{A}$ and $\boldsymbol{Y}$. Conditioning on the confounders, the treatment assignment is randomized, i.e., $\mathrm{Y}_i(\boldsymbol{a}) \perp\!\!\!\perp \mathrm{A}_i|\mathrm{Z}_i$ [1] for any treatment assignment $\boldsymbol{a} \in \{0, 1\}^m$. Following [Wang and Blei, 2019], we also assume that for each instance i, the assignment of different treatments is independent with each other conditioned on the confounders, i.e., $\mathrm{A}_{i,1} \perp\!\!\!\perp ... \perp\!\!\!\perp \mathrm{A}_{i,m}|\mathrm{Z}_i$. Other assumptions in this work include the Positivity, Consistency, and SUTVA assumptions [Rubin, 2005], which are widely-adopted in causal inference.

3 The Proposed Framework

The causal graph of the studied problem is shown in Fig. 1, where Z_i denotes the confounders which influence the outcome and the assignment of at least one treatment for each instance i. Our framework DIRECT learns disentangled representations of confounders $\boldsymbol{Z}$ with a carefully designed variational autoencoder (VAE) architecture. At the macro-level, we learn the embedded cluster structure of different treatments to better learn and interpret the confounder representations. At the micro-level, we force each dimension of the learned representation to capture an isolated factor. Furthermore, by leveraging the dependencies among treatments, we learn a parameterized function to obtain the representation of each treatment. In this way, the model can be generalized to estimate the effects of treatment assignment including new treatments without retraining from scratch.

3.1 Model Description

Our framework includes three sets of latent variables: $\boldsymbol{Z}, \boldsymbol{T}$, and $\boldsymbol{C}$, where $\boldsymbol{T} = \{\boldsymbol{t}_j\}_{j=1}^m$ is the representation of treatments, $\boldsymbol{C} = \{\boldsymbol{c}_j\}_{j=1}^m$ is the cluster assignment of the treatments, each $\boldsymbol{c}_j$ is an one-hot vector and $c_{j,k} = 1$ denotes that treatment j belongs to cluster k. As shown in Fig. 1, the distribution $p(\mathrm{A}, \mathrm{C}, \mathrm{T}, \mathrm{Z})$ can be factorized as:

$$p(\Theta) = \prod\nolimits_{i,j} p(\mathrm{A}_{i,j}|\mathrm{Z}_i, \mathrm{T}_j, \mathrm{C}_j)p(\mathrm{C}_j)p(\mathrm{T}_j|\mathrm{C}_j)p(\mathrm{Z}_i), \quad (1)$$

where $\Theta = \{\mathrm{A}, \mathrm{C}, \mathrm{T}, \mathrm{Z}\}$. Assume that the treatments can be divided into K clusters, we then divide the treatment assignments into K groups corresponding to the treatment clusters. The confounders are learned separately for the treatment assignments in each group for a macro-level disentanglement. The confounder representation learned from the assignment of treatments in cluster k is denoted by $\boldsymbol{Z}^{(k)} = \{\boldsymbol{z}_1^{(k)}, \boldsymbol{z}_2^{(k)}, ..., \boldsymbol{z}_n^{(k)}\}$. Each $\mathrm{Z}_i^{(k)}$ is assumed to follow an isotropic unit Gaussian prior: $\mathrm{Z}_i^{(k)} \sim \mathcal{N}(0, \boldsymbol{I})$. Thus the distribution in Eq. (1) can be further factorized as:

$$p(\Theta) = \prod_{i,j} p(\mathrm{A}_{i,j}|\mathrm{Z}_i, \mathrm{T}_j, \mathrm{C}_j)p(\mathrm{C}_j)p(\mathrm{T}_j|\mathrm{C}_j) \prod_{k=1}^{K} p(\mathrm{Z}_i^{(k)}). \quad (2)$$

An illustration of the proposed framework DIRECT is shown in Fig. 2, which follows a classical VAE architecture with the inference network and the generation network. The inference network infers the variational distributions of treatments and confounders based on the treatment assignment. The generation network reconstructs the input (i.e., treatment assignment), and predicts the potential outcome of each instance.

3.2 Inference Network

Since the true posterior distribution of the latent variables $q(\mathrm{Z}, \mathrm{T}, \mathrm{C}|\boldsymbol{A})$ is intractable, we use an inference network to

[1] We use non-italicized capital letters to denote random variables, and italicized letters to denote specific realization. Among the italicized letters, non-bold letters denote scalars, bold lowercase letters denote vectors, and bold uppercase letters denote matrices or sets. For example, Z_i is a randomly chosen vector of confounders, $\boldsymbol{Z}$ is the set which contains confounders of all instances, and $\boldsymbol{z}_i$ denotes the values of confounders for instance i.

