Machine Learning (2024) 113:1799–1817
https://doi.org/10.1007/s10994-022-06246-0

InfoCEVAE: treatment effect estimation with hidden confounding variables matching

Shonosuke Harada[1] · Hisashi Kashima[1]

Received: 1 June 2022 / Revised: 16 August 2022 / Accepted: 12 September 2022 /
Published online: 2 November 2022

Abstract
Treatment effect estimation is a fundamental problem in various domains for effective decision making. While many studies assume that observational data include all the confounding variables, we cannot practically guarantee that observational data include such confounding variables, and there might be confounding variables that are not included in observational data, referred to as hidden confounding variables. Recently, variational autencoder (VAE) based methods have been successfully applied to treatment effect estimation problem. However, although they can recover a large class of latent variable models, they do not give the correct treatment effect, even when they achieve an optimal solution due to the nature of VAE loss function. We propose an efficient VAE-based method that employs information theory to estimate treatment effect and combines it with a matching technique. To the best of our knowledge, this is the first work that gives the correct treatment effect given an optimal solution using VAE-based methods. Experiments on a semi-real dataset and synthetic dataset demonstrate that the proposed method mitigates VAE problems and observational bias effectively, even under hidden confounding variables, and outperforms strong baseline methods.

Keywords Causal inference · Treatment effect estimation · Generative model

1 Introduction

Treatment effect estimation plays an essential role in decision making in various domains, such as healthcare, economic policy, and education. The goal of treatment effect estimation is to estimate the effect of an action by a decision maker. The main difficulty of treatment effect estimation based on observational data is that a treatment

Editors: Yu-Feng Li, Prateek Jain.

✉ Shonosuke Harada
sh1108@ml.ist.i.kyoto-u.ac.jp

Hisashi Kashima
kashima@i.kyoto-u.ac.jp

[1] Kyoto University, Kyoto, Japan

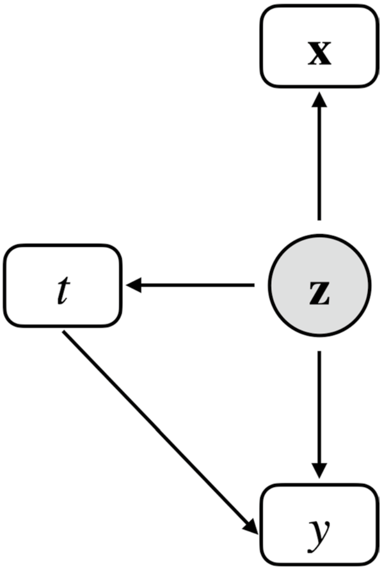


Fig. 1 A graphical model for the treatment effect estimation methods with hidden confounding variables. Hidden confounding variable **z** has an effect on treatment assignment and outcome. Treating proxy variables **x** as normal confounding variables gives incorrect treatment effect estimation

assignment is not randomized, which is often referred to as observational or selection bias. For example, elderly people might be more likely to receive drug treatment than younger people. In this example, age is a variable that impacts treatment assignment and outcome. This variable is called a confounding variable. We need to find such confounding variables to mitigate bias and give appropriate treatment effect.

In the context of treatment effect estimation, many studies usually assume that observational data include all the confounding variables. However, this assumption seems too strong and is not realistic because we cannot always practically obtain sufficient information regarding individuals to guarantee that we observe all the confounding variables. Confounding variables that are not included in observational data are often referred to as hidden confounding variables. For example, private and sensitive individual information like income might be difficult to obtain, but this variable can have an effect on treatment assignment and outcome. Without knowing confounding variables, it is impossible to know the true treatment effect, and treating proxy variables as confounding variables will lead to incorrect estimands (Rothman et al., 2008; Louizos et al., 2017). Fig. 1 illustrates a graphical model of the data generation process. In this graphical model, it is indispensable to infer **z** correctly to know the true treatment effect. Prior studies have used strong assumptions that they have knowledge regarding the nature of hidden confounding variables beforehand, like the number of categories of hidden confounding variables (Cai and Kuroki, 2008). These assumptions limit the application range of these approaches.

Recently, the Causal Effect Variational Autoencoder (CEVAE), the variational autencoder (VAE)-based method has been successfully incorporated into treatment effect estimation with the existence of hidden confounding variables (Louizos et al., 2017; Zhang et al., 2021). One of the advantages of VAE is that it can recover a large class of hidden confounding variable models thanks to the expressive power of neural networks (Tran et al., 2015). Previous researches require that we know the nature of hidden confounding variables, such as the number of categories.

Xu et al. (2021) employed a deep learning-based technique but they also assumed that they can distinguish variables that have an effect only on treatment assignment from variables which have an effect only on outcome. This assumption requires prior knowledge and seems unrealistic.

However, recent theoretical analysis revealed that the global optimum of VAE evidence lower bound (ELBO) does not correctly model the data generation process (Zhao et al., 2019) because VAE focus on reconstruction loss too much, which becomes severer when input variables have much higher dimensions than latent variables. To mitigate this problem, InfoVAE (Zhao et al., 2019), which adds a mutual information regularizer to the VAE loss function, was proposed.

This phenomenon obviously arises in VAE-based methods for treatment effect estimation and makes recovering hidden confounding variables by VAE difficult. We first remark there are datasets that the optimal solution of VAE-based methods, such as CEVAE (Louizos et al., 2017), does not give the correct treatment effect. This is a strict limitation without any guarantee when they achieve optimal solution even though they are capable of recovering them.

To mitigate these problems, we propose hidden confounding variable matching VAE, which combines VAE with information regularization and matching to give appropriate treatment effect. The proposed method obtains the correct treatment effect when it achieves the optimal solution of its loss function, even under the existence of hidden confounding variables. We summarize the contribution of this study as follows:

- To the best of our knowledge, this is the first work that shows the optimal solution of naive VAE-based methods is not a correct average treatment effect (ATE) for types of datasets.
- We propose an effective method based on information regularization and matching algorithm to mitigate hidden confounding variables and bias, with theoretical guarantee.
- In experiments using semi-synthetic and synthetic datasets, the proposed method significantly outperformed existing methods.

2 Related work

2.1 Treatment effect estimation

Treatment effect estimation plays a essential role in decision making across various domains, such as healthcare (Eichler et al., 2016; Sekhon, 2009), economic policy (LaLonde, 1986), and education (Zhao and Heffernan, 2017). We outline important studies, ranging from established methods to modern deep learning-based methods. The goal of treatment effect estimation is to understand the effect of a specific action, i.e., treatment. One of the classical methods for treatment effect estimation is matching (Rubin, 1973; Abadie & Imbens, 2006; King & Nielsen, 2019). Matching methods estimate the counterfactual outcomes by the nearest neighbor of each individual in terms of covariates. Because the curse of dimensionality makes finding appropriate nearest neighbors of each individual more difficult, propensity score matching, which defines nearest neighbors in terms of propensity score, was developed (Rosenbaum & Rubin, 1983, 1985). Tree-based methods, such as Random forest and Bayesian additive regression trees (BART), have also been applied . (Chipman et al., 2010; Hill, 2011)

Recently, deep learning-based methods have been successfully applied to the treatment effect estimation problem (Shalit et al., 2017; Johansson et al., 2016; Yao et al., 2018; Yoon et al., 2018; Louizos et al., 2017; Zhang et al., 2021; Guo et al., 2020; Harada and Kashima, 2020, 2021). Counterfactual regression (CFR) encourages individual representation of each treatment group extracted by neural networks to get closer to each other. Perfect matching combines neural networks and propensity score matching (Schwab et al., 2018), and Counterfactual propagation, which also integrates matching and graph-based semi-supervised learning, aims to estimate treatment effect using a large number of unlabeled individual data (Harada and Kashima, 2020). In particular, VAE-based methods (Louizos et al., 2017; Zhang et al., 2021) have been developed to mitigate the hidden confounding variable problem. They aim to recover hidden confounding variables by the strong expressive power of neural networks. Network structured-data also have been utilized to infer hidden confounding variables effectively (Guo et al., 2020).

