Causal effect estimation using variational information bottleneck

Zhenyu Lu, Yurong Cheng, Mingjun Zhong, George Stoian, Ye Cheng, Guoren Wang
School of Computer Science, Beijing Institute of Technology
Department of Computing Science, University of Aberdeen

Abstract—**Causal inference is to estimate the causal effect in a causal relationship when intervention is applied. Precisely, in a causal model with binary interventions, i.e., control and treatment, the causal effect is simply the difference between the factual and counterfactual. The difficulty is that the counterfactual may never been obtained which has to be estimated and so the causal effect could only be an estimate. The key challenge for estimating the counterfactual is to identify confounders which effect both outcomes and treatments. A typical approach is to formulate causal inference as a supervised learning problem and so counterfactual could be predicted. Including linear regression and deep learning models, recent machine learning methods have been adapted to causal inference. In this paper, we propose to estimate Causal Effect by using Variational Information Bottleneck (CEVIB). The promising point is that VIB is able to naturally distill confounding variables from the data, which enables estimating causal effect by using observational data. We have compared CEVIB to other methods by applying them to three data sets showing that our approach achieved the best performance. We also experimentally showed the robustness of our method.**

Index Terms—**Causal Inference, Causal Effect, Variational Information Bottleneck, Confounding Variables, Intervention.**

I. INTRODUCTION

Causal inference [7], [15], [16] is a task to estimate the effect in a causal relationship when an intervention action is made, which could be used to guide decision-making. In general, causal inference is to estimate the level of outcome changes when their causes are intervened. For example, in medical science we require to estimate the causal effect of a treatment to understand the effectiveness of the treatment applied for treating a particular disease. Causal inference is hard because we could only observe the factual, but would never be able to obtain the counterfactual. The key to estimate the causal effect is to estimate the counterfactual, and so the causal effect is simply the difference between factual and counterfactual. A possible approach for causal inference is the randomized controlled trails (RCT), which is a standard approach in clinical trials. In RCT, two groups of participants are randomly chosen and they are the control and treatment groups. The effect of the treatment is then computed by using the outcomes of the two groups, i.e., the average difference between their outcomes. Although RCT is considered as a most reliable approach for causal inference in practice, it could cost expensive and would be impossible in some situations. In RCT, large number of participants may be required to control the effect of confounding factors which affect both the treatment and outcome. For estimating the causal effect, it is crucial to identify the confounding variables and then eliminate their effects on the outcomes so that the effect of the treatment could be correctly estimated.

Simply, causal inference could be formulated as a linear regression model, where the outcome is represented as the linear regression of the treatment and any other feature variables [7] including confounding variables. The major disadvantage of linear regression approach is its limited ability to handle big data, which is the favorable circumstance in the big data era. Recent advances in deep learning techniques provide an excellent opportunity to utilise large observational data to estimate causal effects. In literature, a couple of deep neural network (DNN) approaches have been proposed for causal inference. For example, Dragonnet [19] proposed an end-to-end DNN architecture with three heads to estimate treatment outcomes. The factual and counterfactual outcomes are respectively represented as two of the three heads. In this architecture, the confounding variables are supposed to be learned from the observational data, which is the main point of the proposed architecture. Variational autoencoder (VAE) is also proposed to modeling the confounding variables so that the causal effect could also be estimated using a neural network [13]. It is interesting to note that all these methods employing deep neural networks attempt to represent the causal inference as a supervised learning problem. The important aspect for such approach is that once the causal inference is posed as a supervised learning problem, the established supervised learning algorithms could then be easily applied boosting the research in this area.

In this paper, we consider to use neural network inspired by variational information bottleneck [2] to not only fit a model to the observed data, but also address hidden confounders that affects treatments or outcomes. Based on this approach, we develop a novel regularization framework that forces the model to "forget" some hidden parameters which are not confounders and learning to extract the confounding variables from data. We present experiments in section V that demonstrate the advantages of this model and show that it outperforms state-of-the-art models in a variety of datasets.