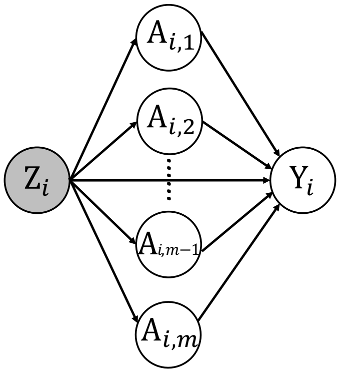


Figure 1: Causal graph of the studied problem.

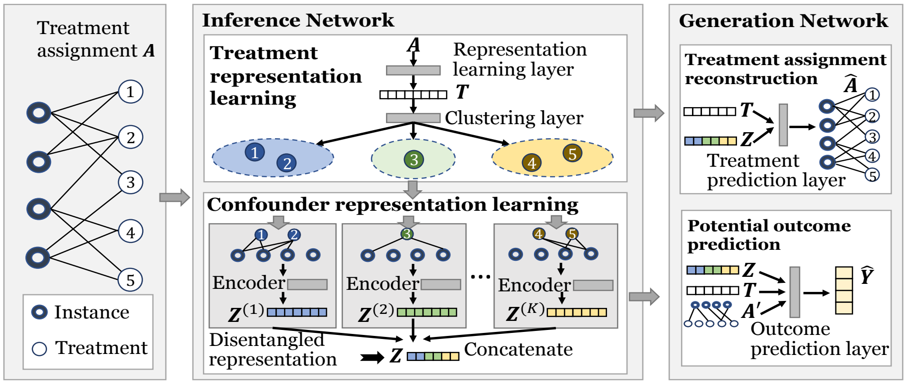


Figure 2: An illustration of the proposed framework DIRECT. It consists of two essential components: inference network and generation network.

approximate it based on the mean-field approximation. The approximate posterior $q(\mathrm{Z}, \mathrm{T}, \mathrm{C}|\boldsymbol{A})$ can be factorized as:

$$q(\mathrm{Z}, \mathrm{T}, \mathrm{C}|\boldsymbol{A})=\prod_{j=1}^{m} q(\mathrm{T}_j|\boldsymbol{a}_{*,j})q(\mathrm{C}_j|\mathrm{T}_j)\prod_{i=1}^{n}\prod_{k=1}^{K} q(\mathrm{Z}_i^{(k)}|\boldsymbol{a}_{i,*}). \tag{3}$$

We learn the representations of treatments and confounders through the variational inference mentioned above, where $*$ represents all the indices, e.g., $\boldsymbol{a}_{i,*}$ refers to $\boldsymbol{a}_{i,1}, ..., \boldsymbol{a}_{i,m}$.

Treatment Representation Learning

The proposed framework explicitly learns the treatments' representations from the observed treatment assignments. Specifically, the inference network specifies the form of the variational distributions of T_j to be Gaussian posteriors: $q(\mathrm{T}_j|\boldsymbol{a}_{*,j}) = \mathcal{N}(\mu_T(\boldsymbol{a}_{*,j}), \mathrm{diag}(\sigma_T^2(\boldsymbol{a}_{*,j})))$. In the inference network, the mean and variance of the posterior are inferred by two separate neural network modules $\mu_T(\cdot)$ and $\sigma_T(\cdot)$. We assume that the representations of treatments in the observational data have an inherent cluster structure composed with K components, where K is a hyperparameter. A clustering module is introduced in the inference network to approximate the cluster distribution $q(\mathrm{C}_j|\boldsymbol{t}_j) = Mult(f_c(\boldsymbol{t}_j))$, where $f_c(\cdot)$ is a function of the clustering module. The output of $f_c(\cdot)$ is a K-dimensional vector, where each element inside corresponds to the probability that the treatment belongs to each cluster. And the multinoulli distribution is implemented by a softmax layer. To enable clustering, the inference network specifies the prior of T_j to be $\mathcal{N}(\boldsymbol{\mu}_{c_j}, \mathrm{diag}(\boldsymbol{\sigma}_{c_j}^2))$, where $\boldsymbol{\mu}_{c_j}$ and $\boldsymbol{\sigma}_{c_j}$ are parameters to be learned, referring to the mean and variance of the distribution of treatments in the cluster containing the j-th treatment. As the latent variable C_j is discrete, the reparameterization trick based sampling is not differentiable for backpropagation, thus we apply Gumbel-Softmax sampling [Jang *et al.*, 2016] to approximate samples from the categorical distribution.