2.2 VAE

VAE is one of the most famous deep generative models (Kingma and Welling, 2013) and has been widely employed in various domains, such as computer vision (Liu et al., 2017), natural language processing (Miao et al., 2016), and chemoinformatics (Liu et al., 2018). One of the advantages of VAE-based generative models is their strong expressive power based on neural networks. VAE has also been successfully applied in treatment effect estimation (Louizos et al., 2017; Zhang et al., 2021). The idea is to recover a joint distribution including hidden confounding variables expressed as latent variables to estimate treatment effect. However, recent theoretical analysis revealed that VAE will ignore the latent variables in the global optimum of the VAE loss function (Zhao et al., 2019). Hence, due to the nature of the VAE loss function, VAE-based treatment effect estimation methods face the unavoidable issue that they do not provide the correct treatment effect estimation even when their loss function achieves the optimal solution, which we will discuss in this paper.

Our goal is to fill the gap between VAE theoretical analysis and VAE-based treatment effect estimation methods, proposing an efficient method that provides theoretical guarantee of treatment effect even when there are hidden confounding variables.

3 Problem statement

In this section, we state the problem setting of treatment effect estimation. Suppose $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^{d_\mathrm{x}}$ is the $d_\mathbf{x}$ dimensional proxy variables of the i-th individual, $t_i \in \mathcal{T} = \{0, 1\}$ is the binary treatment applied to the i-th individual, and $y_i^{t_i} \in \mathcal{Y} \subset R$ is its outcome of the i-th individual. We omit the notation i of a variable when the variable can represent any individual. Given a dataset $\mathcal{D} := (\mathbf{x}_i, t_i, y_i^{t_i})_{i=1}^N$, which includes N individuals, our goal is to estimate the conditional ATE (CATE) and ATE, defined as:

$$\mathrm{CATE}\left(\mathbf{x}_i\right) := \mathbb{E}\left[\mathrm{y}_\mathrm{i}^1 | \mathbf{x}_\mathrm{i}, \mathrm{do(t = 1)}\right] - \mathbb{E}\left[\mathrm{y}_\mathrm{i}^0 | \mathbf{x}_\mathrm{i}, \mathrm{do(t = 0)}\right], \mathrm{ATE} := \mathbb{E}\left[\mathrm{CATE}\left(\mathbf{x}_\mathrm{i}\right)\right]. \quad (1)$$

We make some basic assumptions in this study: (i) stable unit treatment value: the outcome of each instance is not affected by the treatment assigned to other instances; (ii) unconfoundedness: the treatment assignment to an individual is independent of the outcome given hidden confounding variables; (iii) overlap: each individual has a positive probability of treatment assignment; (iv) smoothness: individuals who have similar hidden confounding variables have similar outcomes; (v) noisy proxy variables: hidden confounding variables can be recovered by noisy proxy variables.

4 Preliminaries

We briefly introduce some notable deep generative models based on VAE as preliminaries for clarity.

VAE (Kingma and Welling, 2013) is a widely used deep generative model that sets a prior distribution as the normal distribution. It maximizes the ELBO, consisting of reconstruction loss and the Kullback-Leibler (KL) divergence loss. It usually parameterizes $p_{\theta_{\mathbf{x}}}$ and q_ϕ by neural networks.

$$p(\mathbf{z}_i) = \prod_{j=1}^{d_{\mathbf{z}}} \mathcal{N}(z_{ij} \mid 0, 1); p_{\theta_{\mathbf{x}}}(\mathbf{x}_i \mid \mathbf{z}_i) = \prod_{j=1}^{d_{\mathbf{x}}} p_{\theta_{\mathbf{x}}}(x_{ij} \mid \mathbf{z}_i); \tag{2}$$

$$\mathcal{L}_{\text{ELBO}} = \sum_{i=1}^{N} \mathbb{E}_{q_\phi(\mathbf{z}_i|\mathbf{x}_i)}[\log p_{\theta_{\mathbf{x}}}(\mathbf{x}_i \mid \mathbf{z}_i) + \log p(\mathbf{z}_{\mathbf{i}}) - \log q_\phi(\mathbf{z}_i \mid \mathbf{x}_i)] \tag{3}$$

$$= \sum_{i=1}^{N} E_{q_\phi(\mathbf{z}_i|\mathbf{x}_i)}[\log p_{\theta_{\mathbf{x}}}(\mathbf{x}_i \mid \mathbf{z}_i) - \text{KL}(q_\phi(\mathbf{z}_i \mid \mathbf{x}_i), p(\mathbf{z}))]. \tag{4}$$

InfoVAE (Zhao et al., 2019) is a VAE with a mutual information regularization term. The mutual information term boils down to the distribution divergence between the prior distribution and marginal distribution of posterior distribution, and the function to be optimized is written as

$$\mathcal{L}_{\text{InfoVAE}} = \sum_{i=1}^{N} \mathbb{E}_{q_\phi(\mathbf{z}_i|\mathbf{x}_i)}[\log p_{\theta_{\mathbf{x}}}(\mathbf{x}_i \mid \mathbf{z}_i) - \text{KL}(q_\phi(\mathbf{z}_i \mid \mathbf{x}_i), p(\mathbf{z}))] - D(q_\phi(\mathbf{z}), p(\mathbf{z})), \tag{5}$$

where $D(q_\phi(\mathbf{z}), p(\mathbf{z}))$ is a divergence between the two distributions $p(\mathbf{z})$ and $q_\phi(\mathbf{z})$, and any divergence can be used given that $D(q_\phi(\mathbf{z}), p(\mathbf{z})) = 0$ if and only if $q_\phi(\mathbf{z}) = p(\mathbf{z})$ (Zhao et al., 2019).

CEVAE (Louizos et al., 2017) is a recently proposed VAE-based methods for CATE and ATE estimation, which aims to identify treatment effect under the presence of hidden confounding variables. To correctly specify treatment effect, we need to deal with hidden confounding variables. CEVAE assumes that such hidden confounding variables can be recovered from proxy variables as many previous studies. It takes inputs $\mathbf{x}_i, t_i, y_i^{t_i}$ to infer hidden confounding variables, $\mathbf{z}_i$.

$$p_{\theta_t}(t_i \mid \mathbf{z}_i) = \text{Bern}(h(g(\mathbf{z}_i))), \tag{6}$$

$$p_{\theta_y}(y_i^{t_i} \mid \mathbf{z}_i, t_i) = \mathcal{N}(\mu = \hat{\mu}_i, \sigma^2 = 1); \hat{\mu}_i = t_i f_1(\mathbf{z}_i) + (1 - t_i) f_0(\mathbf{z}_i), \tag{7}$$

$$q_\phi(\mathbf{z}_i \mid \mathbf{x}_i, t_i, y_i^{t_i}) = \prod_{j=1}^{d_\mathbf{z}} \mathcal{N}(\mu_{ij} = \bar{\mu}_{ij}, \sigma_j^2 = \bar{\sigma}_{ij}^2), \tag{8}$$

$$\bar{\boldsymbol{\mu}}_i = t_i \bar{\boldsymbol{\mu}}_{t=0,i} + (1 - t_i)\bar{\boldsymbol{\mu}}_{t=1,i}, \bar{\boldsymbol{\sigma}}_i^2 = t_i \boldsymbol{\sigma}_{t=0,i}^2 + (1 - t_i)\boldsymbol{\sigma}_{t=1,i}^2, \tag{9}$$

