IISE Transactions

ISSN: 2472-5854 (Print) 2472-5862 (Online) Journal homepage: www.tandfonline.com/journals/uiie21

Simultaneous treatment effect estimation and variable selection for observational data

Eun-Yeol Ma, Uichin Lee & Heeyoung Kim

To cite this article: Eun-Yeol Ma, Uichin Lee & Heeyoung Kim (2025) Simultaneous treatment effect estimation and variable selection for observational data, IISE Transactions, 57:4, 380-392, DOI: 10.1080/24725854.2024.2330085

To link to this article: https://doi.org/10.1080/24725854.2024.2330085

IISE TRANSACTIONS
2025, VOL. 57, NO. 4, 380–392
https://doi.org/10.1080/24725854.2024.2330085

Simultaneous treatment effect estimation and variable selection for observational data

Eun-Yeol Ma[a], Uichin Lee[b], and Heeyoung Kim[a]

[a]Department of Industrial and Systems Engineering, KAIST, Daejeon, South Korea; [b]School of Computing, KAIST, Daejeon, South Korea

ABSTRACT

Due to the difficulties inherent in conducting controlled experiments, recent causal inference studies have focused on developing treatment effect estimation using observational data. One major difficulty in causal inference from observational data is that the underlying causal structure is unknown. This may result in the misidentification of potential sources of causal estimation bias, such as confounders, which must be controlled for accurate estimation. To consider all possible confounders and other relevant information, conventional methods predominantly use all observed covariates indiscriminately. However, previous studies have shown that including all covariates without considering their causal relationships may deteriorate the estimation performance. Although several data-driven variable selection methods have been proposed for treatment effect estimation, they cannot distinguish the confounders from other outcome-related covariates and are limited to simple regression forms. In this study, we propose a method called the Variable Selection Causal Estimation Network (VSCEN) that performs treatment effect estimation and causal variable selection simultaneously. Through end-to-end differentiable training, the VSCEN selects only the covariates beneficial for effect estimation and uses only those selected for effect estimation. Experimental evaluations on fully synthetic, semi-synthetic, and real datasets demonstrate the VSCEN's superior performance in conditional average treatment effect estimation and competitive performance in average treatment effect estimation along with its accurate variable selection capabilities.

ARTICLE HISTORY
Received 16 July 2023
Accepted 4 March 2024

KEYWORDS
Causal inference; treatment effect estimation; observational data; variable selection; end-to-end training

1. Introduction

Causal inference, in particular, estimation of treatment (causal) effects, has drawn great attention in recent years in various fields, including medicine, econometrics, and social sciences (Imbens and Rubin, 2015). Under the potential outcomes framework (Rubin, 2005), the treatment effect is defined as the difference between the potential outcomes of possible treatment values. Although experimental studies, such as Randomized Controlled Trials (RCTs) are the gold standard for treatment effect estimation, there are many cases where RCTs are often difficult to conduct, because of multiple practical issues (Frieden, 2017; Deaton and Cartwright, 2018; Goldstein *et al.*, 2018). The increasingly available observational data, or data that were gathered without control for any factors, have served as an alternative to RCT data and have been utilized to explore various causal relationships in question. Beyond the classic methods that have focused on population-level or group-level Average Treatment Effect (ATE), the progressively increasing number of studies on observational data has led to advancements in individual-level treatment effect estimation (Guo *et al.*, 2020).

An essential problem in causal inference using observational data is that the true underlying causal relationships between the observed variables are unknown. For instance, one central task in treatment effect estimation is to identify and control for confounders, which are variables causal of both the treatment and the outcome. Uncontrolled confounders induce unwanted associations between treatment and outcome, leading to biased estimation of treatment effects. Although the backdoor criterion (Pearl, 2009) guarantees a minimal sufficient adjustment set, a complete and accurate causal diagram of the data is unavailable in most cases, making it difficult to precisely specify the necessary variables in the analysis. Therefore, an easy option has been to include all the observed covariates in a causal model in a "throw-in-the-kitchen-sink" manner. However, some covariates in fact hinder the estimation process when they are included. For example, including covariates related to neither the treatment nor the outcome (which we call spurious variables) and covariates predictive of only the treatment and not the outcome (instrumental variables) may decrease the estimation efficiency (Schisterman *et al.*, 2009; De Luna *et al.*, 2011; Rotnitzky and Smucler, 2020) while potentially increasing estimation bias (Myers *et al.*, 2011). On the other hand, previous studies have shown that including covariates

CONTACT Heeyoung Kim heeyoungkim@kaist.ac.kr
Supplemental data for this article can be accessed online at https://doi.org/10.1080/24725854.2024.2330085.

associated solely with the outcome and not the treatment (outcome-predictors) and confounders may be beneficial in causal modeling (Brookhart *et al.*, 2006; Rotnitzky and Smucler, 2020).

In addition to improving treatment effect estimation performance, the knowledge regarding the causal relationships between observed covariates and treatment/outcome can increase our understanding of the conclusions given by causal models, which is particularly important in real-world scenarios. For instance, in the field of healthcare, machine learning has experienced significant growth over the past years (Beaulieu-Jones *et al.*, 2019). Although machine learning models have achieved remarkable success in various clinical tasks, the lack of interpretability has raised skepticism among clinicians regarding the clinical conclusions derived from these approaches (Beam and Kohane, 2018). This issue becomes more pronounced when the question at hand requires causal knowledge, as spurious correlations induced by unintended causal relationships, such as confounders, can lead to erroneous conclusions. Therefore, obtaining and incorporating the information about the potential causality of covariates while estimating the treatment effect in question is an advancement towards making interpretable decisions. Furthermore, such knowledge may also help highlight the covariates that should be considered with priority in future validatory experiments, in which their needs were emphasized by multiple previous studies (Gentzel *et al.*, 2019; Gordon *et al.*, 2019; Zhao *et al.*, 2019).

In this study, we propose the Variable Selection Causal Estimation Network (VSCEN), an end-to-end differentiable neural network model that selects the important causal variables directly from the input data for more accurate estimation of the individual-level treatment effect and increased interpretability regarding the causal relationships within the data. We focus on selecting confounders and outcome-predictors, as these are the covariate subsets known to be beneficial in treatment effect estimation (Brookhart *et al.*, 2006; Rotnitzky and Smucler, 2020). The VSCEN selects the confounders and the outcome-predictors through two disjoint Concrete selector layers (Balın *et al.*, 2019), which are then used as inputs for treatment effect estimation through a neural network. Each selection layer is trained in a distinct manner suitable for selecting only the intended causal variables. We tested VSCEN on fully synthetic data, semi-synthetic data, and real observational data to verify the model's treatment effect estimation and variable selection capabilities. The results validate that VSCEN can select the intended causal variables and accurately estimate both individual-level and group-level treatment effects.

2. Related work

2.1. Causal models with variable selection

The full causal structure of data is needed to identify the covariates necessary for treatment effect estimation. However, defining the causal structure of the observed data based on domain knowledge is challenging, even with only a few covariates, as we need to handcraft the underlying causal relationships (Imbens, 2020; Butcher *et al.*, 2021; Kyrimi *et al.*, 2021). Alternatively, previous studies explored data-driven methods to select only the appropriate variables to be included in causal estimation models. Shortreed and Ertefaie (2017) proposed the outcome-adaptive lasso, which uses the adaptive lasso regularization to select only the confounders and the outcome-predictors to be included in the treatment assignment model and estimate the average treatment effect using the obtained treatment model. Similarly, Koch *et al.* (2018) proposed using the group lasso to select only the confounders and outcome-predictors. The utilization of group sparsity for data-driven selection of outcome-predictors and confounders in causal modeling was also used in Greenewald *et al.* (2021). However, while these methods demonstrate the joint selection of confounders and outcome-predictors, they fail to distinguish between the two. Although Kuang *et al.* (2017) proposed an ATE estimation method capable of distinctly selecting confounders and outcome-predictors, the method overlooks the presence of instrumental variables. Furthermore, all of these methods focus on estimating the ATE only and have not been extended to estimate the Conditional Average Treatment Effect (CATE), which quantifies the individual-level treatment effect. On a different note, Makar *et al.* (2019) proposed a method for data-efficient treatment effect estimation based on input variable distillation, in which they discriminate the confounders from other covariates predictive of the treatment effect during the estimation process. Although their work is capable of estimating the CATE, their method relies on regression trees for interpretability, which has limited estimation capabilities.