The treatment representation learning module enables the model to handle unseen treatments. For a new treatment $m+1$, we can obtain its representation $\boldsymbol{t}_{m+1}$ and predict its cluster based on its treatment assignments and the trained model. Then, the potential outcome of those treatment assignments which involve the new treatment can be predicted with $\boldsymbol{t}_{m+1}$ and the learned confounder representations.

Disentangled Confounder Representation Learning

Inspired by [Ma *et al.*, 2019], we learn disentangled representations of the confounders at two different levels. At the macro-level, the treatment assignment is divided into K groups according to the sampled $\boldsymbol{C}$. We learn the representation $\boldsymbol{Z}^{(k)}$ of confounders from each group k separately, and the final representation $\boldsymbol{Z}$ is the concatenation of $\boldsymbol{Z}^{(1)}, ..., \boldsymbol{Z}^{(K)}$. Specifically, the learned $\boldsymbol{Z}^k$ is expected to correspond to the treatments in cluster k. For each $\boldsymbol{Z}^k$, we infer the posteriors distributions as: $q(\mathrm{Z}_i^{(k)}|\boldsymbol{a}_{i,*}) = q(\mathrm{Z}_i^{(k)}|\boldsymbol{a}_{i,*}^{(k)}) = \mathcal{N}(\mu_I(\boldsymbol{a}_{i,*}^{(k)}), \mathrm{diag}(\sigma_I^2(\boldsymbol{a}_{i,*}^{(k)})))$, where $\boldsymbol{a}_{i,*}^{(k)}$ is the i-th instance's treatment assignment which only contains the treatments in cluster k. $\mu_I(\cdot)$ and $\sigma_I(\cdot)$ are two neural network modules to infer the mean and variance of the distribution of $q(\mathrm{Z}_i^{(k)}|\boldsymbol{a}_{i,*}^{(k)})$, respectively. At the micro-level, to achieve disentanglement among dimensions of learned representations, we specify a weight $\beta \gg 1$ for the Kullback–Leibler (KL) divergence between the isotropic unit Gaussian prior and the learned distribution of each $\mathrm{Z}^{(k)}$ to encourage the dimensions to reflect isolated latent factors.

3.3 Generation Network

In the generation model, we reconstruct the treatment assignment with a neural network module f_a: $p(\mathrm{A}_{i,j}|\boldsymbol{z}_i, \boldsymbol{t}_j, \boldsymbol{c}_j) = Ber(\mathrm{sigmoid}(f_a(\boldsymbol{c}_j, \boldsymbol{z}_i, \boldsymbol{t}_j)))$. We use a sigmoid layer to map the output of network $f_a(\cdot)$ into $(0, 1)$ as the probability of taking the treatment $a_{i,j}$. In order to better capture the latent confounders, we also use the observed outcomes as a supervision signal. Specifically, we use a neural network module $f_y(\mathrm{Z}_i, \mathrm{A}_i, \mathrm{T})$ to predict the potential outcome $y_i(\boldsymbol{a})$ for any treatment assignment $\boldsymbol{a}$. We assume the prediction $\hat{\mathrm{Y}}_i(\boldsymbol{a})$ follows the Gaussian distribution $\mathcal{N}(y_i(\boldsymbol{a}), \sigma_e^2)$, where σ_e^2 is the variance of the prediction error. We use the observed outcome y_i as target and minimize the outcome prediction loss: $\mathcal{L}_y = -\sum_{i=1}^{n} \log p(\hat{\mathrm{Y}}_i = y_i|\boldsymbol{z}_i, \boldsymbol{a}_i, \boldsymbol{T})$.

3.4 Optimization

Following the classical VAE schema, the evidence lower bound (ELBO) $\mathcal{L}_{ELBO}$ can be derived as (the subscript q

Dataset	Synthetic	Amazon-3C	Amazon-6C
# of instances	$2,500$	$3,000$	$6,000$
# of treatments	500	104	325
# of clusters	4	3	6
Avg ratio of treated	42.3%	21.4%	18.6%

Table 1: Detailed statistics of the datasets.

denotes $q(\mathrm{Z}, \mathrm{T}, \mathrm{C}|\boldsymbol{A})$ by default, and we drop the instance index i and treatment index j for notation simplicity):

$$\begin{aligned}&\mathbb{E}_q[\log p(\boldsymbol{A}|\mathrm{Z}, \mathrm{T}, \mathrm{C})] - KL(q(\mathrm{Z}, \mathrm{T}, \mathrm{C}|\boldsymbol{A})||p(\mathrm{Z}, \mathrm{T}, \mathrm{C})) \\ =\ &\mathbb{E}_q[\log p(\boldsymbol{A}|\mathrm{Z}, \mathrm{T}, \mathrm{C})] - \mathbb{E}_{q(\mathrm{T}|\boldsymbol{A})} KL(q(\mathrm{C}|\mathrm{T})||p(\mathrm{C})) \\ -\ &\mathbb{E}_{q(\mathrm{C}|\mathrm{T})} KL(q(\mathrm{T}|\boldsymbol{A})||p(\mathrm{T}|\mathrm{C})) \\ -\ &\sum_{k=1}^{K} \mathbb{E}_{q(\mathrm{T}|\boldsymbol{A})q(\mathrm{C}|\mathrm{T})} KL(q(\mathrm{Z}^{(k)}|\boldsymbol{A})||p(\mathrm{Z}^{(k)})). \qquad (4)\end{aligned}$$