$$\bar{\boldsymbol{\mu}}_{t=0,i}, \boldsymbol{\sigma}_{t=0,i}^2 = f_3 \circ f_2(\mathbf{x}_i, y_i), \bar{\boldsymbol{\mu}}_{t=1,i}, \boldsymbol{\sigma}_{t=1,i}^2 = f_4 \circ f_2(\mathbf{x}_i, y_i), \tag{10}$$

where $h(x)$ is a sigmoid function defined as $h(x) := \frac{1}{1+\exp^{-x}}$, and g, f_0, f_1, f_2, f_3 and f_4 are neural networks. The variational lower bound is given as

$$\mathcal{L}_{\mathrm{ELBO(CEVAE)}} = \sum_{\mathrm{i=i}}^{\mathrm{N}} \mathbb{E}_{\mathrm{q}_\phi(\mathbf{z}_\mathrm{i}|\mathbf{x}_\mathrm{i},\mathrm{t}_\mathrm{i},\mathrm{y}_\mathrm{i})} \log \mathrm{p}_{\theta_{\mathbf{x},\mathrm{t}}}\left(\mathbf{x}_\mathrm{i}, \mathrm{t}_\mathrm{i}|\mathbf{z}_\mathrm{i}\right) + \log \mathrm{p}_{\theta_\mathrm{y}}\left(\mathrm{y}_\mathrm{i}^{\mathrm{t}_\mathrm{i}}\mathrm{t}_\mathrm{i}, \mathbf{z}_\mathrm{i}\right) - \mathrm{KL}\left(\mathrm{q}_\phi\left(\mathbf{z}_\mathrm{i}|\mathbf{x}_\mathrm{i}, \mathrm{t}_\mathrm{i}, \mathrm{y}_\mathrm{i}\right), \mathrm{p}(\mathbf{z})\right) \tag{11}$$

where $\log p_{\theta_{\mathbf{x},t}}(\mathbf{x}_i, t_i \mid \mathbf{z}_i) = \log p_{\theta_\mathbf{x}}(\mathbf{x}_i \mid \mathbf{z}_i) + \log p_{\theta_t}(t_i \mid \mathbf{z}_i)$. To give outcomes for new individuals, CEAVE is required to have the treatment assignment and outcome beforehand. Therefore, it employs two auxiliary loss functions to deal with new individuals. Finally, the objective function of CEVAE is given as

$$\mathcal{L}_{\mathrm{CEVAE}} = \mathcal{L}_{\mathrm{ELBO(CEVAE)}} + \sum_{i=i}^{N} \log q(t_i \mid \mathbf{x}_i) + \log q(y_i^{t_i} \mid \mathbf{x}_i, t_i). \tag{12}$$

5 CEVAE fails to estimate CATE

Treatment effect estimation with hidden confounding variables is an essential problem. CEVAE (Louizos et al., 2017) enabled us to estimate treatment effect with hidden confounding variables without any strong assumption because VAE can recover a larger function class. Prior studies have made strong assumptions, such as on the properties of proxy variables and hidden confounding variables. CEVAE can identify CATE and ATE when it recovers the joint distribution $p(\mathbf{z}, \mathbf{x}, t, y)$.

Theorem 1 *We can recover CATE and ATE when we recover the joint distribution* $p(\mathbf{z}, \mathbf{x}, t, y)$ *in Fig.* (1).*(Louizos et al.,* 2017).

Proof The proof is completed by applying the rules of do-calculus to Fig. (1). See CEVAE paper for the details (Louizos et al., 2017).

However, one of the major drawbacks of previous VAE-based methods, including CEVAE, is that they do not guarantee that they can recover the hidden confounding variables, even when when they achieve the optimal solution even though they have a capability to recover them. As a motivating example, we first note that there is a dataset for which the optimal solution of CEVAE does not give the correct CATE and ATE for new individuals.

Note that we consider the case that we use only the proxy variables $\mathbf{x}$ because assuming that we have correct outcomes y for new individuals is not realistic.

Theorem 2 *Suppose we have a dataset* $\mathcal{D} = \{\mathbf{x}_i, t_i, y_i^{t_i}\}_{i=1}^N$, *where* $\mathbf{z}_i \sim \mathcal{N}(0, 1)$, $\mathbf{x}_i \sim \mathcal{N}(\mathbf{z}_i, 1)$, $t_i \sim \mathrm{Bern}(\rho_t)$, $y_i \sim \mathcal{N}(\mathbb{I}(Cz_i > 0)t, 1)$, *where* ρ_t *is a probability of of receiving treatment and* C *is a constant value. Suppose we only observe* $\mathbf{x}_i = 1$ *or* $\mathbf{x}_i = -1$ *and* $y_i = 1$ *or* $y_i = -1$. *The optimal solution of CEVAE for this dataset does not give correct CATE and ATE.*

Proof Appendix.

This result demonstrates the insufficiency of naive VAE-based methods to recover hidden confounding variables and estimate treatment effect. Because there are numerous situations where observational data are limited and over-fitting to observational data may occur, we need to treat this problem carefully. Here we demonstrate a specific dataset, but we leave the proof of a more general form for future work.

6 InfoCEVAE with hidden confounding variables matching

The phenomenon described above arises because of the nature that VAE pushes masses away from each other and focuses on reconstruction loss too much. This becomes more crucial when we have higher dimensional proxy variables and a lower number of hidden confounding variables compared to proxy variables (i.e, $d_{\mathbf{x}} \gg d_{\mathbf{z}}$), especially when we have limited data. Some readers might think a larger number of proxy variables makes an unconfoundedness assumption, i.e., non-hidden confounding assumption, more reasonable; however, we usually can not guarantee that there are no hidden confounding variables in practice, and moreover, sometimes we never have access to the hidden confounding variables (e.g., variables including sensitive privacy information) even when we can easily obtain some proxy variables.

The straightforward solution to obtain the correct ATE using VAE-based methods is to employ the theoretical analysis of InfoVAE (Zhao et al., 2019), which adds the mutual information regularization term to the original ELBO of VAE.

The ELBO of InfoCEVAE will be adding the information regularization term to CEVAE given as

$$\begin{aligned}\mathcal{L} = \sum_{i=i}^{N} \mathbb{E}_{q_\phi(\mathbf{z}_i \mid \mathbf{x}_i, t_i, y_i)}[\log p_{\theta_{\mathbf{x},t}}(\mathbf{x}_i, t_i \mid \mathbf{z}_i) + \log p(y_i^{t_i} \mid t_i, \mathbf{z}_i) \\ - \mathrm{KL}(q_\phi(\mathbf{z}_i \mid \mathbf{x}_i, t_i, y_i^{t_i}), p(\mathbf{z}))] - D(q_\phi(\mathbf{z}), p(\mathbf{z})).\end{aligned} \tag{13}$$

We can employ the several measures of divergence D between two probability distributions, such as 2-Wasserstein distance given that $D(q(\mathbf{z}), p(\mathbf{z})) = 0$ if and only if $q(\mathbf{z}) = p(\mathbf{z})$. We use the 2-Wasserstein distance as D, and the 2-Wasserstein distance for two Gaussian distributions is written as:

$$D(\mathcal{N}(\mu_1, \sigma_1), \mathcal{N}(\mu_2, \sigma_2)) = \|\mu_1 - \mu_2\|^2 + \|\sigma_1 - \sigma_2\|^2. \tag{14}$$

We can also get correct CATE and ATE when the model achieves the optimal solution of the objective function $q_\phi(\mathbf{z}) = p(\mathbf{z})$.