II. CAUSAL EFFECT MODELS

Suppose a dataset $D = \{X_i, T_i, Y_i\}_{i=1}^N$ is given, where X represents the covariates of a subject, e.g., the health status, T represents the treatment applied to the patient, e.g., a

medication, and Y represents the outcome after the treatment is applied. Note that we consider a binary variable for T and so $T \in \{0,1\}$ where $T = 0$ means that no medication is applied to the control group, and $T = 1$ the medication applied to the treatment group. For a subject X_i, $T_i = 0$ or $T_i = 1$. For a control subject i, since $T_i = 0$ is implemented, the outcome y_{i0} is called the factual outcome which is observed; however, $T_i = 1$ was never implemented, the potential outcome y_{i1} is called the counterfactual outcome which is thus never able to be observed. Conversely, for a treatment subject j, the outcome y_{j1} is the factual outcome which is observed, and y_{j0} is the counterfactual outcome, which is never observed. For computing the causal treatment effect for a subject i, i.e., $y_{i1} - y_{i0}$, it is required to know both factual and counterfactual outcomes of a subject, but the counterfactual outcome is not obtained. Our objective is to estimate the average treatment effect (ATE) denoted by ψ and individual treatment effect (ITE) when a treatment T, e.g., a medication, is applied to a patient.

For a subject i with a d-dimensional covariate $X_i \in R^d$ and its outcome being $Y_i \in R^1$ after a treatment T_i is applied, we assume that the observed covariates X_i include all possible causes for outcome and treatment. Those causal variables, i.e., confounders, for both treatment and outcome are hard to identify. We therefore assume the cofounders to be a collection of latent variables which could be distilled from covariates. Figure 1 (left) represents an observational data model using this assumption, in which X, T and Y are observed, and Z are the confounders. To estimate causal effect, it needs to use the intervention model (the right graph in Figure 1), in which the intervention treatment, i.e., Pearl's $do()$ action, is applied which does not effected by any confounders. We would also assume the following Assumptions 2.1-2.3, which is sufficient to identify the treatment effect from the observed data [8]:

Assumption 2.1: The potential outcome of treatment T equals the observed outcome if the actual treatment received is T. i.e. $\forall t \in \{0,1\}, if[T = t], then[Y = Y(t)]$.

Assumption 2.2: For any set of covariates X the probability to receive treatment 0 or 1 is positive. i.e. $\forall X \in R^d : P(T = 0|X) \in (0,1) \; and \; P(T = 1|X) \in (0,1)$.

Assumption 2.3: The potential outcomes $Y(1)$ and $Y(0)$ are independent of the treatment assignment T given the covariates X. i.e. $Y \perp\!\!\!\perp T|X$.

Under these assumptions, the individual treatment effect could be represented as

$$
\begin{aligned}
ITE(x_i) = &E_{Y|X,do(T_i=1)}[Y|X = x_i, do(T_i = 1)] \\
&- E_{Y|X,do(T_i=0)}[Y|X = x_i, do(T_i = 0)].
\end{aligned} \tag{1}
$$

The ATE is then the expectation with respect to all the individuals and so

$$
\begin{aligned}
\psi &= E_X[\; ITE(X)] \\
&= E_X\left[E_{Y|X,do(T_i=1)}[Y|X = x_i, do(T_i = 1)]\right. \\
&\quad \left. - E_{Y|X,do(T_i=0)}[Y|X = x_i, do(T_i = 0)]\right].
\end{aligned} \tag{2}
$$

Note that Pearl's $do()$ notation is used to represent the causal relationship indicating the potential effect when the subject is intervened by applying the treatment. Note that Assumption 2.1 is equivalent to the principle of independent mechanism (PIM) [16], which assumes that the condional distributions in the system do not influence each other. Under the PIM assumption, the conditionals in the intervention model do not change and so are identical to those in the observational data model. This assumption allows us to apply the learned conditionals by using observation data model into the intervention model for inference.
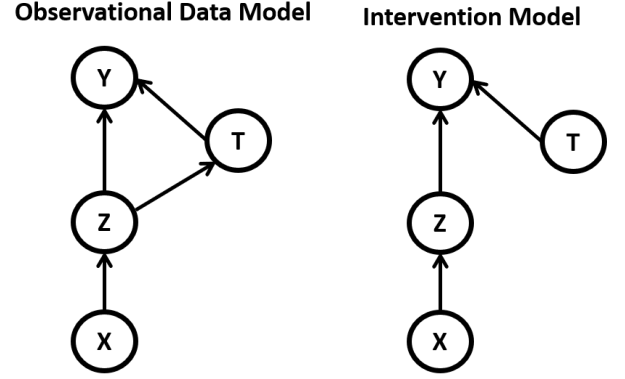

Figure 1. The observation model (left) and the intervention model (right). X are covariates, Z are the latent confounders, T is a treatment, and Y is an outcome.