The ELBO consists of the reconstruction term and the KL term, which contains three terms: 1) the clustering prior term, where we use the uniform prior; 2) the treatment prior term, which drives the treatment clustering as described in the section of treatment representation learning; 3) the confounder prior term, which leads to disentanglement among dimensions by utilizing the isotropic nature of the prior. It is impractical to calculate the expectations over the variational distribution analytically, thus these terms are instead estimated by Monte Carlo samples from $q(\mathrm{Z}, \mathrm{T}, \mathrm{C}|\boldsymbol{A})$. By putting all the aforementioned components together, we obtain the loss function of the proposed framework:

$$\begin{aligned}\mathcal{L} = &- \mathbb{E}_q[\log p(\boldsymbol{A}|\mathrm{Z}, \mathrm{T}, \mathrm{C})] + \mathbb{E}_{q(\mathrm{T}|\boldsymbol{A})} KL(q(\mathrm{C}|\mathrm{T})||p(\mathrm{C})) \\ &+ \mathbb{E}_{q(\mathrm{C}|\mathrm{T})} KL(q(\mathrm{T}|\boldsymbol{A})||p(\mathrm{T}|\mathrm{C})) + \lambda \mathcal{L}_y \\ &+ \beta \sum_{k=1}^{K} \mathbb{E}_{q(\mathrm{T}|\boldsymbol{A})q(\mathrm{C}|\mathrm{T})} KL(q(\mathrm{Z}^{(k)}|\boldsymbol{A})||p(\mathrm{Z}^{(k)})). \qquad (5)\end{aligned}$$

Hyperparameters β and λ are used to control the effect of different parts of the objective function.

4 Experiments

4.1 Datasets

We evaluate the proposed method on one synthetic dataset and two semi-sythetic datasets from real-world scenarios. The detailed statistics of these datasets are shown in Table 1, including the number of instances, treatments, treatment clusters, and the average ratio of treatments assigned to instances.

Synthetic Dataset. We first conduct experiments on a synthetic dataset. This dataset is generated as follows:

$$\mathrm{Z}_i^{(k)} \sim \mathcal{N}(0, \boldsymbol{I}),\ \mathrm{C}_j \sim Mult(\boldsymbol{\pi}),$$
$$\mathrm{T}_j|\boldsymbol{c}_j \sim \prod\nolimits_{k=1}^{K} \mathcal{N}(\boldsymbol{\mu}_k, \mathrm{diag}(\boldsymbol{\sigma}_k^2))^{c_{j,k}},\ \mu_k \sim \mathcal{N}(0, \boldsymbol{I}),$$
$$\sigma_k \sim \mathrm{rand}(0, \boldsymbol{I}), \mathrm{A}_{i,j} \sim Ber(\mathrm{sigmoid}(\boldsymbol{z}_i^{(c_j)} \boldsymbol{t}_j)), \qquad (6)$$

where $i = 1, ..., n$ and $j = 1, ..., m$, $\boldsymbol{\pi}$ is a K-dimensional vector, corresponding to the probability that the treatment belongs to each cluster. We set d_I (the dimension of each $\boldsymbol{z}_i^{(k)}$) and d_T (the dimension of T_j) both as 20. The potential outcome of instance i under a treatment assignment $\boldsymbol{a}$ is simulated as $y_i(\boldsymbol{a}) = \boldsymbol{a}^T \boldsymbol{T} \boldsymbol{W}_1 \boldsymbol{z}_i$, where $\boldsymbol{z}_i$ is the concatenation of $\boldsymbol{z}_i^{(k)}$ $(k = 1, ..., K)$, and $\boldsymbol{T} = [\boldsymbol{t}_1, ..., \boldsymbol{t}_m]^T$. $\boldsymbol{W}_1$ is a matrix of parameters with dimensions $d_T \times K d_I$.