Theorem 3 *The optimal solution of InfoCEVAE gives the correct CATE and ATE.*

Proof According to the Proposition of InfoVAE, we obtain the optimal solution when we achieve $q_\phi(y \mid t, \mathbf{z}) = p(y \mid t, \mathbf{z})$ and $q_\phi(\mathbf{z} \mid \mathbf{x}, t, y) = p(\mathbf{z} \mid \mathbf{x}, t, y)$. Therefore,

$$\widehat{CATE}(\mathbf{x}) = p_\theta(y \mid t = 1, \mathbf{x}) - p_\theta(y \mid t = 0, \mathbf{x}) \tag{15}$$

$$= \int_{\mathcal{Z}} p_\theta(y = 1 \mid t = 1, \mathbf{z}) q_\phi(\mathbf{z} \mid \mathbf{x}, t = 0, y) - p_\theta(y = 1 \mid t = 0, \mathbf{z}) q_\phi(\mathbf{z} \mid \mathbf{x}, t = 1, y) dz \tag{16}$$

$$= \int_{\mathcal{Z}} p(y \mid t = 1, \mathbf{z}) p(\mathbf{z} \mid \mathbf{x}, t = 0, y) - p(y \mid t = 0, \mathbf{z}) p(\mathbf{z} \mid \mathbf{x}, t = 1, y) dz \tag{17}$$

$$\begin{aligned} = \int_{\mathcal{Z}} & p(y \mid do(t = 1), \mathbf{z}) p(\mathbf{z} \mid \mathbf{x}, do(t = 0), y) \\ & - p(y \mid do(t = 0), \mathbf{z}) p(\mathbf{z} \mid \mathbf{x}, do(t = 1), y) dz \end{aligned} \tag{18}$$

$$= p(y \mid \mathbf{x}, do(t = 1)) - p(y \mid \mathbf{x}, do(t = 0)) \tag{19}$$

$$= CATE(\mathbf{x}). \tag{20}$$

□

However, this naive approach requires that we obtain the correct outcome function, i.e., $p(y \mid \mathbf{z}, t) = p_\theta(y \mid \mathbf{z}, t)$ as well as the propensity score function $p(t \mid \mathbf{z})$. Obtaining the correct outcome function is challenging, especially when we need to consider observational bias. Say we obtain $q_\phi(\mathbf{z}) = p(\mathbf{z})$ once, and then our goal is to recover the joint distribution $\int_z q_\phi(\mathbf{z}, \mathbf{x}, t, y) dz = p(\mathbf{x}, t, y)$. Therefore we need to ensure that we have $q(\mathbf{x}, t, y \mid \mathbf{z}) = p(\mathbf{x}, t, y \mid \mathbf{z})$. Hence, to achieve the optimal solution of InfoCEVAE, we need to learn θ such that $p_\theta(\mathbf{x}, t, y \mid \mathbf{z}) = p(\mathbf{x}, t, y \mid \mathbf{z})$, which means we need to learn the correct outcome function only by skewed observational data. This is almost impossible without modification. The estimator θ_y given observational data is given as

$$\theta_y^{obs} = \mathrm{argmin}_{\theta_y \in \Theta} - \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{q_\phi(\mathbf{z}_i \mid \mathbf{x}_i, t_i, y_i)} [\log p_{\theta_y}(y_i \mid t_i, \mathbf{z}_i)] \tag{21}$$

$$\simeq \mathrm{argmin}_{\theta_y \in \Theta} - \mathbb{E}_{p_{\mathcal{D}_{\text{train}}}(t,y)} [\mathbb{E}_{q_\phi(\mathbf{z}_i \mid \mathbf{x}, t_i, y_i)} [\log p_{\theta_y}(y_i \mid t_i, \mathbf{z}_i)]]. \tag{22}$$

However, this estimator is not consistent because of observational bias caused by hidden confounding variables.

$$\lim_{N \to \infty} \theta_y^{obs} = \mathrm{argmin}_{\theta_y \in \Theta} - \mathbb{E}_{p_{\mathcal{D}_{\text{train}}}(t,y)} [\mathbb{E}_{q_\phi(\mathbf{z} \mid \mathbf{x}, t, y)} [\log p_{\theta_y}(y_i \mid t_i, \mathbf{z}_i)]] \tag{23}$$

$$\neq \mathrm{argmin}_{\theta_y \in \Theta} - \mathbb{E}_{p(t)p(y)} [\mathbb{E}_{q_\phi(\mathbf{z} \mid \mathbf{x}, t, y)} [\log p_{\theta_y}(y_i \mid t_i, \mathbf{z}_i)]]. \tag{24}$$

$$\because p_{\mathcal{D}_{\text{train}}}(t,y) = \int_{\mathcal{Z}} p(y \mid t, \mathbf{z}) p(t \mid \mathbf{z}) p(\mathbf{z}) d\mathbf{z} \neq \int_{\mathcal{Z}} p(y \mid t, \mathbf{z}) p(t) p(\mathbf{z}) d\mathbf{z} \tag{25}$$

$$= p(t)p(y). \tag{26}$$

Note that we assume the treatment assignment is randomized when evaluating the model. To resolve this problem, we propose an effective algorithm based on latent variables and a matching algorithm. Note that InfoCEVAE guarantees the correct treatment effect when it achieves the optimal solution, although it is challenging to obtain. However, CEVAE cannot provide the optimal treatment effect, even when it achieves the optimal solution.

6.1 Hidden confounding variables matching

To mitigate the above issue, we aim to recover hidden confounding variables by only proxy variables, not using outcomes like CEVAE. This approach sounds reasonable because the assumption that we can recover hidden confounding variables only by proxy variables when we have such high dimensional proxy variables is quite valid (Zhang et al., 2021). Moreover, the advantage of using only proxy variables is that we do not need to predict outcomes for new individuals. Hence, hidden confounding variables are inferred as:

$$q_\phi(\mathbf{z}_i \mid \mathbf{x}_i) = \prod_{j=1}^{d_\mathbf{z}} \mathcal{N}(\mu = \mu_{ij}, \sigma = \sigma_{ij}); p(\mathbf{z}_i) = \mathcal{N}(0, 1). \tag{27}$$

The ELBO is given as:

$$\begin{aligned} \mathcal{L}_{\text{InfoCEVAE}} = \sum_{i=i}^{N} \mathbb{E}_{q_\phi(\mathbf{z}_i|\mathbf{x}_i)} \Big[\log p_{\theta_{\mathbf{x},t}}(\mathbf{x}_i, t_i|\mathbf{z}_i) + \log p_{\theta_y}\big(y_i|t_i, \mathbf{z}_i\big) \\ -\text{KL}\big(\log q_\phi\big(\mathbf{z}_i|\mathbf{x}_i\big), p(\mathbf{z})\big)\Big] - \lambda D\big(q_\phi(\mathbf{z}), p(\mathbf{z})\big), \end{aligned} \tag{28}$$

where λ is a hyper-parameter that controls the strength of regularization.

For bias mitigation, we propose latent variable matching, which makes use of latent variables to match individuals. Thanks to the theoretical advantage of InfoCEVAE, we can find the matching based on the some metric using latent variables. By nearest neighbor matching, we construct the counterfactual outcome for each individual i as

$$\hat{y}_i^{\bar{t}_i} = \frac{1}{k} \sum_{j \in \text{NN}(\mathbf{z}_i, k)} y_j^{t_j}, \tag{29}$$

where $\text{NN}(\mathbf{z}_i, k) = \{i_1, \dots, i_k\}$ is a set of indices ordered by a similarity that defines nearest neighbors of $\mathbf{z}_i$, and $\bar{t}_i \in \mathcal{T}$ represents the other treatment of t_i. Here, we consider two variants of nearest neighbor selection: (i) Euclidean distance of means of the two latent variables: (ii) propensity score matching. The advantage of (i) is that we can use all the information of latent variables and does not need to infer propensity score, while (i) might fail to find good matching in higher dimensions of latent variables. The pros and cons of (ii) are the opposite of those of (i). Note that under the smoothness assumption and when we achieve the optimal solution of InfoCEVAE, both hidden confounding variable matching methods yield consistency estimators. We compute the log-likelihood of counterfactual outcome as

$$\mathcal{L}_{\mathrm{cf}} = \sum_{i=i}^{N} \mathbb{E}_{q_\phi(\mathbf{z}_i|\mathbf{x}_i)}[\log p(\hat{y}_i^{\bar{t}_i} \mid \bar{t}_i, \mathbf{z}_i)]. \tag{30}$$

Finally, the objective function to be optimized is given as

$$\mathcal{L} = \mathcal{L}_{\mathrm{cf}} + \mathcal{L}_{\mathrm{InfoCEVAE}}. \tag{31}$$

Theorem 4 *The optimal solution of InfoCEVAE with hidden variables matching gives the consistent treatment effect estimator under the smoothness assumption.*

Proof According to the theorem of InfoVAE, we can obtain the correct posterior function when we obtain the optimal solution. Using correct hidden confounding variables, we can obtain correct counterfactual outcomes under the smoothness assumption. Using the correct counterfactual outcomes as well as factual outcomes, we can obtain a consistent estimator, which yields the correct ATE. □

7 Experiments

We validated the performance of the proposed method, especially when there are hidden confounding variables. First, we introduce the datasets used in the experiments, and detail the experimental settings.