III. RELATED WORKS

In this section, we briefly discuss causal effect estimation methods. Statistical methods have been proposed for causal effect estimation. For example, regression methods fit the treatment assignment and as well as the covariates to represent the outcomes [3], [7], [11], [15]. Sample re-weighting methods aim to correct the treatment assignment using observational data in order to overcome the subject selection bias. For instance, these methods include Inverse Propensity Weighting (IPW) based on propensity score [17] and confounder balancing methods [12]. Other methods include doubly-robust approaches which combine covariate adjustment with propensity score weighting [3]–[5].

In recent years, many machine learning methods have been applied to the problem of estimating the potential outcomes and treatment effects. Most of them could be viewed as discriminative (or supervised) learning methods. For example, Bayesian Additive Regression Trees (BART) [6] and Causal Forests [21] have been used to estimate causal effect. The work in [9] proposed Balanced Linear Regression (BLR) and Balanced Neural Networks (BNN) to learn a balanced covariate representation for causal outcomes. Other approaches include Causal Multi-task Gaussian Processes(CMGP) where the factual and counterfactual outcomes are modelled using a vector-valued reproducing kernel Hilbert space [1] and GANITE which uses generative adversarial networks [23] to estimate causal effect.
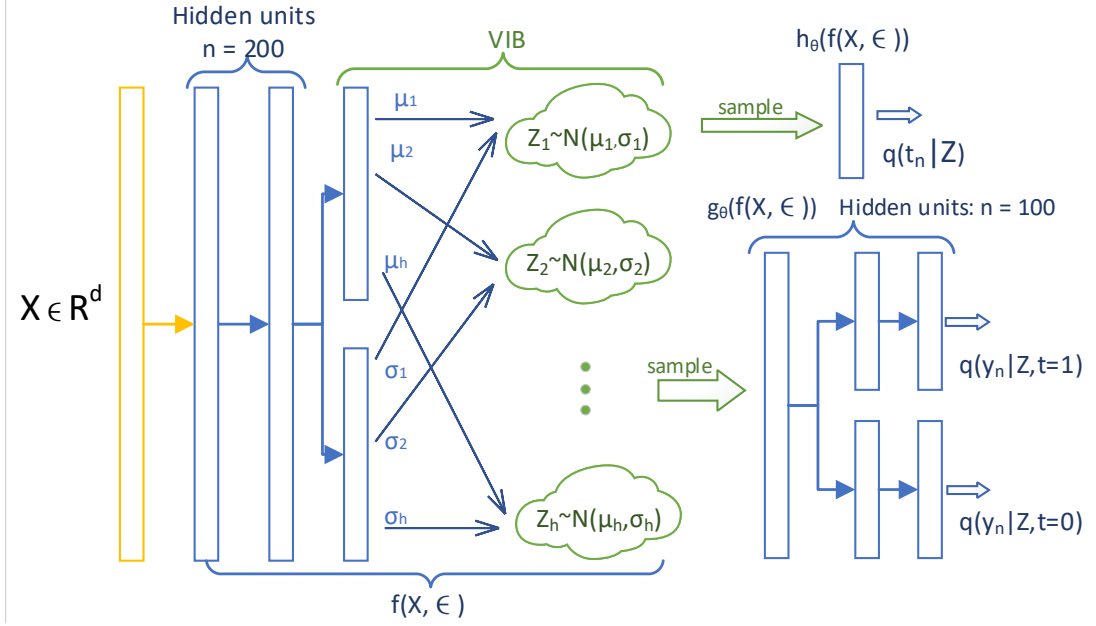
Figure 2. The architecture framework for estimating causal effect using variational information bottleneck.

Methods that have similarities with CEVIB are CEVAE [14] and Dragonnet [19]. All these methods have a similar neural networks structure to that of TARnet [18]. An advantage of CEVIB is that this method uses variatinal bottleneck approach to represent the confounding variables as a distribution which could be able to take into account the uncertainty of the model. In the experiment section, we show that CEVIB is robust in terms of subject sample bias to support such advantage.

IV. VARIATIONAL INFORMATION BOTTLENECK FOR CAUSAL INFERENCE

Under the sufficient assumptions provided in Section II, the observed data $D = \{X, T, Y\}$ were generated from the observational data model illustrated in Figure 1 with confounding variables Z. The confounders Z effect both treatment T and outcomes Y. To estimate causal effects, instead of using the obervational data model, the intervention model should be used where the treatment is intervened. However, all the conditional distributions in the intervention model is unknown, and they need to be learned from data. According to the principle of independent mechanism, the conditionals in intervention model are identical to those in observational data model. There is hope that these conditionals could be learned from the data. According to our observational data model, we hope that all the confounding variables could be distilled from the data X because X contain all the possible confounders. We employ the idea of vational information bottleneck (VIB) which attempts to distill sufficient informaton from X which maximumly inform both T and Y (see the Figure 1). Our VIB network architecture (CEVIB) is shown in the Figure 2. Our framework is most relevant to the Dragonnet [19]. CEVIB provides an end-to-end procedure for predicting treatment and outcome. We firstly use neural network to learn the representation $Z(x) \sim N(\mu^k, \theta^k)$ following a Normal distribution,