Real-world Datasets. It is notoriously hard to obtain the ground truth treatment effect as we only observe one of the potential outcomes for each instance. Thus, we create two semi-synthetic datasets (Amazon-3C and Amazon-6C) based on the real-world Amazon review data[2]. In each dataset, we select three/six categories of items. In each category, we select the top-1000 products with the highest number of reviews as instances. We aim to investigate the effect of the keywords in reviews on the future sales of each product: (1) *Treatment*: We first generate a dictionary of keywords by performing unsupervised feature selection [Li *et al.*, 2017] on the bag-of-words features of reviews, then randomly select three words from the dictionary as a treatment, (e.g., if an item receives reviews containing all the three words in a treatment, then we say the treatment is assigned to the item). (2) *Potential outcome*: The future amount of sales of each product is the outcome and is simulated in the same way as that for the synthetic data. (3) *Condounders*: The confounders are the latent attributes of the products, which affect what words would appear in the reviews, as well as the product sales. We simulate the confounders by training a neural network to fit the treatment assignment, and take the output of a middle layer as confounders, then use it to simulate the potential outcome.

4.2 Experiment Settings

To evaluate our proposed framework in MTE estimation, we compare it with several state-of-the-art baselines in the following three categories: (1) traditional regression methods: least square regression (**OLS/LR**) and random forest (**RF**). These methods can take treatment assignment as features and predict the outcomes; (2) representation learning based ITE esitmation methods for single treatment: Causal Effect Variational Autoencoder (**CEVAE**) [Louizos *et al.*, 2017], Treatment-Agnostic Representation Network (**TARNet**) [Shalit *et al.*, 2017], and counterfactual regression with Wasserstein metric (**CFR**) [Shalit *et al.*, 2017]. (3) Multiple treatment effect estimation methods: Bayesian Additive Regression Trees (**BART**) [Hill, 2011] – though widely used in single-cause ITE estimation, can be naturally extended to multi-treatment setting by extending the input vectors in the Bayesian regression tree. Multi-cause deconfounder [Wang and Blei, 2019] utilizes the dependencies among the assigned causes to capture the confounders. We apply two different forms (linear and quadratic) in the potential outcome prediction, denoted as **Deconf-l** and **Deconf-q** respectively. As an ablation study of our proposed method, we disable the disentanglement by setting $K = 1$ and $\beta = 1.0$, maintaining the same dimension of representation. This variant of our method is denoted by **DIRECT-ND**.

Setup. Each dataset is randomly split into $60\%/20\%/20\%$ training/validation/test set. Unless otherwise specified, hy-

[2] http://jmcauley.ucsd.edu/data/amazon/index_2014.html

Method	Synthetic		Amazon-3C		Amazon-6C	
	$\overline{PEHE}$	$\overline{\epsilon_{ATE}}$	$\overline{PEHE}$	$\overline{\epsilon_{ATE}}$	$\overline{PEHE}$	$\overline{\epsilon_{ATE}}$
OLS/LR	10.07 ± 1.28	5.31 ± 0.88	11.15 ± 1.93	5.21 ± 1.13	11.27 ± 1.48	6.51 ± 0.52
RF	10.26 ± 1.23	5.22 ± 0.64	10.25 ± 1.81	5.17 ± 0.75	10.53 ± 1.62	5.82 ± 0.41
CEVAE	16.58 ± 1.99	7.38 ± 0.51	19.11 ± 1.21	9.93 ± 0.69	17.52 ± 0.92	8.08 ± 0.26
TARNET	12.07 ± 1.30	5.77 ± 0.84	9.27 ± 1.26	5.12 ± 1.24	9.51 ± 0.34	4.31 ± 0.20
CFR	12.78 ± 1.49	6.03 ± 0.94	8.37 ± 0.43	3.92 ± 1.04	9.32 ± 0.92	4.29 ± 0.19
BART	10.92 ± 2.20	5.30 ± 1.05	9.91 ± 1.77	6.03 ± 1.49	11.02 ± 2.15	5.11 ± 1.41
Deconf-l	8.26 ± 1.37	3.16 ± 0.24	7.34 ± 0.48	3.86 ± 0.41	8.16 ± 0.65	4.53 ± 0.53
Deconf-q	8.54 ± 1.28	3.42 ± 0.33	7.18 ± 0.52	3.21 ± 0.30	8.68 ± 0.72	4.25 ± 0.28
DIRECT-ND (ours)	4.91 ± 0.36	2.26 ± 0.08	5.89 ± 0.34	2.85 ± 0.16	6.37 ± 0.13	3.38 ± 0.12
DIRECT (ours)	$\mathbf{3.42 \pm 0.12}$	$\mathbf{1.33 \pm 0.08}$	$\mathbf{4.57 \pm 0.31}$	$\mathbf{2.04 \pm 0.14}$	$\mathbf{5.04 \pm 0.09}$	$\mathbf{2.37 \pm 0.08}$

Table 2: Performance of multiple treatment effect estimation for different methods.