7.1 Datasets

We rarely have real-world datasets due to the counterfactual nature of treatment effect estimation problem. We employed a widely-used semi-synthetic dataset and a synthetic dataset for this experiment.

7.1.1 News dataset (Johansson et al., 2016)

This is a dataset including opinions of media consumers for news articles (Johansson et al., 2016).[1] It contains 5, 000 news articles and outcomes generated from the NY Times corpus[2]. Each article is consumed on desktop ($t = 0$) or mobile ($t = 1$), and it is assumed that media consumers prefer to read some articles on mobile than desktop. We use the News dataset by setting the scale parameter for outcome in previous research (Johansson et al., 2016) as 200. Each article is generated by a topic model and represented in the bag-of-words representation. The size of the vocabulary is 3, 477. As preprocessing, we apply principal component analysis (PCA) with $d_{\mathbf{z}} = 30$. To simulate hidden confounding variables situation, we generate proxy variables using these variables after PCA. More

[1] https://www.fredjo.com/.

[2] https://archive.ics.uci.edu/ml/datasets/Bag+of+Words.

concretely, we treat these variables as hidden confounding variables z_{ij} and generate proxy variables as

$$x_{i,j\times 1,\dots,j\times d_{proxy}} \sim \mathcal{N}\big(z_{ij}, \sigma_{\mathbf{z}}^2\big), \tag{32}$$

$$\mathbf{x}_i = [x_{i,1}, \dots, x_{i,30\times d_{\text{proxy}}}], \tag{33}$$

where $\sigma_{\mathbf{z}}$ is a standard deviation of the entire variables after PCA , d_{proxy} stands for the number of proxy variables per hidden confounding variables and [] represents the concatenation. We set d_{proxy} as 30 for the News dataset.

7.1.2 Synthetic dataset

The synthetic dataset is a benchmark generated in this study. This dataset includes 5, 000 individuals, binary treatment, and continuous outcomes. We generated the dataset according to the following procedure:

$$z_{ij} \sim \mathcal{N}(0, 1) \;\; (j = 1, \dots, 5), \tag{34}$$

$$x_{i,j\times 1,\dots,j\times d_{\text{proxy}}} \sim \mathcal{N}(10z_{ij}, 1), \tag{35}$$

$$\mathbf{x}_i = [x_{i,1}, \dots, x_{i,5\times d_{\text{proxy}}}], \tag{36}$$

$$t_i \sim \mathrm{Bern}\left(\alpha h\left(\sum_{j=1}^{5} z_{ij}\right)\right), y_i \sim \mathcal{N}\left(3\mathbb{I}\left(\sum_{\mathrm{j}=1}^{5} \mathrm{z}_{\mathrm{ij}} \geq 0\right) \times \mathrm{t}_{\mathrm{i}} + 5\mathrm{t}_{\mathrm{i}}, 1\right) \tag{37}$$

where $\alpha \geq 0$ is a variable that controls the strength of observational bias, and $\mathbb{I}(x)$ is an indicator function that is 1 if x is True and 0 otherwise. Note again that h is a sigmoid function. Larger α means we have severer observational bias, and setting α as 0 represents a randomized controlled trial. We clamped the treatment assignment probability at 0.01 and 0.99. We change d_{proxy} as ranging from 10 to 500 for the Synthetic dataset. Unless otherwise stated, we report the results when $d_{\text{proxy}} = 500$.

In the experiments, we investigated the robustness against the bias strength by changing the value of α.

7.2 Experimental settings

We split the all individuals into 20%, 40%, and 40% train, validation, and test data, respectively. Note that we especially focus on the case when train data are limited because overestimation becomes severer. As base neural network models including VAE-based methods, we use two-layer neural networks. We also set the number of neurons (i.e, the number of representations) as 50 for TARNet and CFR. We use the *elu* function (Clevert et al., 2015) as the activation function for all neural networks.

As evaluation metrics, we employ *ATE error* defined as

$$\epsilon_{\mathrm{ATE}} = \frac{1}{N}\sum_{i=1}^{N}\left|\left(y_i^1 - y_i^0 - (\hat{y}_i^1 - \hat{y}_i^0)\right)\right|$$

and *precision in estimation of heterogeneous effect (PEHE)* used in previous researches (Hill, 2011; Johansson et al., 2016). ϵ_{PEHE} is the estimation error of individual treatment effects and is defined as

$$\sqrt{\epsilon_{\mathrm{PEHE}}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i^1 - y_i^0 - (\hat{y}_i^1 - \hat{y}_i^0))^2}.$$

The hyper-parameters are tuned based on the prediction loss using the observed outcomes on the validation data. We log-uniform randomly choose the hyper-paramters λ ranging from $1e-3$ to $1e3$ ten times, and the final hyper-parameter is selected based on the prediction loss using the outcomes on the validation data. For CEVAE, we compute the ELBO using validation data and use the model at the epoch when the ELBO for validation data achieves the maximum value. We report the average results of 10 trials on the Synthetic dataset and 20 trials on the News dataset.

7.3 Baseline methods

We compare the proposed method with the following baseline methods including VAE-based methods. Unless otherwise stated, we use the concatenation of proxy variables and treatment assignment coded as a one-hot vector as the input of predictive models of (i) and (ii).

(i) Ridge is the ordinary linear regression methods with L2 regularization.

(ii) Random forest (RF) (Breiman, 2001) and BART (Chipman et al., 2010; Hill, 2011) are the predictive models based on the decision tree.

(iii) TARNet (Shalit et al., 2017) is a deep neural network model that has shared layers for representation learning and different layers for outcome prediction for treatment and control instances. Counterfactual regression (CFR) (Shalit et al., 2017) is a state-of-the-art deep neural network model based on balanced representations between treatment and control instances. We use the Wasserstein distance.

(iv) CEVAE (Louizos et al., 2017) is a VAE-based treatment effect estimation method.

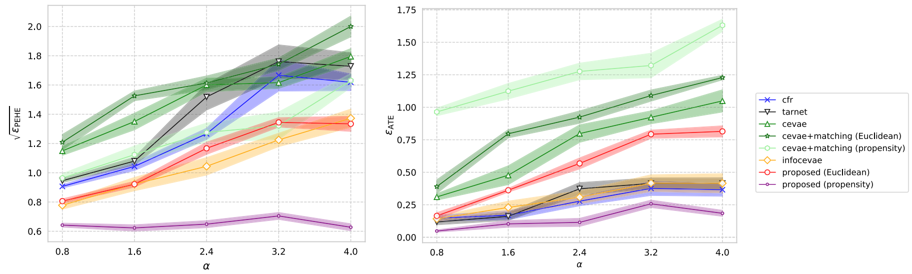
7.4 Results

We first assess the full results in comparison with the baseline methods, and then we investigate how the performance changes as we change the size of proxy variables or the strength of observational bias. Table 1 gives a performance comparison of the proposed method with the baseline methods. Overall, the proposed method outperforms baseline methods significantly. On the News dataset, the both approaches of proposed method show significant improvement from the baseline methods. On the Synthetic dataset, the proposed method with propensity score matching works better. This result makes sense because the propensity score and outcome have strong correlation in this dataset. However, the proposed method with the Euclidean matching does not work

Table 1 Performance comparison on the News dataset and the Synthetic dataset in terms of PEHE and ATE. Lower is better.