2.2. CATE estimation using neural networks

Recent advances in deep learning have led to the development of causal inference methods that utilize neural networks for accurate CATE estimation. Johansson *et al.* (2016) proposed using neural networks to estimate the treatment effect by first learning representations that ensure covariate distribution balance between treated samples and controlled samples and then using the representations to predict both potential outcomes. Shalit *et al.* (2017) built upon the work of Johansson *et al.* (2016) and proposed the counterfactual regression, in which the representations were learned by penalizing the integral probability metric between the treated and controlled representations. Shi *et al.* (2019) also proposed a neural network model for causal representation learning, named the Dragonnet, which utilizes propensity score estimation in building the representations. Du *et al.* (2021) integrated adversarial learning and a mutual information constraint to learn balanced representations. Other works that use neural networks include Alaa *et al.* (2017); Louizos *et al.* (2017); Yoon *et al.* (2018); Li and Yao (2022); Zhou *et al.* (2022). However, they focus on the expressivity of neural networks without considering the causal relationships between the covariates and treatment/outcome variables. Hassanpour and Greiner (2019), Zhang *et al.* (2021), and Wu *et al.* (2022) proposed methods to decompose the

input data into latent representations that act as confounders, instrumental variables, and adjustment variables through neural networks for more accurate estimation of treatment effects. However, such latent representations generated via neural networks are unidentifiable in most cases (Roeder *et al.*, 2021; Wu and Fukumizu, 2021). Thus, even if these methods can accurately estimate the CATE, every model specification and initialization may result in drastically different latent representations, which may lead to a misinterpretation of the underlying causality of the observed covariates. Furthermore, while the learned network weights are used to determine the causal contribution of the observed covariates to the latent variables (Hassanpour and Greiner, 2019; Wu *et al.*, 2022), it is difficult to trust that these weights correctly reflect the true effect sizes, since multiple weight matrices may yield the same result (Roeder *et al.*, 2021). Instead, this work aims to clearly highlight the causality of the observed covariates by identifying the confounders and outcome-predictors directly from the input variables, and then using these selected covariates to estimate the CATE, which is the main departure from earlier studies (Johansson *et al.*, 2016; Hassanpour and Greiner, 2019; Zhang *et al.*, 2021; Wu *et al.*, 2022). On a separate note, similar to our work, Chu *et al.* (2020) proposed a method for simultaneous variable selection and treatment effect estimation entitled Feature Selection Representation Matching (FSRM), which employs a sparse one-to-one layer and incorporates sample matching inside a Wasserstein distance-regulated representation space in a network that jointly estimates the potential outcomes and treatment assignment. However, their method relies on the elastic net regularization for variable selection (Zou and Hastie, 2005), limiting its ability to select variables based on the desired causal relationships between the observed covariates.

3. Background

3.1. Treatment effect estimation under the potential outcomes framework

We work under the Neyman–Rubin potential outcomes framework (Rubin, 2005) and assume binary treatment. Let Y_i^a denote the potential outcome of subject i with observed treatment $A = a$, that is, the outcome if the subject were to be given treatment value a. For binary treatment, the factual (observed) outcome can be defined as follows:

$$Y = (1 - a)Y^0 + aY^1. \tag{1}$$

The potential outcome of the unobserved treatment assignment Y^{1-a} is commonly referred to as the counterfactual outcome. Then, we can define the Individual Treatment Effect (ITE) as follows:

$$ITE_i = Y_i^1 - Y_i^0. \tag{2}$$

However, because we do not have access to the true counterfactual outcome, we cannot calculate the ITE directly. Therefore, the individual-level treatment effect for subject i characterized by covariates $\mathbf{x}_i$ is instead estimated through the CATE:

$$CATE(\mathbf{x}_i) = \mathbb{E}(Y_i^1 - Y_i^0 | \mathbf{X}_i = \mathbf{x}_i). \tag{3}$$

The ATE can then estimated via:

$$ATE = \mathbb{E}(CATE(\mathbf{x}_i)). \tag{4}$$

3.2. Causal inference with observational data

We consider the standard assumptions for treatment effect estimation under observational data.

- ***Consistency:*** For observed treatment $A = a$, $Y = Y^a$.
- ***Overlap:*** If $P(\mathbf{X} = \mathbf{x}) > 0$, then $0 < P(A = a|\mathbf{X} = \mathbf{x}) < 1$ for all possible values of a.
- ***Unconfoundedness:*** $Y^a \perp\!\!\!\perp A|\mathbf{X}$, i.e., there are no unobserved confounders.
- ***Stable unit treatment value assumption:*** The potential outcomes of unit i do not vary with respect to the treatment assignment of any $j \neq i$.

These assumptions are sufficient conditions for causal identifiability (Rosenbaum and Rubin, 1983) with observational data.

Further, we assume the causal structure in Figure 1, in which the observed covariates $\mathbf{X}$ can be partitioned into the following variable sets based on their relationship with the treatment A and outcome Y: confounders $\mathbf{X}^C$, outcome-predictors $\mathbf{X}^P$, instrumental variables $\mathbf{X}^I$, and spurious variables $\mathbf{X}^N$. The assumed causal graph is natural for observational data for which we do not know the true underlying causal relationships, as we consider all possible direct causes of the treatment and outcome.

Assuming such a causal structure, Greenewald *et al.* (2021) provide a proof for the c-equivalence (Pearl and Paz, 2014) between the union set of confounders and outcome-predictors $\mathbf{X}^C \cup \mathbf{X}^P$ and the observed covariate set $\mathbf{X}$. In other words, it is sufficient to control for only the covariate set $\mathbf{X}^C \cup \mathbf{X}^P$ from the full observed set $\mathbf{X}$ to make an unbiased estimation of the treatment effect.

3.3. End-to-end variable selection using concrete random-variable sampling

The proposed model incorporates the Concrete selector layer (Balın *et al.*, 2019), which selects input variables through Concrete (Gumbel-softmax) random variable sampling (Jang *et al.*, 2017; Maddison *et al.*, 2017), for end-to-end selection of the intended causal variables. Concrete random variables

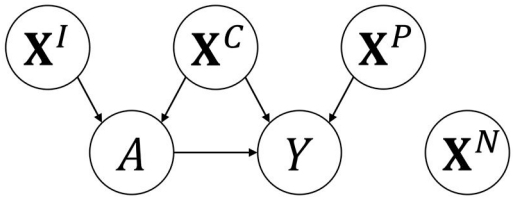

Figure 1. Partitioning of observed covariates $\mathbf{X}$ based on its causal relationship to A and Y. Each subset of $\mathbf{X}$ must be discovered from the data. (A: treatment assignment, Y: outcome, $\mathbf{X}^I$: instrumental variables, $\mathbf{X}^C$: confounders, $\mathbf{X}^P$: outcome-predictors, $\mathbf{X}^N$: spurious variables.)

are continuous relaxations of discrete random variables, in which a Concrete vector $\mathbf{m}$ is defined as:

$$\mathbf{m} \sim Concrete(\boldsymbol{\alpha}, T), \mathbf{m}_j = \frac{\exp\left((\log \boldsymbol{\alpha}_j + \mathbf{g}_j)/T\right)}{\sum_{d=1}^{D} \exp\left((\log \boldsymbol{\alpha}_d + \mathbf{g}_d)/T\right)}, \tag{5}$$

where $\mathbf{g}$ is a D-dimensional vector of i.i.d. Gumbel samples. T is the temperature parameter determining the discreteness of the Concrete random variable. As $T \to 0$, $\mathbf{m}$ approaches a one-hot vector with $P(\mathbf{m}_j = 1) = \boldsymbol{\alpha}_j / \sum_d \boldsymbol{\alpha}_d$. Concrete random variables can be learned within an end-to-end differentiable framework through the reparameterization trick (Kingma and Welling, 2013), by learning the location parameter $\boldsymbol{\alpha}$ and sampling $\mathbf{g}$.

Variable selection in neural networks using the Concrete selector layer is performed by first sampling a Concrete random variable $\mathbf{m}^{(k)}$, and then taking its dot product with the input variables to obtain $\mathbf{u}^{(k)} = \mathbf{X} \cdot \mathbf{m}^{(k)}$. These variables are then fed into further neural network layers for a downstream task (Doo and Kim, 2023), which is treatment effect estimation in our case. As $T \to 0$, $\mathbf{u}^{(k)}$ becomes a single variable from the input variables as determined by the nonzero element of the one-hot vector $\mathbf{m}$. We apply the exponential temperature annealing schedule presented in Balın *et al.* (2019) with the modifications described in Section 4. The annealing schedule sets T at epoch b as $T_0(T_B/T_0)^{b/B}$, where T_0 and T_B are the initial and final temperatures, respectively, and B is the total number of training epochs. By applying such an annealing schedule, the selector layers are trained to explore various variable combinations in the beginning but later exploit the selection of the variables most effective in minimizing the loss.