Treatment	Hold out 20%		Together	
	$\overline{PEHE}$	$\overline{\epsilon_{ATE}}$	$\overline{PEHE}$	$\overline{\epsilon_{ATE}}$
Synthetic	3.21 ± 0.07	1.26 ± 0.05	$\mathbf{3.04 \pm 0.08}$	$\mathbf{1.22 \pm 0.05}$
Amazon-3C	4.62 ± 0.48	2.30 ± 0.13	$\mathbf{4.59 \pm 0.55}$	$\mathbf{2.24 \pm 0.41}$
Amazon-6C	5.41 ± 0.09	2.52 ± 0.12	$\mathbf{5.23 \pm 0.12}$	$\mathbf{2.49 \pm 0.13}$

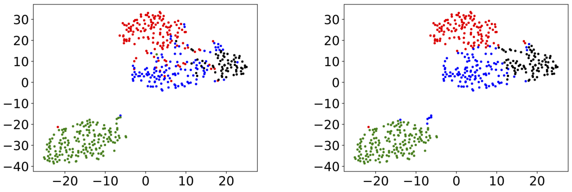
Table 3: Model generalization for new treatments.

perparameters are set as $\beta = 20$, $\lambda = 0.4$. By default, we set K as the same number of true treatment clusters, then we alter K to test the performance and disentanglement in Section 4.4. All the results are averaged over ten executions.

Metrics. Two evaluation metrics are widely used in treatment effect estimation – Rooted Precision in Estimation of Heterogeneous Effect (PEHE) [Hill, 2011] and Mean Absolute Error on ATE (ϵ_{ATE}) [Willmott and Matsuura, 2005]. Following [Saini *et al.*, 2019], we extend them into the multi-treatment setting. The evaluation is performed on a predefined set of R different treatment assignments, $\mathcal{A} = \{\boldsymbol{a}^1, ..., \boldsymbol{a}^R\}$, where $0 < R < 2^m$. For each $\boldsymbol{a}^r \in \mathcal{A}$, we have: $PEHE^r = \sqrt{\sum_{i=1}^{n}(\tau_{i,\boldsymbol{a}^r} - \hat{\tau}_{i,\boldsymbol{a}^r})^2/n}$, where $\hat{\tau}_{i,\boldsymbol{a}^r} = \hat{y}_i(\boldsymbol{a}^r) - \hat{y}_i(\mathbf{0})$ is the predicted treatment effect over $\boldsymbol{a}^r$. The average over the R treatment assignments is: $\overline{PEHE} = \frac{1}{R}\sum_{r=1}^{R} PEHE^r$. Similarly, another metric ϵ_{ATE} can also be extended to the multiple treatment setting: $\overline{\epsilon_{ATE}} = \frac{1}{R}\sum_{r=1}^{R} |\frac{1}{n}\sum_{i=1}^{n}\tau_{i,\boldsymbol{a}^r} - \frac{1}{n}\sum_{i=1}^{n}\hat{\tau}_{i,\boldsymbol{a}^r}|$.

4.3 MTE Estimation

To evaluate the proposed method in MTE estimation, we compare it with the aforementioned baselines. CEVAE, TARNET and CFR are designed for single cause ITE estimation, following [Yoon *et al.*, 2018], we apply them into multi-cause setting: we randomly select three treatments, choose the assignment $\{0, 0, 0\}$ as the control group, and other seven treatment assignments as treated group. In this way, we create seven separate single cause ITE estimation tasks, and calculate the averaged PEHE and ϵ_{ATE} over the seven tasks.We show the results of all methods when we randomly select three treatments in Table 2. We observe that DIRECT consistently outperforms the baselines. The regression methods OLS/LR and RF cannot capture the confounders and thus suffer from the confounding bias. CEVAE, TARNET and CFR model each treatment separately, thus cannot capture the dependency among treatments. BART is limited in the strong ignorability assumption. Deconf-l and Deconf-q may capture latent confounders by utilizing the assignment of multiple treamtents, but they do not utilize the observed outcome, and also lack disentanglement, which is also the limitation of DIRECT-ND. We attribute the superiority of DIRECT to two key factors: (1) our framework leverages the multiple treatment assignment and observed outcome to capture more latent confounders; (2) the disentangled representation often leads to higher performance, which is in line with the conclusion in [Ma *et al.*, 2019].