Method	News		Synthetic	
	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}
Mean	$^{\dagger}14.325_{\pm 0.128}$	$^{\dagger}3.921_{\pm 0.551}$	$^{\dagger}1.980_{\pm 0.010}$	$^{\dagger}1.292_{\pm 0.015}$
Ridge	$^{\dagger}13.764_{\pm 0.959}$	$0.911_{\pm 0.190}$	$^{\dagger}1.570_{\pm 0.019}$	$^{\dagger}0.438_{\pm 0.061}$
RF	$^{\dagger}10.246_{\pm 0.959}$	$^{\dagger}2.211_{\pm 0.385}$	$^{\dagger}1.465_{\pm 0.021}$	$^{\dagger}0.854_{\pm 0.024}$
BART	$^{\dagger}13.618_{\pm 0.921}$	$^{\dagger}1.310_{\pm 0.221}$	$^{\dagger}2.758_{\pm 0.332}$	$^{\dagger}1.829_{\pm 0.332}$
TARNet	$^{\dagger}8.988_{\pm 0.488}$	$^{\dagger}1.135_{\pm 0.200}$	$^{\dagger}1.729_{\pm 0.093}$	$^{\dagger}0.415_{\pm 0.043}$
CFR	$^{\dagger}9.125_{\pm 0.488}$	$^{\dagger}1.643_{\pm 0.268}$	$^{\dagger}1.619_{\pm 0.057}$	$^{\dagger}0.366_{\pm 0.049}$
CEVAE	$^{\dagger}9.389_{\pm 0.600}$	$^{\dagger}2.319_{\pm 0.381}$	$^{\dagger}1.795_{\pm 0.053}$	$^{\dagger}1.048_{\pm 0.085}$
CEVAE w/Euclidean	$^{\dagger}8.659_{\pm 0.524}$	$^{\dagger}1.196_{\pm 0.250}$	$^{\dagger}2.000_{\pm 0.053}$	$^{\dagger}1.229_{\pm 0.017}$
CEVAE w/propensity	$^{\dagger}8.642_{\pm 0.523}$	$^{\dagger}1.136_{\pm 0.254}$	$^{\dagger}1.630_{\pm 0.046}$	$^{\dagger}0.683_{\pm 0.013}$
InfoCEVAE	$^{\dagger}8.453_{\pm 0.510}$	$^{\dagger}1.742_{\pm 0.242}$	$^{\dagger}1.373_{\pm 0.062}$	$^{\dagger}0.415_{\pm 0.073}$
InfoCEVAE w/Euclidean	$\mathbf{7.934}_{\pm 0.478}$	$0.928_{\pm 0.172}$	$^{\dagger}1.334_{\pm 0.032}$	$^{\dagger}0.815_{\pm 0.042}$
InfoCEVAE w/propensity	$\mathbf{7.930}_{\pm 0.476}$	$\mathbf{0.835}_{\pm 0.147}$	$\mathbf{0.626}_{\pm 0.023}$	$\mathbf{0.184}_{\pm 0.022}$

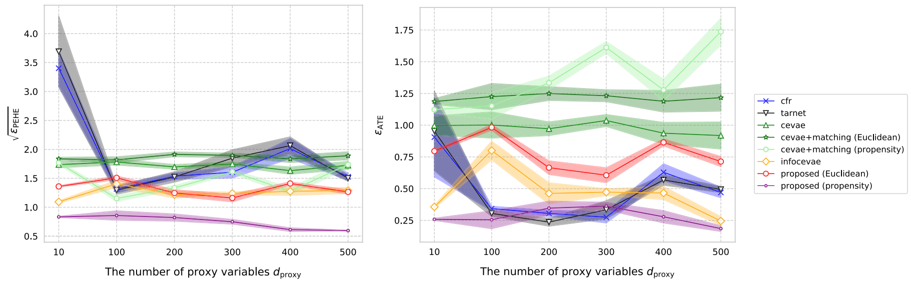
† indicates that the proposed method show statistically significantly better result by the paired *t*-test with $p < 0.05$. Bold results show the best results in term of average. We also show standard errors for 20 and 10 times repeated experiments for the News dataset and the Synthetic dataset, respectively



(a) $\sqrt{\epsilon_{PEHE}}$ **(b)** ϵ_{ATE}

Fig. 2 Performance comparison as the change of observational bias α. Lower is better. Whereas baseline methods suffered a observational bias and get degrade its performance, the proposed method demonstrates its robustness to the observational bias and almost entirely surpass the baseline methods in the both metrics. Especially, the proposed method consistently shows the affordable performance in ATE

because nearby individuals in terms of the Euclidean distance of hidden confounding variables do not necessarily become the good matching unless we have a large amount of individuals. Meanwhile the predictive performance deteriorates as selection bias becomes stronger, the proposed method shows robustness to selection bias and consistently outperforms the baseline methods. Figures 2 and 3 demonstrate the change of predictive performances as we change the strength of bias α and the number of proxy variables d_{proxy}. Whereas the baseline methods suffer from observational bias, the proposed method show robustness to it. Although, the baseline methods result in limited improvement, the proposed method also can deal with and make use of high dimensional proxy variables and improve its predictive performance.



(a) $\sqrt{\epsilon_{PEHE}}$ **(b)** ϵ_{ATE}

Fig. 3 Performance comparison as the change of the number of proxy variables. Lower is better. While the baseline methods do not improve their predictive performances as the number of proxy variables increase, the proposed method with propensity score matching achieves almost entirely the best results, especially significant in $\sqrt{\epsilon_{\rm PEHE}}$

8 Conclusion

In this study, we considered treatment effect estimation problem with hidden confounding variables using VAE. VAE has been used to recover hidden confounding variables by making use of its large capability. We first pointed out that the optimal solution of CEVAE is not the correct ATE. We propose an efficient algorithm to recover hidden confounding variables and estimate treatment effect making use of mutual information and matching techniques. Experiments on semi-synthetic and synthetic datasets demonstrate the effectiveness of the proposed method, especially when we have higher dimensional proxy variables but still hidden confounding variables.

Appendix a proof of Theorem 2.