4. Proposed model

Our goal is to estimate the CATE of the supposed treatment by predicting each counterfactual outcome using only observational data. Because the exact causal relationships among the observed variables are unknown in observational data, we aim to select only the covariates that are helpful in accurate estimation of potential outcomes, namely the confounders and the outcome-predictors, and use only those covariates for treatment effect estimation. We propose to achieve both tasks (treatment effect estimation and causal variable selection) simultaneously through an end-to-end differentiable training method using neural networks. The VSCEN distinctively selects confounders and outcome-predictors through its respective Concrete selector layers specifically designed to select only the intended covariate sets. This design ensures a clear and straightforward interpretation of the selected covariates. Furthermore, this separate selection process ensures the use of both intended covariate sets in the treatment effect estimation.

4.1. Overall architecture

Figure 2 shows the overall architecture of the VSCEN. Let $\mathbf{X}_i \in \mathbb{R}^D$ be the input covariates. For notational simplicity, we hereafter omit the subscript i unless necessary. The VSCEN first takes the input covariates and selects K_C potential confounders and K_P potential outcome-predictors through two separate Concrete selector layers. Each variable selected by the Concrete selector layer $\tilde{\mathbf{u}}^{(k_1)}, k_1 = 1, ..., K_C$, and $\tilde{\mathbf{v}}^{(k_2)}, k_2 = 1, ..., K_P$, is initially a weighted linear combination of $\mathbf{x}^1, ..., \mathbf{x}^D$, and approaches a one-hot vector as training progresses. Then, at test time, the arguments of the maxima are taken to obtain $\mathbf{U}$, which consists of the observed covariates assumed to be confounders $\mathbf{X}^C$, and $\mathbf{V}$, which consists of the observed covariates assumed to be outcome-predictors $\mathbf{X}^P$:

$$\mathbf{U} = \left[\mathbf{u}^{(1)}, ..., \mathbf{u}^{(K_C)}\right], \mathbf{u}^{(k)} = \mathbf{X}^{\mathrm{argmax}_j \tilde{\mathbf{u}}_j^{(k)}}, \tag{6}$$

$$\mathbf{V} = \left[\mathbf{v}^{(1)}, ..., \mathbf{v}^{(K_P)}\right], \mathbf{v}^{(k)} = \mathbf{X}^{\mathrm{argmax}_j \tilde{\mathbf{v}}_j^{(k)}}. \tag{7}$$

On one side of the architecture, the selected variables $\mathbf{U}$ and $\mathbf{V}$ are concatenated and fed into the further layers to

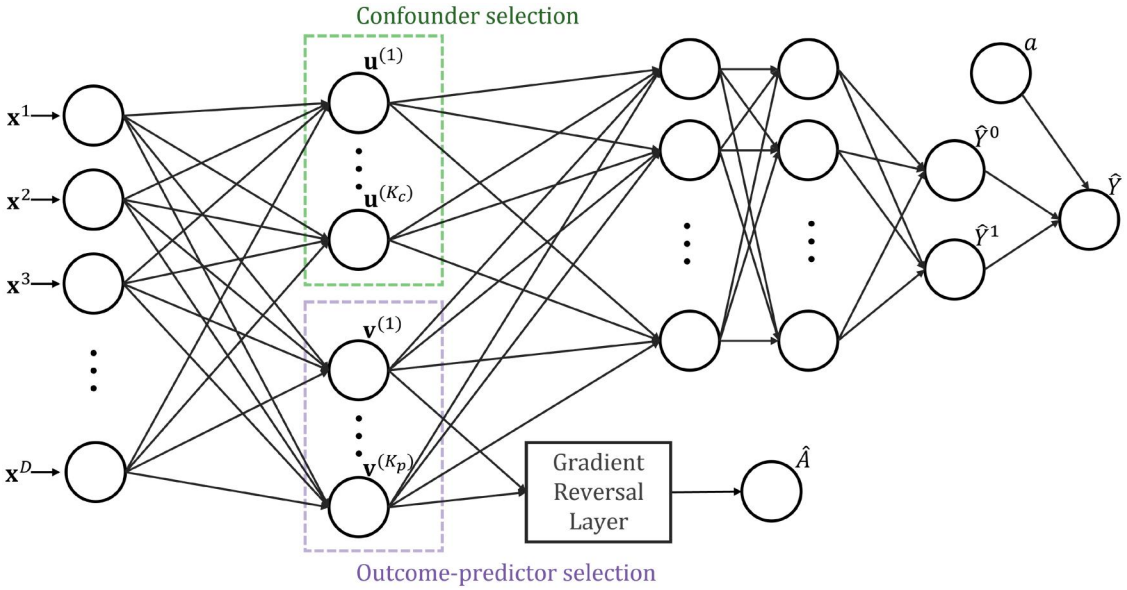

Figure 2. Network structure of the proposed VSCEN.

obtain the estimated potential outcomes $\hat{Y}^{(1)}$ and $\hat{Y}^{(0)}$. The estimated potential outcomes are then used to calculate the estimated factual outcome $\hat{Y}$ using the observed binary treatment assignment a through (1). On the other side of the architecture, $\mathbf{V}$ goes through the gradient reversal layer (Ganin and Lempitsky, 2015) in predicting the treatment assignment $\hat{A}$.

4.2. Training the outcome-predictor selector layer

Each selector layer is trained in a distinct manner such that one layer prefers the selection of confounders $\mathbf{X}^C$ and the other layer prefers the selection of outcome-predictors $\mathbf{X}^P$. This means that the instrumental variables $\mathbf{X}^I$ and spurious variables $\mathbf{X}^N$ will not be selected by these selector layers. To train the outcome-predictor selector layer to select outcome-predictors as $\mathbf{V}$ as intended, we leverage the fact that among confounders, outcome-predictors, and instrumental variables, only the outcome-predictors are not predictive of the treatment assignment (see Figure 1); the other two subsets of covariates (i.e., confounders and instrumental variables) are directly causal of the treatment assignment, as indicated by the edges going into A from $\mathbf{X}^C$ and $\mathbf{X}^I$. This observation hints that the gradient reversal layer (Ganin and Lempitsky, 2015) can be placed on the network head that predicts A using only the supposed outcome-predictors to enforce the selector layer to select variables that are least predictive of the treatment assignment while most predictive of the outcome. That is, for every epoch, the network is optimized in the direction opposite of minimizing the treatment assignment prediction error, resulting in the exclusion of instrumental variables and confounders for $\mathbf{V}$. Because we use $\mathbf{V}$ in estimating an outcome, spurious variables that are causal of neither the outcome nor the treatment are naturally excluded, resulting in the selection of only the outcome-predictors as $\mathbf{V}$.

4.3. Training the confounder selector layer

The confounder selector layer is trained by discriminating the confounders from $\mathbf{X}^C \cup \mathbf{X}^P$ as the training progresses. As the network is trained to estimate the outcome, the confounder selector layer is likely to select either confounders or outcome-predictors, although it has no knowledge of whether the selected covariate is a confounder or an outcome-predictor. Assuming that the network has already trained to select outcome-related covariates with some training, we then force the confounder selector layer to prefer the selection of confounders over outcome-predictors as $\mathbf{U}$ through the use of *covariate-treatment correlation*. As shown in Figure 1, there exists an edge from the confounders $\mathbf{X}^C$ to treatment A, whereas no edge exists between outcome-predictors $\mathbf{X}^P$ and A. Hence, it is likely that the confounders are more strongly associated with the treatment than the outcome-predictors, making the covariate-treatment correlation an obvious criterion that distinguishes the two.

Assuming we train the model for a total of B epochs, we first train the confounder selector layer with the standard exponential temperature annealing schedule for B' epochs (e.g., $B' = B/2$). We then incorporate the covariate-treatment correlation into the Concrete random variable sampling scheme at epoch $b \geq B'$ as follows:

$$\mathbf{m}_j(b) = \frac{\exp\left((\log \boldsymbol{\alpha}_j \times \exp\left(\beta|\boldsymbol{\rho}_j|\right) + \mathbf{g}_j)/T(b)\right)}{\sum_{d=1}^{D} \exp\left((\log \boldsymbol{\alpha}_d \times \exp\left(\beta|\boldsymbol{\rho}_d|\right) + \mathbf{g}_d)/T(b)\right)},$$

where

$$\boldsymbol{\rho}_k = \begin{cases} Corr(\mathbf{X}^{(k)}, A) & \text{if } \mathbf{X}^{(k)} \in \hat{\mathbf{X}}^C \text{at epoch } (b-1), \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

The hyperparameter β controls the extent to which the correlation is considered. Such adjustments inflate the location parameter $\boldsymbol{\alpha}$ for only the covariates correlated with treatment assignment. Therefore, with the confounder selector already trained to select some covariates predictive of the outcome, the incorporation of covariate-treatment correlation ρ encourages the confounder selector layer to select confounders over outcome-predictors. This additional inclination towards confounders using ρ is important for confounder selection regardless of what variables are selected by the outcome-predictor selector layer. In essence, ρ reduces the chances of only outcome-predictors being selected by both selector layers when selection is solely based on prediction. The incorporation of ρ can also improve training efficiency by guiding the confounder selector to restrict searches on irrelevant variables. Note that while increasing the number of training epochs may help the confounder selector prefer the desired covariates through enhanced prediction in general, a mere increase in training epochs may not be enough, as there always remains the possibility of selecting outcome-predictors over confounders when selection is solely driven by prediction. Although instrumental variables may also be associated with the outcome as they are connected in the assumed causal structure (Figure 1), we expect their effects to be negligible in predicting the outcome since the observed treatment value $A = a$ is included in the estimation process. Hence, it is unlikely that the confounder selector layer would select instrumental variables over confounders as $\mathbf{U}$. Similar to the outcome-predictor selector layer, spurious variables are naturally excluded because they are not causal of the outcome.