where k denotes the dimension of hidden confounders. The hidden confounders are regularized by information bottleneck and predict the treatment and outcome. The regulation of variational information bottleneck enforces the model to extract hidden confounding variables as much as possible and forget those variables which are not confounders.

A. Training the observational model

In this section, we apply the idea of VIB to train the observation model using the data. Denote $\widetilde{Y} = \{Y, T\}$. Our aim is to optimize the following problem which attempts to maximizing the mutual information between $\widetilde{Y}$ and the confounders Z whist threshold the mutual information between the covariates X and Z:

$$\max_{\theta} I(Z, \widetilde{Y}; \theta), \text{ subject to, } I(X, Z; \theta) \leq I_c, \quad (3)$$

where $I(Z, \widetilde{Y}; \theta) = \int p(Z, \widetilde{Y}) \log \frac{p(Z, \widetilde{Y})}{p(Z)p(\widetilde{Y})} d\widetilde{Y} dZ$ represents the mutual information, and I_c is a constant. This is equivalent to maximizing the following objective function

$$R_{IB}(\theta) = I(Z, \widetilde{Y}; \theta) - \beta I(X, Z; \theta). \quad (4)$$

Instead we could maximize the following lower bound

$$R_{IB}(\theta) = I(Z, \widetilde{Y}; \theta) - \beta I(X, Z; \theta)$$
$$\geq \int p(x)p(\widetilde{y}|x)p(z|x) \log q(\widetilde{y}|z) dx d\widetilde{y} dz$$
$$- \beta \int p(x)p(z|x) \log \frac{p(z|x)}{r(z)} dx dz$$
$$\approx \frac{1}{N} \sum_{n=1}^{N} \left\{ \int p(z|x_n) \left[\log q(\widetilde{y_n}|z) - \beta \log \frac{p(z|x_n)}{r(z)} \right] dz \right\}$$
$$= L. \quad (5)$$

$$L = \frac{1}{N}\sum_{n=1}^{N}\left\{\int p(z|x_n)\left[t_n\log q(y_n|z,t_n=1)+(1-t_n)\log q(y_n|z,t_n=0)+\log q(t_n|z)-\beta\log\frac{p(z|x_n)}{r(z)}\right]dz\right\}$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left\{\int p(z|x_n)\left[t_n\log q(y_n|z,t_n=1)+(1-t_n)\log q(y_n|z,t_n=0)+\log q(t_n|z)\right]dz+\beta KL[p(z|x_n)||r(z)]\right\}$$

$$= L_1 + L_2 \tag{9}$$

$$L_1 = \frac{1}{N}\sum_{n=1}^{N}\left\{\frac{1}{M}\sum_{m=1}^{M}\left[t_n\log q(y_n|f(x_n,\epsilon_m),t_n=1)+(1-t_n)\log q(y_n|f(x_n,\epsilon_m),t_n=0)+\log q(t_n|f(x_n,\epsilon_m))\right]\right\} \tag{10}$$

$$L_2 = \frac{\beta}{N}\sum_{n=1}^{N}KL[p(z|x_n)||r(z)] \tag{11}$$

In the causal effect model (see the observation model in Figure 1), our outputs given the latent variable has the form

$$q(\widetilde{y}|z) = q(y|z,t=1)^t q(y|z,t=0)^{1-t} q(t|z)$$

So the lower bound can be represented as the equation (9).

Suppose the latent variable z has dimension K, we assume a diagonal multivariate Normal distribution,

$$p(z|x_n) = \prod_{k=1}^{K}p(z_k|x_n) = \prod_{k=1}^{K}N(z_k|\mu_k(x_n),\sigma_k^2(x_n))$$

where μ_k and σ_k are neural networks. Therefore, we can use change of variables to draw a sample $z_k = f_k(x_n,\epsilon) = \mu_k(x_n)+\sigma_k(x_n)\epsilon$ where $\epsilon \sim N(0,1)$. Denote $f(x_n,\epsilon) = [f_1(x_n,\epsilon),\cdots,f_K(x_n,\epsilon)]^T$. So $L1$ can be expressed as equation (10). We define the following distributions for outcome and treatment variables