Proof Note that **x** and **z** represent vectors in the main paper but they are also scalar values in this proof. The ATE of this dataset is

$$\mathbb{E}[y^1] - \mathbb{E}[y^0] = p(\mathbf{z}_i \geq 0)\mathrm{C} - 0 \tag{A1}$$

$$= p(\mathbf{z}_i \geq 0)\mathrm{C}. \tag{A2}$$

We first show naive CEVAE loss has unbounded reward if the proxy variables come from Gaussian distribution family. This step mainly follows the same procedure as InfoVAE (Zhao et al., 2019). We consider the following restricted a Gaussian models and if we achieve the infinite ELBO in this model, we can achieve the infinite ELBO in any model with more expresiveness than this model.

$$p(\mathbf{x} \mid \mathbf{z}) = \begin{cases} \mathcal{N}(1, \sigma^2) & (\mathbf{z} \geq 0) \\ \mathcal{N}(-1, \sigma^2) & (\mathbf{z} < 0)' \end{cases}$$

$$q(\mathbf{z} \mid \mathbf{x}) = \begin{cases} \mathcal{N}(a, \sigma_q^2) & (\mathbf{x} \geq 0) \\ \mathcal{N}(-a, \sigma_q^2) & (\mathbf{x} < 0)' \end{cases}$$

$$p(t \mid \mathbf{z}) = \begin{cases} p_1 & (\mathbf{z} \geq 0) \\ p_0 & (\mathbf{z} < 0) \end{cases}, \; p(y \mid \mathbf{z}) = \begin{cases} \mathcal{N}(\mathrm{C}, 1) & (\mathbf{z} \geq 0, t = 1) \\ \mathcal{N}(0, 1) & (\mathbf{z} < 0, t = 1) \\ \mathcal{N}(0, 1) & (t = 0). \end{cases}$$

The ELBO for $\mathbf{x} = 1$ is

$$\mathcal{L}_{AE}(\mathbf{x} = 1) \equiv \mathbb{E}_{q(\mathbf{z}|\mathbf{x}=1)}[\log p(\mathbf{x} = 1 \mid \mathbf{z})] + \mathbb{E}_{q(\mathbf{z}|\mathbf{x}=1)}[\log p(\mathbf{x} = 1 \mid \mathbf{z})] \tag{A3}$$

$$\mathbb{E}_{q(\mathbf{z}|\mathbf{x}=1)}[\log p(t \mid \mathbf{z})] + \mathbb{E}_{q(\mathbf{z}|\mathbf{x}=1)}[\log p(y \mid \mathbf{z}, t)]. \tag{A4}$$

Taking the gradient of $\mathcal{L}_{AE}(\mathbf{x} = 1)$,

$$\frac{\partial \mathcal{L}_{AE}(\mathbf{x} = 1)}{\sigma} = -\frac{1}{\sigma} + \frac{4}{\sigma^3} q(\mathbf{z} \leq 0 \mid \mathbf{x} = 1) = 0, \tag{A5}$$

and the optimal solution for $\mathcal{L}_{AE}(\mathbf{x} = 1)$ is achieved when $\sigma = 2\sqrt{q(\mathbf{z} \leq 0 \mid \mathbf{x} = 1)}$. Therefore,

$$\mathcal{L}^*_{AE}(x = 1) = -\frac{1}{2} \log q(\mathbf{z} \leq 0 \mid \mathbf{x} = 1) + \text{Constant}. \tag{A6}$$

$q(\mathbf{z} \leq 0 \mid \mathbf{x} = 1)$ is the sum of Gaussian tail probabilities. Hence in the limit $\sigma_q \rightarrow 0$, $a \rightarrow \infty$,

$$\mathcal{L}^*_{AE}(\mathbf{x} = 1) = \Theta(\frac{a^2}{\sigma_q^2}). \tag{A7}$$

$$\mathcal{L}_{REG} = -\mathrm{KL}(q_\phi(\mathbf{z} \mid \mathbf{x} = 1)||p(\mathbf{z})) \tag{A8}$$

$$= \log \sigma_q - \frac{\sigma_q^2}{2} - \frac{a^2}{2} + \frac{1}{2}. \tag{A9}$$

Therefore, we can achieve unbounded ELBO.

$$\lim_{\sigma_q \rightarrow 0, a \rightarrow \infty} \mathcal{L}_{\mathrm{ELBO}}(\mathbf{x} = 1) = \lim_{\sigma_q \rightarrow 0, a \rightarrow \infty} \mathcal{L}^*_{\mathrm{AE}}(\mathbf{x} = 1) + \mathcal{L}_{\mathrm{REG}}(\mathbf{x} = 1) \tag{A10}$$

$$\rightarrow \infty. \tag{A11}$$

Next, we show that treating them as normal confounding variables will not give the correct treatment effect.

$$\mathbb{E}[y^1 \mid t = 1] = \int_{\mathcal{X}} p(y \mid t = 1, \mathbf{x}) p(\mathbf{x}) d\mathbf{x} \tag{A12}$$

$$= \int_{\mathcal{X}} \frac{p(t = 1, \mathbf{x} \mid y) p(y)}{p(t = 1, \mathbf{x})} p(\mathbf{x}) d\mathbf{x} \tag{A13}$$

$$= \int_{\mathcal{X}} \frac{\int_{\mathcal{Z}} p(y, t = 1, \mathbf{x} \mid \mathbf{z}) p(\mathbf{z}) d\mathbf{z}}{\int_{\mathcal{Z}} p(t = 1, \mathbf{x} \mid \mathbf{z}) p(\mathbf{z}) d\mathbf{z}} p(\mathbf{x}) d\mathbf{x} \tag{A14}$$

$$= \int_{\mathcal{X}} \frac{\int_{\mathcal{Z}} p(y, t = 1, \mathbf{x} \mid \mathbf{z} \geq \mathrm{c}) p(\mathbf{z} \geq 0) d\mathbf{z} + \int_{\mathcal{Z}} p(y, t = 1, \mathbf{x} \mid \mathbf{z} < 0) p(\mathbf{z} < 0) d\mathbf{z}}{\int_{\mathcal{Z}} p(t = 1, \mathbf{x} \mid \mathbf{z} \geq 0) p(\mathbf{z} \geq \mathrm{c}) d\mathbf{z} + \int_{\mathcal{Z}} p(t = 1, \mathbf{x} \mid \mathbf{z} < 0) p(\mathbf{z} < 0) d\mathbf{z}} p(\mathbf{x}) d\mathbf{x} \tag{A15}$$

$$= \int_{\mathcal{X}} \frac{\int_{\mathcal{Z}} p(y, t = 1, \mathbf{x} \mid \mathbf{z} \geq 0) p(\mathbf{z} \geq 0) d\mathbf{z}}{\int_{\mathcal{Z}} p(t = 1, \mathbf{x} \mid \mathbf{z} \geq 0) p(\mathbf{z} \geq 0) d\mathbf{z} + \int_{\mathcal{Z}} p(t = 1, \mathbf{x} \mid \mathbf{z} < 0) p(\mathbf{z} < 0) d\mathbf{z}} p(\mathbf{x}) d\mathbf{x} \tag{A16}$$

$$= \int_{\mathcal{Z}} \frac{\rho_t \mathrm{C} p(\mathbf{x} \mid \mathbf{z} \geq 0) + \rho_t' p(\mathbf{x} \mid \mathbf{z} < 0) 0}{\rho_t p(\mathbf{x} \mid \mathbf{z} \geq 0) + \rho_t' p(\mathbf{x} \mid z < 0)} p(\mathbf{x}) d\mathbf{x} \tag{A17}$$

$$= \int_{\mathcal{X}} \frac{\rho_t \mathrm{C} p(\mathbf{x} \mid \mathbf{z} \geq 0)}{\rho_t p(\mathbf{x} \mid \mathbf{z} \geq 0) + \rho_t' p(\mathbf{x} \mid \mathbf{z} < 0)} p(\mathbf{x}) d\mathbf{x}. \tag{A18}$$

One case where this procedure gives the correct estimand is the case when the treatment assignment is randomized, i.e., $\rho_t = \rho'$.

$$\mathbb{E}[\hat{y}^1 \mid t = 1] = \int_{\mathcal{X}} \frac{\rho_t \mathrm{C} p(\mathbf{x} \mid \mathbf{z} \geq 0)}{\rho_t p(\mathbf{x} \mid \mathbf{z} \geq 0) + \rho_t' p(\mathbf{x} \mid \mathbf{z} < 0)} p(\mathbf{x}) d\mathbf{x} \tag{A19}$$

$$= \int_{\mathcal{X}} \frac{\mathrm{C} p(\mathbf{x} \mid \mathbf{z} \geq 0)}{p(\mathbf{x} \mid \mathbf{z} \geq 0) + p(\mathbf{x} \mid \mathbf{z} < 0)} p(\mathbf{x}) d\mathbf{x} \tag{A20}$$

$$= \mathrm{C} p(\mathbf{x} \mid \mathbf{z} \geq 0). \tag{A21}$$

Next, we try to estimate treatment effect using CEVAE and prove the estimand is wrong even if we obtain the correct outcome function.