4.4. End-to-end training of VSCEN

The VSCEN is trained by minimizing the following loss:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_{out}(y_i, f(\mathbf{U}, \mathbf{V}, a)) + \gamma \mathcal{L}_{treat}(a_i, h(\mathbf{V})), \tag{9}$$

where $\mathcal{L}_{out}$ is set as the mean squared error loss of the factual outcome estimation and $\mathcal{L}_{treat}$ is set as the binary cross-entropy loss of the treatment assignment prediction. Here, f and h are neural network functions that estimate the outcome and treatment, respectively. Note that $\mathcal{L}_{out}$ is calculated using both selected covariate sets, whereas $\mathcal{L}_{treat}$ is calculated using only the selected outcome-predictor set. At every epoch of training, the network trains to minimize the

total loss $\mathcal{L}$ and is optimized in the direction of minimizing the factual outcome estimation error while maximizing the treatment assignment prediction error, resulting in the desired selection of confounders and outcome-predictors. Training of all network parameters, including both the neural network parameters and variable selection location parameters, is performed through standard backpropagation using gradient descent, allowing fully differentiable end-to-end training for simultaneous estimation of the treatment effect and selection of causal variables. The training algorithm is detailed in Section 1 in the Supplementary Materials.

5. Experiments

To test the abilities for treatment effect estimation and causal variable selection, we conducted experiments on three types of data: fully synthetic data, semi-synthetic data, and real data. For each dataset, we aim to answer the following evaluation questions:

- *Fully synthetic data*: How precisely can VSCEN estimate the individual and population-level treatment effect? How correctly can VSCEN identify the true confounders and outcome-predictors?
- *Semi-synthetic data*: How precisely can VSCEN estimate the treatment effect on popular benchmark data compared to other methods? How correctly can VSCEN identify the true outcome-generating covariates?
- *Real data*: Can VSCEN suggest reasonable effect size and confounders/outcome-predictors on real-world data?

We relied on fully synthetic and semi-synthetic data due to the innate difficulties of evaluating causal estimation models on real-world observational data, in which neither the true treatment effect nor the true causality between observed variables is known for real data (Holland, 1986; Shalit *et al.*, 2017). We also conducted experiments on real data to investigate the model's ability to discover meaningful causal covariates while estimating treatment effects.

We compared our method with previous methods for treatment effect estimation, including linear regression with Lasso Regularization (LR lasso), Causal Forest (CF) (Wager and Athey, 2018), Orthogonal Random Forest (ORF) (Oprescu *et al.*, 2019), Outcome-Adaptive Lasso (OAL)[1] (Shortreed and Ertefaie, 2017), Group Lasso and Doubly robust estimation (GLiDer) (Koch *et al.*, 2018), Data-driven Variable Decomposition (D2VD)[2] (Kuang *et al.*, 2017, Kuang *et al.*, 2022), Treatment-Agnostic Representation network (TARnet), CounterFactual Regression network (CFR) (Shalit *et al.*, 2017), DragonNet (Shi *et al.*, 2019), Treatment Effect by Disentangled Variational AutoEncoder (TEDVAE) (Zhang *et al.*, 2021), Adversarial Balancing-based representation learning for Causal Effect Inference (ABCEI) (Du *et al.*, 2021), and Feature Selection Representation Matching (FSRM) (Chu *et al.*, 2020). Details on hyperparameter ranges and implementation are provided in Section 2 in the Supplementary Materials.

For the fully synthetic and semi-synthetic data, we evaluated the methods primarily on two treatment effect estimation metrics:

- Root Precision in the Estimation of Heterogeneous Effects (PEHE)

$$\epsilon_{\sqrt{PEHE}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}((\hat{Y}^{(1)} - \hat{Y}^{(0)}) - (Y^{(1)} - Y^{(0)}))^2}, \quad (10)$$

- Absolute ATE error

$$\epsilon_{ATE} = \left|\widehat{ATE} - ATE\right|. \quad (11)$$

where $\hat{Y}^{(a)}$ is the estimated potential outcome for treatment value a and $\widehat{ATE}$ is the estimated ATE. The PEHE measures the individual-level effect estimation error, and the ATE error measures the population-level effect estimation error.

Furthermore, the variable selection results are presented through the False Discovery Rate (FDR), which measures the proportion of falsely selected variables:

$$FDR = \mathbb{E}\left[\frac{|\hat{S} \cap S_0^c|}{|\hat{S}|}\right], \quad (12)$$

where $\hat{S}$ is the set of selected variables, S_0 is the true set, and $|\cdot|$ is the cardinality of the set. For the fully synthetic and semi-synthetic data, we report the results for both in-samples (training samples) and out-samples (test samples). Note that in-sample estimation is still a meaningful task for causal inference because the counterfactual outcomes remain unknown to the model during training. For the real data, we report the estimated treatment effect size and qualitatively assess the selected covariates based on well-known domain knowledge. However, the true treatment effect in real data is unknown because the counterfactual outcomes are unavailable.

5.1. Fully synthetic data-based evaluation

First, we conducted experiments on a fully synthetic dataset, in which the full data generation scheme including all counterfactual outcomes is known completely. This allows us to evaluate the methods for treatment effect estimation since the ground-truth effect sizes are known, both individual-wise and population-wise. It also allows us to evaluate the methods for variable selection, because the true confounder and outcome-predictors in the data-generating process are known.

5.1.1. Data generation

We model after Zigler and Dominici (2014) and Shortreed and Ertefaie (2017) for the fully synthetic data generation. For each sample, we sampled the covariates $\mathbf{X}$ from a 100-dimensional multivariate standard Gaussian distribution. For interpretational simplicity, we fixed covariates $\{\mathbf{X}^1, ..., \mathbf{X}^5\}$ as

[1]Because OAL was originally presented with an ATE estimator and not an outcome-model, we used the OAL to calculate the Inverse Probability of Treatment Weights (IPTW) but fit a separate linear outcome model.
[2]We only present the ATE error for D2VD as it is unsuitable for CATE estimation.

the set of confounders, $\{\mathbf{X}^6, ..., \mathbf{X}^{10}\}$ as the set of outcome-predictors, and $\{\mathbf{X}^{11}, ..., \mathbf{X}^{15}\}$ as the set of treatment-predictors. The remaining covariates $\{\mathbf{X}^{16}, ..., \mathbf{X}^{100}\}$ are spurious variables. Note that the results are general regardless of the fixed covariate indices, since all covariates are generated via the standard Gaussian distribution.

We next set two coefficient vectors $\mathbf{c}_o$ and $\mathbf{c}_t$ for outcome generation and treatment assignment, respectively, in which their elements were set as 0.5 if the corresponding covariate is causal and zero otherwise (i.e., $\mathbf{c}_o^1 = ... = \mathbf{c}_o^{10} = 0.5, \mathbf{c}_o^{11} = ... = \mathbf{c}_o^{100} = 0$ and $\mathbf{c}_t^1 = ... = \mathbf{c}_t^5 = \mathbf{c}_t^{11} = ... = \mathbf{c}_t^{15} = 0.5$, $\mathbf{c}_t^6, ..., \mathbf{c}_t^{10} = \mathbf{c}_t^{16}, ..., \mathbf{c}_t^{100} = 0$). We sampled binary treatment assignment from a Bernoulli distribution with $p = \sigma(\sum_{d=1}^{D} \mathbf{c}_t \mathbf{X})$, where σ is the sigmoid function. Outcomes were generated using a linear model $Y^a = \sum_d \mathbf{c}_o^d \mathbf{X}^d + 4a + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 1)$. As seen in the outcome generation model, the treatment effect size was set to four. For the experiments on the fully synthetic dataset, we report the mean (standard error) of the mentioned performance metrics over 100 replications of data generation, each with $N = 2000$. We report the mean (standard error) FDR for only the VSCEN and D2VD, as the other methods cannot distinguish the selected covariates and thus cannot calculate the FDR of confounders and outcome-predictors separately.