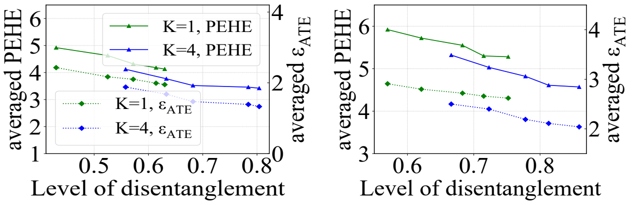
(a) Ground-truth clusters (b) Predicted clusters

Figure 3: Treatment clusters in the synthetic dataset.

Generalization for New Treatments. We assess how the proposed framework can be generalized to predict the effects of treatments that are unseen in the training data. Since none of the baseline methods can handle unseen treatments, in each dataset, we randomly hold out 20% treatments and compare the performance of our framework in predicting the causal effect of the treatment assignment over the held out treatments with/without their assignment data. The results in Table 3 show that our model can achieve comparable performance for new treatments without retraining, which benefits from the trainable network for treatment representation learning.

4.4 Disentanglement & Interpretability

We visualize the treatment representation and color them w.r.t. their true/predicted clusters in Fig. 3. We observe that the predicted clusters are very close to the ground-truth. As the treatments' representation is aligned with the confounders' representation, it indicates good macro-level disentanglement of the representations of confounders. Due to



(a) Synthetic

(b) Amazon-3C

Figure 4: MTE estimation performance w.r.t. different levels of disentanglement of the representations of confounders.

Manipulated dimension	Top-5 treatments				
In Cluster 1	**tune** musician capo	**tune** **tuner** recording	loud **tune** bass	**tune** bass price	finger player **tune**
In Cluster 2	**size** **sizing** width	**long** **size** old	**longer** little classic	**sizing** felt price	**size** **longer** feel
In Cluster 3	**battery** **charge** phone	long **battery** old	headphone **battery** quick	light **charger** cost	access connect **battery**

Table 4: Examples of the top-5 influenced treatments after modifying a dimension of confounder representation.

space limit, we only show the results in the synthetic dataset, but the observations are similar in other datasets.

To investigate the relation between the disentanglement of confounders' representations and the MTE estimation performance, we vary the hyperparameter K and β to control the level of disentanglement, and Fig. 4 shows how the estimation performance varies w.r.t. different levels of disentanglement of the confounders' representations. Here the level of disentanglement of representation with dimension d is calculated by $1 - \frac{2}{d(d-1)} \sum_{i,j} |corr(i,j)|$, where $corr(i,j)$ is the correlation between dimension i and j. Due to space limit, we only show the results on datasets Synthetic and Amazon-3C, but similar observations can be found on the other dataset. As shown in Fig. 4, treatment clustering ($K > 1$) benefits the disentanglement, and higher levels of disentanglement often leads to better MTE estimation performance.

To further show the interpretability of the learned disentangled representations, we investigate their semantics in micro-level. On Amazon-3C, after training, we use a similar way to evaluate the micro-level disentanglement as [Ma *et al.*, 2019; Wang *et al.*, 2020]. We modify one dimension of the learned confounder representations by multiplying it with a temperature factor $\tau = 10$, while keep all other dimensions fixed. Then we list the treatments with the biggest changes w.r.t. the predicted treatment assignment after modification in Table 4. Generally, we have two observations: 1) the treatments that are significantly affected can match the cluster of the modified dimension. This indicates a high-level interpretation, e.g., when we modify a dimension in $\boldsymbol{z}^{(k)}$, most of the top influenced treatments are about the musical instruments. This may imply that the cluster k corresponds to latent attributes related to musical instrument products; 2) most of the top influenced treatments contain a common word or semantically related words, which indicates that the model can capture the fine-grained latent factors by disentangled representation, e.g., when we modify a dimension, the top influenced treatments share the word "tune", which provides human-understandable semantics for the modified dimension.

5 Related Work

Multiple Treatment Effect Estimation. Traditional methods for the single cause can be extended to the multi-cause setting [Lopez *et al.*, 2017; Zanutto *et al.*, 2005; Lechner, 2001], and recent work [Sharma *et al.*, 2020] applies the neural network. However, these works are still based on the strong ignorability assumption. To mitigate this problem, a relaxed assumption called single strong ignorability is proposed in [Wang and Blei, 2019] for the multi-cause scenarios, which assumes that there do not exist unobserved single-cause confounders that causally affect the outcome and only one of the treatments. Despite its success in applications such as recommender systems [Wang *et al.*, 2018] and medication analysis [Zhang *et al.*, 2019], their captured latent confounders might be highly entangled and hard to interpret.