$$\mathbb{E}[\hat{y}^1 \mid t = 1] = \int_{\mathcal{X}} \int_{\mathcal{Z}} p(y \mid t = 1, \mathbf{z}) p(\mathbf{z} \mid \mathbf{x}) p(\mathbf{x}) d\mathbf{z} d\mathbf{x} \tag{A22}$$

$$= \frac{1}{2} \int_{\mathcal{Z}} p(y \mid t = 1, \mathbf{z}) p(\mathbf{z} \mid \mathbf{x} = 1) d\mathbf{z} + \frac{1}{2} \int_{\mathcal{Z}} p(y \mid t = 1, \mathbf{z}) p(\mathbf{z} \mid \mathbf{x} = -1) d\mathbf{z} \tag{A23}$$

$$= \frac{1}{2} \mathrm{C} p(\mathbf{z}_i \geq 0 \mid \mathbf{x} = 1) + \frac{1}{2} \mathrm{C} p(\mathbf{z}_i \geq 0 \mid \mathbf{x} = -1) \tag{A24}$$

$$\simeq \frac{1}{2} \mathrm{C}. \tag{A25}$$

$$\mathbb{E}[\hat{y}^0 \mid t = 0] = \int_{\mathcal{X}} \int_{\mathcal{Z}} p(y \mid t = 0, \mathbf{z}) p(\mathbf{z} \mid \mathbf{x}) p(\mathbf{x}) d\mathbf{z} d\mathbf{x} \tag{A26}$$

$$= \frac{1}{2}\int_{\mathcal{Z}} p(y \mid t=0, \mathbf{z})p(\mathbf{z} \mid \mathbf{x}=1)d\mathbf{z} + \frac{1}{2}\int_{\mathcal{Z}} p(y \mid t=0, \mathbf{z})p(\mathbf{z} \mid \mathbf{x}=-1)d\mathbf{z} \tag{A27}$$

$$= \frac{1}{2}\mathrm{C}p(\mathbf{z}_i \geq 0 \mid \mathbf{x}=1) + \frac{1}{2}\mathrm{C}p(\mathbf{z}_i \geq 0 \mid \mathbf{x}=-1) \tag{A28}$$

$$\simeq \frac{1}{2}\mathrm{C} \neq 0. \tag{A29}$$

$$\widehat{\mathrm{ATE}} = \mathbb{E}[\hat{y} \mid t=1] - \mathbb{E}[\hat{y} \mid t=0] = 0 \tag{A30}$$

$$\neq \mathrm{C}p(\mathbf{z}_i \geq 0) = \mathrm{ATE}. \tag{A31}$$

Author Contribution Conceptualization: SH; Methodology: SH, HK; Formal analysis and investigation: SH; Writing - original draft preparation: SH; Writing - review and editing: HK; Funding acquisition: SH, HK; Resources: HK; Supervision: HK.

Funding This work was supported by JSPS KAKENHI Grant Number 21J14882 and 20H04244.

Data availibility statement The semi-synthetic dataset used in the experiment are publicly available. https://www.fredjo.com/

Code availability We are preparing to publish the source code.

Declarations

Conflict of interest Not applicable.

Ethics approval Not applicable.

Consent to participate Not applicable

References

Abadie, A., & Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica, 74*(1), 235–267.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Cai, Z, Kuroki, M. (2008). On identifying total effects in the presence of latent variables and selection bias. In *Proceedings of the twenty-fourth conference on uncertainty in artificial intelligence (UAI)*, p 62–69.

Chipman, H. A., George, E. I., McCulloch, R. E., et al. (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics, 4*(1), 266–298.

Clevert, D.A., Unterthiner T., Hochreiter, S. (2015). Fast and accurate deep network learning by exponential linear units (elus). arXiv preprint arXiv:1511.07289.

Eichler, H. G., Bloechl-Daum, B., Bauer, P., et al. (2016). Threshold-crossing: A useful way to establish the counterfactual in clinical trials? *Clinical Pharmacology & Therapeutics, 100*(6), 699–712.

Guo, R. Li, J. Liu, H. (2020). Learning individual causal effects from networked observational data. In *Proceedings of the 13th international conference on web search and data mining (WSDM)*, pp 232–240.

Harada, S. Kashima, H. (2020). Counterfactual propagation for semi-supervised individual treatment effect estimation. In *Joint European conference on machine learning and knowledge discovery in databases*, Springer, pp 542–558.

Harada, S. Kashima, H. (2021). Graphite: Estimating individual effects of graph-structured treatments. In *Proceedings of the 30th ACM international conference on information & knowledge management (CIKM)*, pp 659–668.

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

Johansson, F. Shalit U, Sontag D. (2016). Learning representations for counterfactual inference. In *Proceedings of the 33rd international conference on machine learning (ICML)*, pp 3020–3029.

Kingma, D.P., Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

King, G., & Nielsen, R. (2019). Why propensity scores should not be used for matching. *Political Analysis, 27*(4), 435–454.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review, 76,* 604–620.

Liu, M.Y., Breuel, T., Kautz, J. (2017). Unsupervised image-to-image translation networks. In textit Advances in Neural Information Processing Systems (NeurIPS).

Liu, Q. Allamanis, M. Brockschmidt, M. et al. (2018). Constrained graph variational autoencoders for molecule design. In *Advances in Neural Information Processing Systems (NeurIPS).*

Louizos, C. Shalit U. Mooij J. et al. (2017). Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems (NeurIPS).*

Miao, Y., Yu, L., Blunsom, P. (2016). Neural variational inference for text processing. In *Proceedings of the 33th international conference on machine learning (ICML)*, pp 1727–1736.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41–55.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician, 39*(1), 33–38.

Rothman, K. J., Greenland, S., Lash, T. L., et al. (2008). *Modern epidemiology*. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins.

Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics, 29,* 159–183.

Schwab, P. Linhardt, L. Karlen, W. (2018). Perfect match: A simple method for learning representations for counterfactual inference with neural networks. arXiv preprint arXiv:1810.00656

Sekhon, J. S. (2009). Opiates for the matches: Matching methods for causal inference. *Annual Review of Political Science, 12,* 487–508.

Shalit, U., Johansson, F.D. Sontag, D. (2017). Estimating individual treatment effect: Generalization bounds and algorithms. In *Proceedings of the 34th international conference on machine learning (ICML)*, pp 3076–3085.

Tran, D. Ranganath, R, Blei, D. M. (2015). The variational gaussian process. arXiv preprint arXiv:1511.06499.

Xu, L. Kanagawa, H. Gretton, A. (2021). Deep proxy causal learning and its application to confounded bandit policy evaluation. In *Advances in Neural Information Processing Systems (NeurIPS).*

Yao, L. Li, S. Li, Y. et al. (2018). Representation learning for treatment effect estimation from observational data. In *Advances in Neural Information Processing Systems (NeurIPS).*

Yoon, J. Jordon, J. van der Schaar, M. (2018). Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *Proceedigns of the 6th international conference on learning representations (ICLR).*

Zhang, W. Liu, L. Li, J. (2021). Treatment effect estimation with disentangled latent factors. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)*, pp 10923–10930.

Zhao, S. Heffernan, N. (2017). Estimating individual treatment effect from educational studies with residual counterfactual networks. In *Proceedings of the 10th international conference on educational data mining (EDM)*.

Zhao, S. Song, J. Ermon, S. (2019). Infovae: Balancing learning and inference in variational autoencoders. In Proceedings of the AAAI conference on artificial intelligence (AAAI), pp 5885–5892.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.