5.1.2. Exploring the behaviors of VSCEN

We begin our analysis of the VSCEN by conducting an ablation study. In this study, we conducted experiments on the fully synthetic dataset, disabling specific components of the VSCEN to clearly highlight its behaviors. Given the known ground-truth data generation process, this ablation study enables us to understand the functionality of each suggested component. Specifically, we tested the impact of each component introduced for variable selection on the estimation and variable selection performance. To achieve this, we performed experiments using the following four models:

1. Full VSCEN with all proposed components.
2. VSCEN without the use of the gradient reversal layer in outcome-predictor selection (i.e., utilizing only covariate-treatment correlation in Concrete random sampling for confounder selection).
3. VSCEN without the use of covariate-treatment correlation (i.e., employing only the gradient reversal layer).
4. VSCEN without both. These two components were incorporated into the VSCEN to train the selector layers, guiding them to select the desired covariates of outcome-predictors and confounders.

Figure 3 summarizes the results from the ablation study, with the evaluation metrics (PEHE, ATE error, FDR for confounder selection, and FDR for outcome-predictor selection) shown for all four models tested. We present the average performance and the corresponding standard error over 100 replications. As shown in Figure 3(a)-(d), the estimation error was the lowest for the VSCEN with all model components (i.e., gradient reversal layer and the covariate-treatment correlation). Furthermore, removing the gradient reversal layer for outcome-predictor selection decreased the ability of both variable selectors and resulted in the worst FDRs. Similarly, without using the covariate-treatment correlation in selecting the confounders, the treatment effect estimation performance decreased, and higher PEHE and ATE error resulted.

Figure 3(e) and (f) shows the proportion of specific covariates selected by the models considered in the ablation study as confounders and outcome-predictors, respectively. Examining the covariates selected by the models in depth, the models without the use of the gradient reversal layer for outcome-predictors selection (cyan: w/o GRL & Corr., red: w/o GRL) often resulted in mixing up the confounders and the outcome-predictors, as they are both predictive of the factual outcome. The high confusion in variable selection was still evident for the model without the gradient reversal layer but with the use of the covariate-treatment correlation (red: w/o GRL), although the difference in effect estimation performance was less pronounced than the difference in variable selection performance (Figure 3). We speculate that this model displayed good estimation in spite of poor variable selection because it actually uses the same covariates as the full VSCEN in place to estimate the outcome, as confounders are identified as outcome-predictors and *vice versa*. Furthermore, the models without the gradient reversal layer occasionally selected treatment-predictors as outcome-predictors, which rarely happened for the models with the gradient reversal layer used (orange: w/o Corr., purple: Full VSCEN). Therefore, we conclude that both components used to train the selector layers help not only variable selection but also estimation.

5.1.3. Comparative results

Table 1 shows the estimation performance on the synthetic dataset of all models considered. The VSCEN had the lowest root PEHE among all contending models, showing its ability to accurately estimate the CATE. In addition, the VSCEN performed comparably with other methods in terms of ATE estimation. Compared to the D2VD (Kuang *et al.*, 2017), which was the only baseline method capable of distinctively selecting confounders and outcome-predictors, the VSCEN exhibited superior variable selection performance for both covariate sets. In particular, the VSCEN outperformed the D2VD significantly in confounder selection. This is expected, as the assumed causal structure for the D2VD does not consider instrumental variables, unlike the VSCEN.

We also examined which covariates were selected as confounders and outcome-predictors using the VSCEN (Figure 4). The left green bars represent the proportion of covariates selected as confounders and the right blue bars represent the proportion of covariates selected as outcome-predictors. As intended, $\mathbf{X}^1, ..., \mathbf{X}^5$ are generally selected as confounders, and $\mathbf{X}^6, ..., \mathbf{X}^{10}$ are generally selected as outcome-predictors. As seen in Figure 4, there were occasions when the confounders and outcome-predictors were

Figure 3. Treatment effect estimation error and variable selection performance of the models in the ablation study. (a) PEHE, (b) ATE error, (c) FDR for confounder selection, (d) FDR for outcome-predictor selection, (e) Proportion of times each covariate was selected as a confounder by the models in the ablation study, (f) Proportion of times each covariate was selected as an outcome-predictor by the models in the ablation study (GRL: gradient reversal layer, Corr: covariate-treatment correlation).

confused with each other, as both selections were trained to estimate the factual outcome well. However, instrumental variables (treatment-predictors), which we want to exclude deliberately, were not selected by the model. Therefore, we can see that each of the selectors was mostly able to select only the intended covariates.

5.2. Semi-synthetic data-based evaluation

Similar to experiments on the fully synthetic data, we conducted experiments on a popular benchmark dataset used to evaluate treatment effect estimation models. In specific, we used the Infant Health and Development Program (IHDP)

Table 1. Results on fully synthetic data.

	In-sample		*Out-sample*		*Variable selection*	
	$\epsilon_{\sqrt{PEHE}}$	ϵ_{ATE}	$\epsilon_{\sqrt{PEHE}}$	ϵ_{ATE}	FDR_C	FDR_P
LR lasso	1.95 (0.02)	1.33 (0.08)	1.96 (0.08)	1.34 (0.10)		
CF	1.03 (0.04)	0.20 (0.08)	1.07 (0.06)	0.22 (0.08)		
ORF	1.01 (0.04)	0.24 (0.09)	0.81 (0.05)	0.26 (0.10)		
OAL	1.18 (0.01)	**0.05 (0.04)**	1.19 (0.03)	0.84 (0.19)		
GLiDer	1.73 (0.01)	2.66 (0.03)	1.74 (0.02)	2.67 (0.08)		
D2VD		0.25 (0.07)		0.25 (0.07)	0.94 (0.12)	0.32 (0.15)
TARnet	0.74 (0.06)	0.16 (0.08)	0.76 (0.07)	0.17 (0.09)		
CFR	0.73 (0.07)	0.16 (0.08)	0.73 (0.07)	0.16 (0.08)		
Dragonnet	0.73 (0.06)	0.15 (0.08)	0.76 (0.06)	0.16 (0.09)		
TEDVAE	1.37 (0.15)	0.15 (0.09)	1.37 (0.15)	0.15 (0.10)		
ABCEI	1.41 (0.03)	0.11 (0.06)	1.48 (0.09)	**0.14 (0.10)**		
FSRM	0.55 (0.07)	0.27 (0.10)	0.53 (0.07)	0.25 (0.11)		
VSCEN	**0.36 (0.08)**	0.15 (0.11)	**0.36 (0.08)**	0.16 (0.11)	**0.19 (0.13)**	**0.24 (0.14)**

*FDR_C : FDR for confounder selection, FDR_P : FDR for outcome-predictor selection. A lower value is better for all metrics presented.

Table 2. Results on semi-synthetic (IHDP) data.

	In-sample		*Out-sample*		
	$\epsilon_{\sqrt{PEHE}}$	ϵ_{ATE}	$\epsilon_{\sqrt{PEHE}}$	ϵ_{ATE}	$FDR_{C\cup P}$
LR lasso	2.15 (0.12)	3.36 (0.29)	2.15 (0.13)	3.37 (0.41)	0.82 (0.23)
CF	1.23 (0.22)	0.33 (0.21)	1.30 (0.25)	0.36 (0.27)	
ORF	1.04 (0.25)	0.39 (0.19)	1.06 (0.31)	0.40 (0.27)	
OAL	1.47 (0.25)	0.25 (0.21)	1.47 (0.26)	0.50 (0.43)	0.61 (0.14)
GLiDer	1.88 (0.15)	2.77 (0.29)	1.88 (0.16)	2.78 (0.41)	0.60 (0.19)
D2VD		0.22 (0.16)		0.25 (0.20)	0.70 (0.16)
TARnet	1.09 (0.19)	0.16 (0.13)	1.17 (0.23)	0.19 (0.15)	
CFR	0.98 (0.17)	0.16 (0.13)	1.07 (0.19)	0.19 (0.14)	
Dragonnet	0.82 (0.11)	0.16 (0.13)	0.87 (0.15)	0.19 (0.14)	
TEDVAE	0.66 (0.13)	0.14 (0.11)	0.66 (0.16)	**0.15 (0.12)**	
ABCEI	1.58 (0.11)	0.22 (0.11)	1.61 (0.17)	0.23 (0.15)	
FSRM	0.92 (0.32)	0.40 (0.33)	0.99 (0.33)	0.42 (0.34)	0.60 (0.31)
VSCEN	**0.60 (0.20)**	**0.14 (0.10)**	**0.58 (0.19)**	0.16 (0.12)	**0.10 (0.10)**

*$FDR_{C\cup P}$: FDR for the union of confounders and outcome-predictors.