Disentangled Representation Learning. Disentangled representation learning has attracted significant attention recently. [Ma *et al.*, 2019] introduces a macro-micro disentangled representation learning framework for recommender systems, which achieves macro disentanglement by inferring high-level user intentions and micro disentanglement to force each dimension capture an isolated factor. We use a similar way of hierarchical disentanglement but focus on the causal inference domain. In causal inference, a line of work [Hassanpour and Greiner, 2019; Zhang *et al.*, 2020] identifies disentangled representations to separate the latent factors which influence the treatment assignment, the outcome, or both of them. Our work differs from them as the disentanglement in our work focuses on the confounders mixed in hierarchical patterns.

6 Conclusion

In this paper, we study a novel problem of disentangled multiple treatment effect estimation, and analyze its importance and challenges. We develop a novel framework DIRECT to learn disentangled representations of latent confounders for MTE estimation. Specifically, we improve the interpretability of the learned representations of confounders at both macro level and micro level. Then, we learn a trainable function to obtain the representation for each treatment by leveraging their inherent dependencies, which can be further generalized to unseen treatments. We conduct extensive experiments on different datasets, and the experimental results validate the effectiveness and interpretability of our proposed framework.

Acknowledgements

This material is, in part, supported by the National Science Foundation (NSF) under grants #2006844 and #1938167.

References

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

[Higgins *et al.*, 2016] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. Beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.

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

[Jang *et al.*, 2016] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[Lechner, 2001] Michael Lechner. Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In *Econometric evaluation of labour market policies*. 2001.

[Li *et al.*, 2017] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. Feature selection: A data perspective. *ACM Computing Surveys (CSUR)*, 50(6), 2017.

[Locatello *et al.*, 2019] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, 2019.

[Lopez *et al.*, 2017] Michael J Lopez, Roee Gutman, et al. Estimation of causal effects with multiple treatments: a review and new ideas. *Statistical Science*, 32(3), 2017.

[Louizos *et al.*, 2017] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, 2017.

[Ma *et al.*, 2019] Jianxin Ma, Chang Zhou, Peng Cui, Hongxia Yang, and Wenwu Zhu. Learning disentangled representations for recommendation. In *Advances in Neural Information Processing Systems*, 2019.

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

[Rubin, 2005] Donald B Rubin. Bayesian inference for causal effects. *Handbook of Statistics*, 25, 2005.

[Saini *et al.*, 2019] Shiv Kumar Saini, Sunny Dhamnani, Akil Arif Ibrahim, and Prithviraj Chavan. Multiple treatment effect estimation using deep generative model with task embedding. In *The World Wide Web Conference*, 2019.

[Schnabel, 2016] Schnabel. Recommendations as treatments: Debiasing learning and evaluation. In *International Conference on Machine Learning*, 2016.

[Shalit *et al.*, 2017] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, 2017.

[Sharma *et al.*, 2020] Ankit Sharma, Garima Gupta, Ranjitha Prasad, Arnab Chatterjee, Lovekesh Vig, and Gautam Shroff. Multimbnn: Matched and balanced causal inference with neural networks. *arXiv preprint arXiv:2004.13446*, 2020.

[Tran *et al.*, 2017] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[Vemuri, 2015] Vijay K. Vemuri. Causal inference for statistics, social, and biomedical sciences: An introduction by guido w. imbens and donald b. rubin. *Journal of Information Technology Case & Application Research*, 17(3-4), 2015.

[Wang and Blei, 2019] Yixin Wang and David M Blei. The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528), 2019.

[Wang *et al.*, 2018] Yixin Wang, Dawen Liang, Laurent Charlin, and David M Blei. The deconfounded recommender: A causal inference approach to recommendation. *arXiv preprint arXiv:1808.06581*, 2018.

[Wang *et al.*, 2020] Xiang Wang, Hongye Jin, An Zhang, Xiangnan He, Tong Xu, and Tat-Seng Chua. Disentangled graph collaborative filtering. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020.

[Willmott and Matsuura, 2005] Cort J Willmott and Kenji Matsuura. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate Research*, 30(1), 2005.

[Yoon *et al.*, 2018] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*, 2018.

[Zanutto *et al.*, 2005] Elaine Zanutto, Bo Lu, and Robert Hornik. Using propensity score subclassification for multiple treatment doses to evaluate a national antidrug media campaign. *Journal of Educational and Behavioral Statistics*, 30(1), 2005.

[Zhang *et al.*, 2019] Linying Zhang, Yixin Wang, Anna Ostropolets, Jami J Mulgrave, David M Blei, and George Hripcsak. The medical deconfounder: Assessing treatment effects with electronic health records. In *Machine Learning for Healthcare Conference*, 2019.

[Zhang *et al.*, 2020] Weijia Zhang, Lin Liu, and Jiuyong Li. Treatment effect estimation with disentangled latent factors. *arXiv preprint arXiv:2001.10652*, 2020.