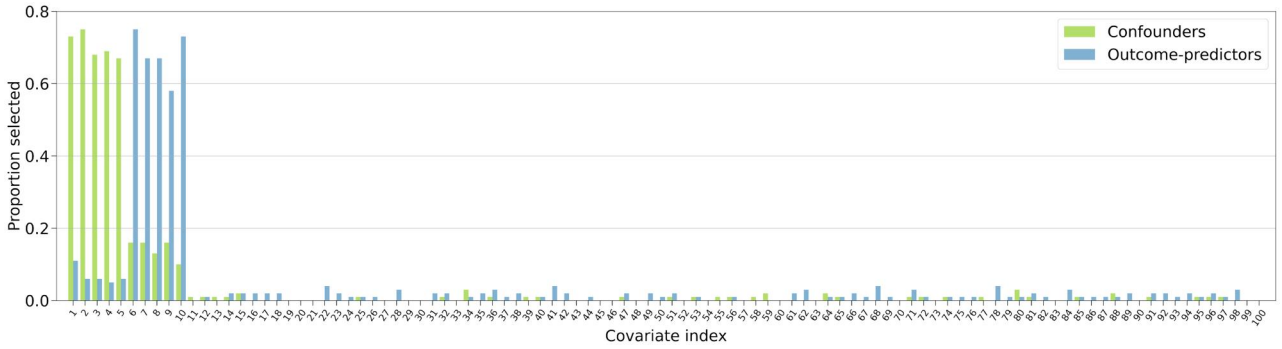

Figure 4. Proportion of times each covariate was selected as a confounder or an outcome predictor over 100 replications of the fully synthetic data.

dataset (Hill, 2011). The IHDP is a suitable benchmark dataset for our purpose because we can identify both potential outcomes and which covariates were used to generate outcome.

5.2.1. Data generation

The IHDP dataset is originally from a real RCT on the effects of home visits by trained experts on child development. Hill (2011) derandomized this dataset by removing treated samples with non-white mothers, resulting in a dataset composed of 747 samples (139 treated and 608 control), with 25 real covariates regarding the child or the mother, such as the child's birth weight and the mother's age. In addition, using these real covariates and treatment values, Hill (2011) devised synthetic outcome generation schemes in which the outcome is generated using a randomly sampled subset of the observed covariates, allowing identification of the true outcome-generating covariates at each replication of outcome generation. We generated the outcomes using the nonlinear generation procedure of Hill (2011).[3] Such a

[3]We generated our own version of the IHDP data in order to identify the outcome-generating (causal) covariates. This is because the publicly available version provided in Johansson *et al.* (2016) does not specify the outcome-generating covariates.

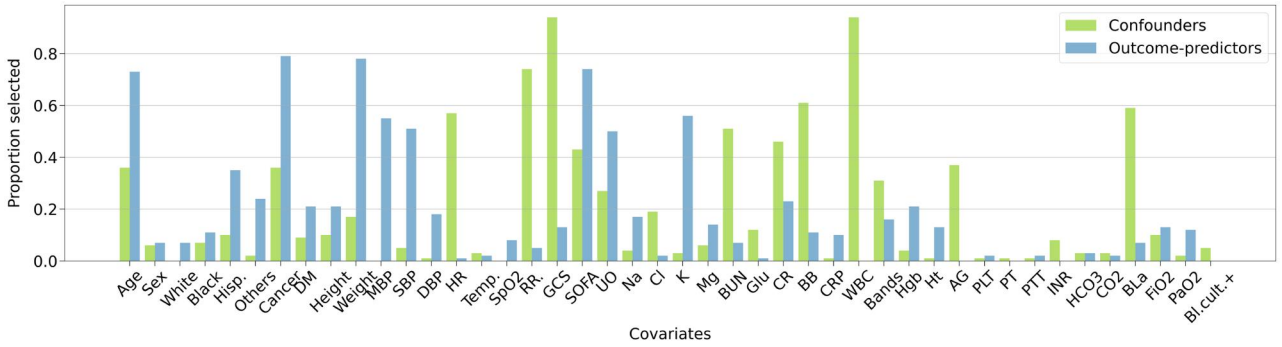

Figure 5. Proportion of selected sepsis-related covariates.

semi-synthetic dataset allows the evaluation of treatment effect estimation models based on the assumed treatment effect size. The IHDP dataset is a well-known benchmark dataset used to evaluate treatment effect estimation models.

We again report the mean (standard error) of the performance metrics over 100 replications of outcome generation using 25 real input covariate values and treatment assignment for 747 samples. Because the treatment assignment is taken from the original experimental data, we cannot distinguish the confounders from the outcome-predictors. Therefore, we report the mean (standard error) FDR for the selection of outcome-generating variables as either confounders or outcome-predictors ($FDR_{C\cup P}$) for the applicable models.

5.2.2. Comparative results

Table 2 shows the estimation performance on the IHDP data. Similar to the fully synthetic data experiments, the VSCEN outperformed all contending methods in terms of root PEHE. Compared with the shallow models (LR Lasso, CF, ORF, OAL, GLiDer), the VSCEN achieved an improvement of at least 40%. Compared with TARNet, CFR, and Dragonnet, which are neural network methods that indiscriminately use all input variables, the VSCEN achieved improvements of approximately 20% to 30%. The VSCEN also outperformed the FSRM, which is a neural network method that incorporates variable selection. The ATE estimation performance was also the best among the contending models. The results provide evidence that incorporating the causal structure regarding the outcome and the treatment may improve estimation performance. Specifically, we find that recovering only the confounders and the outcome-predictors from the data is sufficient for treatment effect estimation, which is the key task of this work. Furthermore, the VSCEN accurately selected the covariates used for outcome generation as either confounders or outcome-predictors whereas the previous variable selection models, both shallow and deep, performed poorly (Table 2). This indicates that the VSCEN is useful in distinguishing the variables that also need to be collected for future effect prediction.

5.3. Real data-based evaluation

Finally, we conducted experiments on real patient data to verify the effectiveness of the model in identifying true causal covariates in real causal situations.

We used the data from MIMIC-III (Johnson *et al.*, 2016), which is a clinical database containing comprehensive deidentified data of approximately 50,000 Intensive Care Unit (ICU) patients from the Beth Israel Deaconess Medical Center in Boston, Massachusetts, to investigate the effect of antibiotics, which is a common treatment option for sepsis, on the 30-day mortality of septic patients.[4] We extracted 41 sepsis-related variables for 2773 patients (see Section 3 in the Supplementary Materials for details) as the covariates, whether the patient passed away within 30-days of ICU stay as the observed outcome, and whether the patient was given antibiotics as the binary treatment. The softmax probabilities of the 30-day mortality for treated and controlled ($P(Y=1|A=a)$ for $a\in\{0,1\}$) were considered to be the estimated potential outcomes. We used the mean values of the covariates before the patient was given the treatment in consideration as the input covariate values.

We report the mean (standard error) estimated effect size and the p-value under the null hypothesis that the average treatment effect is zero over 100 repeated experiments on the same dataset with different train-test splits. The VSCEN estimated a statistically significant (p-value < 0.001) negative treatment effect of size 0.03 as the treatment effect of antibiotics. In other words, the use of antibiotics on sepsis patients decreased the risk of mortality and was indeed causal to a patient's survival. Although we cannot know how accurate the estimated effect size is as the ground-truth effect size is unknown, such a negative treatment effect was similarly suggested by other baseline models as well (estimated effect sizes by other baseline models are given in Section 3 in the Supplementary Materials).

We examine in depth the covariates selected by the VSCEN as confounders and outcome-predictors when the use of antibiotics is the treatment of interest. Figure 5 shows the proportion of times each covariate was selected as either confounder (left green bars) or outcome-predictor (right blue bars) over 100 repeated experiments. The most dominant confounder selected was the white blood cell count, which is one of the major symptoms of an infection. Thus, white blood cell count is a clear cause of antibiotic usage while also being a cause of increased mortality rate. Another dominant confounder identified by the VSCEN was the

[4]Sepsis is an illness due to physiologic, pathologic, and biochemical abnormalities caused by infection (Singer *et al.*, 2016).

Glasgow Coma Scale (GCS), which measures the consciousness of patients. Indeed, GCS is another symptom that is commonly observed for patients with infection and is concurrently highly predictive of mortality (Udekwu *et al.*, 2004). Other covariates that were often selected as confounders were vital signs, such as respiratory rate and heart rate, which are common predictors of mortality while being indicative of an infection. On the other hand, covariates such as age, weight, and comorbid conditions (e.g., cancer) may not be associated with antibiotic usage, but they are typical predictors of patient mortality. The VSCEN appropriately selected these covariates as outcome-predictors instead of confounders. In conclusion, the VSCEN was able to simultaneously estimate treatment effects while selecting clinically meaningful confounding and outcome-predicting causal covariates.

6. Conclusion

In this study, we proposed the VSCEN for simultaneous treatment effect estimation and causal variable selection. Specifically, we incorporate the Concrete random sampling layers in a neural network for potential outcomes estimation to select confounders and outcome-predictors, allowing both estimation and selection to be performed in an end-to-end differentiable manner. Achieving both tasks simultaneously allows us to highlight the more important covariates for the treatment effect while quantifying the treatment effect size of interest.

We conducted experiments on various datasets to evaluate the proposed method in terms of both treatment effect estimation and causal variable selection. First, through experiments on fully synthetic data for which we know the complete underlying causal structure, we showed that the VSCEN is able to accurately estimate treatment effects while selecting the intended covariate subsets. Second, through experiments on the IHDP dataset, which is a commonly used benchmark dataset, we demonstrated the effective performance of the VSCEN for both treatment effect estimation and variable selection in comparison to contending models. Finally, through experiments on the MIMIC-III sepsis data, we displayed the ability of the VSCEN to select causally meaningful covariates in real clinical data while quantifying the treatment effect size.

Despite the success of the VSCEN in simultaneous treatment effect estimation and causal variable selection on both synthetic and real datasets, there are a few limitations that must be noted. First, the proposed model and its results are only applicable if the causal structure assumed in this study (Figure 1) is true. Hence, the results are not universal and cannot handle other causal structures (e.g., when there is selection bias due to a common child of the treatment and outcome). However, we believe this causal structure is practical enough for most real observational data, because all direct causes of the treatment and outcome are taken into account. In addition, because the confounder selection utilizes the correlation between the covariates and the treatment assignment, the selection may be focused on those covariates with a stronger linear correlation with the treatment assignment. Although correlation is a distinguishing factor between confounders and outcome-predictors in the assumed causal structure, there may be situations in which the model over-selects a few dominant confounders in terms of correlation instead of selecting from a wider pool of candidates. Lastly, this study focused solely on empirically suggesting the intended variable subsets and cannot guarantee their identification theoretically.

Despite these limitations, the VSCEN can be useful in providing accurate treatment effect estimation for various data, while increasing knowledge about the underlying causal structure. This can aid causal decision-making in situations such as drug prescription, in which important causal variables must be primarily considered. In future research, we plan to expand the work to more complex causal scenarios, such as those with selection bias or effect modifiers. Furthermore, we plan to extend our model to handle multiple treatment scenarios.

Funding

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (2023R1A2C2005453, RS-2023-00218913).

Notes on contributors

Eun-Yeol Ma received his BS degree in 2017 and MS degree in 2019 in industrial engineering from KAIST. He is currently pursuing a PhD degree in the Department of Industrial and Systems Engineering, KAIST. His research interests include machine learning, applied statistics, and causal inference.

Uichin Lee is a full professor in the School of Computing at KAIST. He received a B.S. in computer engineering from Chonbuk National University in 2001, an M.S. degree in computer science from KAIST in 2003, and a PhD degree in computer science from UCLA in 2008. He continued his studies at UCLA as a postdoctoral research scientist (2008–2009) and then worked for Alcatel-Lucent Bell Labs as a member of the technical staff until 2010. His major research areas are human-computer interaction (HCI), ubiquitous computing (UbiComp), IoT data science, and data visualization. In 2023, he was named to the ACM SIGCHI Academy in recognition of his contributions to the field of HCI. He has been regularly serving as a program committee member of the key HCI conferences and journals such as ACM CHI, PACM IMWUT/ACM Ubicomp, and PACM HCI/ACM CSCW. He served as a program committee chair of ACM Ubicomp 2021, IEEE/IFIP WONS 2021, and general co-chairs of HCI Korea 2021 and IEEE ICMU 2021. He received the best paper awards at ACM CHI'16, AAAI ICWSM'13, IEEE CCGrid'11, and IEEE PerCom'07, and an impact award from the IEEE IoT Forum'19.

Heeyoung Kim received her BS and MS degrees in industrial engineering from KAIST, an MS degree in statistics and a PhD degree in industrial engineering from the Georgia Institute of Technology. She is an associate professor with the Department of Industrial and Systems Engineering, KAIST. She was a Senior Member of Technical Staff with AT&T Laboratories. Her research interests include applied statistics and machine learning.

Data availability statement

The simulated datasets used in this study are available at https://github.com/eyma3115/vscen. The MIMIC-III dataset (Johnson *et al.*,

2016) is available at https://mimic.mit.edu/ upon request for Physionet credentialed users (Goldberger *et al.*, 2000)

References

Alaa, A.M., Weisz, M. and van der Schaar, M. (2017) Deep counterfactual networks with propensity-dropout, in *ICML 2017 Workshop on Principled Approaches to Deep Learning*. Online proceedings without a specified publisher, Sydney, Australia, pp. 1–5.

Balın, M.F., Abid, A. and Zou, J. (2019, 9–15 Jun). Concrete autoencoders: Differentiable feature selection and reconstruction, in *Proceedings of the 36th International Conference on Machine Learning*, PMLR, Long Beach, CA, USA, pp. 444–453.

Beam, A.L. and Kohane, I.S. (2018, 04) Big data and machine learning in health care. *The Journal of the American Medical Association*, **319**(13), 1317–1318.

Beaulieu-Jones, B., Finlayson, S.G., Chivers, C., Chen, I., McDermott, M., Kandola, J., Dalca, A.V., Beam, A., Fiterau, M. and Naumann, T. (2019, 10) Trends and focus of machine learning applications for health research. *JAMA Network Open*, **2**(10), e1914051–e1914051.

Brookhart, M.A., Schneeweiss, S., Rothman, K.J., Glynn, R.J., Avorn, J. and Stürmer, T. (2006, 04) Variable selection for propensity score models. *American Journal of Epidemiology*, **163**(12), 1149–1156.

Butcher, B., Huang, V.S., Robinson, C., Reffin, J., Sgaier, S.K., Charles, G. and Quadrianto, N. (2021) Causal datasheet for datasets: An evaluation guide for real-world data analysis and data collection design using Bayesian networks. *Frontiers in Artificial Intelligence*, **4**, 612551.

Chu, Z., Rathbun, S.L. and Li, S. (2020) Matching in selective and balanced representation space for treatment effects estimation, in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, Association for Computing Machinery, New York, NY, USA, pp. 205–214.

De Luna, X., Waernbaum, I. and Richardson, T.S. (2011, 10) Covariate selection for the nonparametric estimation of an average treatment effect. *Biometrika*, **98**(4), 861–875.

Deaton, A. and Cartwright, N. (2018) Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, **210**, 2–21.

Doo, W. and Kim, H. (2023) Simultaneous deep clustering and feature selection via k-concrete autoencoder. *IEEE Transactions on Knowledge and Data Engineering*, 1–17. https://doi.org/10.1109/TKDE.2023.3323580

Du, X., Sun, L., Duivesteijn, W., Nikolaev, A. and Pechenizkiy, M. (2021) Adversarial balancing-based representation learning for causal effect inference with observational data. *Data Mining and Knowledge Discovery*, **35**(4), 1713–1738.

Frieden, T.R. (2017) Evidence for health decision making—beyond randomized, controlled trials. *New England Journal of Medicine*, **377**(5), 465–475.

Ganin, Y. and Lempitsky, V. (2015, 07–09 Jul) Unsupervised domain adaptation by backpropagation, in *Proceedings of the 32nd International Conference on Machine Learning*, Curran Associates, Inc., Red Hook, NY, USA, pp. 1180–1189.

Gentzel, A., Garant, D. and Jensen, D. (2019) The case for evaluating causal models using interventional measures and empirical data, in *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, Volume 32, Curran Associates, Inc., Red Hook, NY, USA.

Goldberger, A.L., Amaral, L.A., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.-K. and Stanley, H.E. (2000) Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation*, **101**(23), e215–e220.

Goldstein, C.E., Weijer, C., Brehaut, J.C., Fergusson, D.A., Grimshaw, J.M., Horn, A.R. and Taljaard, M. (2018) Ethical issues in pragmatic randomized controlled trials: A review of the recent literature identifies gaps in ethical argumentation. *BMC Medical Ethics*, **19**(1), 1–10.

Gordon, B.R., Zettelmeyer, F., Bhargava, N. and Chapsky, D. (2019) A comparison of approaches to advertising measurement: Evidence from big field experiments at facebook. *Marketing Science*, **38**(2), 193–225.

Greenewald, K., Shanmugam, K. and Katz, D. (2021, 13–15 Apr) High-dimensional feature selection for sample efficient treatment effect estimation, in *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, *Proceedings of Machine Learning Research*, **130**, 2224–2232.

Guo, R., Cheng, L., Li, J., Hahn, P.R. and Liu, H. (2020) A survey of learning causality with data: Problems and methods. *ACM Computing Surveys (CSUR),* **53**(4), 1–37.

Hassanpour, N. and Greiner, R. (2019, 26 Apr – 01 May) Learning disentangled representations for counterfactual regression, in *International Conference on Learning Representations*. Openreview.net, Addis Ababa, Ethiopia, pp. 1–11.

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

Holland, P.W. (1986) Statistics and causal inference. *Journal of the American Statistical Association*, **81**(396), 945–960.

Imbens, G.W. (2020) Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *Journal of Economic Literature*, **58**(4), 1129–1179.

Imbens, G.W. and Rubin, D.B. (2015) *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, New York, NY.

Jang, E., Gu, S. and Poole, B. (2017, 24 – 26 Apr) Categorical reparameterization with gumbel-softmax, in *International Conference of Learning Representations*. Openreview.net, Toulon, France, pp. 1–12.

Johansson, F., Shalit, U. and Sontag, D. (2016, 20–22 Jun) Learning representations for counterfactual inference, in *Proceedings of The 33rd International Conference on Machine Learning*, PMLR, New York City, NY, USA, pp. 3020–3029.

Johnson, A.E., Pollard, T.J., Shen, L., Li-Wei, H.L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A. and Mark, R.G. (2016) Mimic-iii, a freely accessible critical care database. *Scientific Data*, **3**(1), 1–9.

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

Koch, B., Vock, D.M. and Wolfson, J. (2018) Covariate selection with group lasso and doubly robust estimation of causal effects. *Biometrics*, **74**(1), 8–17.

Kuang, K., Cui, P., Li, B., Jiang, M., Yang, S. and Wang, F. (2017, Feb) Treatment effect estimation with data-driven variable decomposition, in *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 31. AAAI Press, Palo Alto, CA, USA, pp. 1–7.

Kuang, K., Cui, P., Zou, H., Li, B., Tao, J., Wu, F. and Yang, S. (2022) Data-driven variable decomposition for treatment effect estimation. *IEEE Transactions on Knowledge and Data Engineering*, **34**(5), 2120–2134.

Kyrimi, E., Dube, K., Fenton, N., Fahmi, A., Neves, M.R., Marsh, W. and McLachlan, S. (2021) Bayesian networks in healthcare: What is preventing their adoption? *Artificial Intelligence in Medicine*, **116**, 102079.

Li, X. and Yao, L. (2022, 28 Nov–1 Dec) Contrastive individual treatment effects estimation, in *2022 IEEE International Conference on Data Mining (ICDM)*, IEEE Press, Piscataway, NJ, pp. 1053–1058.

Louizos, C., Shalit, U., Mooij, J., Sontag, D., Zemel, R. and Welling, M. (2017) Causal effect inference with deep latent-variable models, in *Advances in Neural Information Processing Systems 31 (NeurIPS 2017)*, Volume 2017-Decem, Curran Associates, Inc., Red Hook, NY, USA, pp. 6447–6457.

Maddison, C.J., Mnih, A. and The, Y.W. (2017, 24 – 26 Apr) The concrete distribution: A continuous relaxation of discrete random variables, in *International Conference of Learning Representations*. Openreview.net, Toulon, France, pp. 1–12.

Makar, M., Swaminathan, A. and Kıcıman, E. (2019) A distillation approach to data efficient individual treatment effect estimation, in *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 33, AAAI Press, Palo Alto, CA, USA, pp. 4544–4551.

Myers, J.A., Rassen, J.A., Gagne, J.J., Huybrechts, K.F., Schneeweiss, S., Rothman, K.J., Joffe, M.M. and Glynn, R.J. (2011) Effects of

adjusting for instrumental variables on bias and precision of effect estimates. *American Journal of Epidemiology*, **174**(11), 1213–1222.
Oprescu, M., Syrgkanis, V. and Wu, Z.S. (2019, 09–15 Jun) Orthogonal random forest for causal inference, in *Proceedings of the 36th International Conference on Machine Learning*, PMLR, Long Beach, CA, USA, pp. 4932–4941.
Pearl, J. (2009) *Causality*. Cambridge University Press, New York, NY.
Pearl, J. and Paz, A. (2014) Confounding equivalence in causal inference. *Journal of Causal Inference*, **2**(1), 75–93.
Roeder, G., Metz, L. and Kingma, D. (2021, 18–24 Jul) On linear identifiability of learned representations, in *Proceedings of the 38th International Conference on Machine Learning*, PMLR, Virtual, pp. 9030–9039.
Rosenbaum, P.R. and Rubin, D.B. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**(1), 41–55.
Rotnitzky, A. and Smucler, E. (2020) Efficient adjustment sets for population average causal treatment effect estimation in graphical models. *Journal of Machine Learning Research*, **21**(188), 1–86.
Rubin, D.B. (2005) Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, **100**(469), 322–331.
Schisterman, E.F., Cole, S.R. and Platt, R.W. (2009) Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology (Cambridge, Mass.)*, **20**(4), 488–495.
Shalit, U., Johansson, F.D. and Sontag, D. (2017, 06–11 Aug) Estimating individual treatment effect: Generalization bounds and algorithms, in *Proceedings of the 34th International Conference on Machine Learning*, PMLR, Sydney, Australia, pp. 3076–3085.
Shi, C., Blei, D. and Veitch, V. (2019) Adapting neural networks for the estimation of treatment effects, in *Advances in Neural Information Processing Systems 33 (NeurIPS 2019)*, Volume 32. Curran Associates, Inc., Red Hook, NY, USA.
Shortreed, S.M. and Ertefaie, A. (2017) Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics*, **73**(4), 1111–1122.
Singer, M., Deutschman, C.S., Seymour, C.W., Shankar-Hari, M., Annane, D., Bauer, M., Bellomo, R., Bernard, G.R., Chiche, J.-D., Coopersmith, C.M., Hotchkiss, R.S., Levy, M.M., Marshall, J.C., Martin, G.S., Opal, S.M., Rubenfeld, G.D., van der Poll, T., Vincent, J.-L. and Angus, D.C. (2016, 02) The third international consensus definitions for sepsis and septic shock (Sepsis-3). *The Journal of the American Medical Association*, **315**(8), 801–810.
Udekwu, P., Kromhout-Schiro, S,. Vaslef, S., Baker, C. and Oller, D. (2004) Glasgow coma scale score, mortality, and functional outcome in head-injured patients. *Journal of Trauma and Acute Care Surgery*, **56**(5), 1084–1089.
Wager, S. and Athey, S. (2018) Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, **113**(523), 1228–1242.
Wu, A., Yuan, J., Kuang, K., Li, B., Wu, R., Zhu, Q., Zhuang, Y.T. and Wu, F. (2022) Learning decomposed representations for treatment effect estimation. *IEEE Transactions on Knowledge and Data Engineering*, **35**, 4989–5001.
Wu, P. and Fukumizu, K. (2021, 03 – 07 May) β-intact-vae: Identifying and estimating causal effects under limited overlap, in *International Conference of Learning Representations*, Virtual.
Yoon, J., Jordon, J. and Van Der Schaar, M. (2018, 30 Apr – 03 May) Ganite: Estimation of individualized treatment effects using generative adversarial nets, in *International Conference on Learning Representations*. Openreview.net, Vancouver, BC, Canada.
Zhang, W., Liu, L. and Li, J. (2021) Treatment effect estimation with disentangled latent factors. *Proceedings of the AAAI Conference on Artificial Intelligence*, **35**(12), 10923–10930.
Zhao, Q., Keele, L.J. and Small, D.S. (2019) Comment: Will competition-winning methods for causal inference also succeed in practice? *Statistical Science*, **34**(1), 72–76.
Zhou, G., Yao, L., Xu, X., Wang, C. and Zhu, L. (2022) Cycle-balanced representation learning for counterfactual inference, in *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*, SIAM, Philadelphia, PA, pp. 442–450.
Zigler, C.M. and Dominici, F. (2014) Uncertainty in propensity score estimation: Bayesian methods for variable selection and model-averaged causal effects. *Journal of the American Statistical Association*, **109**(505), 95–107.
Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **67**(2), 301–320.