$$\log q(y_n|f(x_n,\epsilon_m),t_n) \approx -(y_n-g_\theta^{t_n}(f(x_n,\epsilon)))^2 \tag{12}$$
$$\log q(t_n|f(x_{n,m})) \approx -(t_n-h_\theta(f(x_n,\epsilon)))^2 \tag{13}$$

where $g_\theta^{t_n}(.)$ and $h_\theta(.)$ are neural networks. Therefore, our task is then to learn the neural networks $g_\theta^{t_n}(.)$, $h_\theta(.)$, $\mu_k(.)$ and $\sigma_k(.)$. For the L_2, we use the identity

$$KL\left(N\left((\mu_1,\cdots,\mu_K)^T,diag(\sigma_1^2,\cdots,\sigma_K^2)\right)||N(0,1)\right)$$
$$= \frac{1}{2}\sum_{k=1}^{K}(\sigma_k^2+\mu_k^2-1-ln(\sigma_k^2))$$

So we have

$$L_2 = \frac{\beta}{N}\sum_{n=1}^{N}KL[p(z|x_n)||r(z)]$$
$$= \frac{\beta}{N}\sum_{n=1}^{N}\left[\frac{1}{2}\sum_{k=1}^{K}(\sigma_k^2(x_n)+\mu_k^2(x_n)-1-ln(\sigma_k^2(x_n)))\right]$$

Our CEVIB is to train to maximize L_1 and L_2 using observational data, and so all the conditional distributions are trained, which can then be used in the intervention model to estimate causal effects, which we will describe in the following subsection.

B. Estimating causal effects

After training, the average treatment affect τ can be calculated using,

$$\tau = \int Y p(Y|do(T=1))dY - \int Y p(Y|do(T=0))dY \tag{14}$$

We can then apply the $do-calculus$ according to the intervention model structure (see the intervention model in Figure 1)

$$p(Y|do(T=t))$$
$$= \int p_{do(T=t)}(Y,Z,X,T=t)dZdX \tag{15}$$
$$= \int p_{do(T=t)}(Y|Z,X,T=t)p_{do(T=t)}(Z,X,T=t)dZdX \tag{16}$$
$$= \int p_{do(T=t)}(Y|Z,T=t)p(Z|X)p(X)p_{do(T=t)}(T=t)dZdX \tag{17}$$
$$= \int p(Y|Z,T=t)p(Z|X)p(X)dZdX \tag{18}$$
$$\approx \int q(Y|Z,T=t)p(Z|X)p(X)dZdX \tag{19}$$

where we have applied the approximation $p(Y|Z,T=t) \approx q(Y|Z,T=t)$. Note that all these conditionals involving computing $p(Y|do(T=t))$ have been trained using observational model. We then draw samples from $p(Y|do(T=t))$ which will be used to calculate the average treatment affect. To draw samples from $p(Y|do(T=t))$, we do the following process

- $x^{(i)} \sim p(X)$, i.e., draw a sample $x^{(i)}$ from $p(X)$
- $z^{(i)} \sim p(Z|x^{(i)})$
- $y_t^{(i)} \sim q(Y|z^{(i)},T=t)$

This provides a Monte Carlo estimate for the average treatment affect,

$$\hat{\tau} = \frac{1}{N_1}\sum_{i=1}^{N_1}y_1^{(i)} - \frac{1}{N_0}\sum_{i=1}^{N_0}y_0^{(i)}$$

We will use this method to estimate causal effects in the following experiments.

V. Experiments

A. The data sets and training setup

Due to the difficulty of gathering treatment effects factuals for both control and treatment, evaluating the methods for estimating causal effects may have to use synthetic or semi-synthetic data sets. Our experiments will apply CEVIB to three semi-synthetic benchmark data sets: **IHDP** [22], **Twins** [23] and **ACIC** [20].

IHDP: This data is constructed from the Infant Health and Development Program (IHDP). There are 100 files in which each file contains 747 subjects[1]. Both factual and counterfactual are provided for each subject, which provides a ground truth for evaluating causal inference algorithms.

Twins: This data is a benchmark task that utilizes data from twin births in the USA between 1989-1991. There are 11399 subjects in this data[2]. The samples in the data set are all twins, $t = 1$ represents the heavier baby, and the outcome Y corresponds to the mortality of each of the twins in their first year of life.

ACIC: This data is a collection of semi-synthetic datasets derived from the linked birth and infant death data (LBIDD) [20]. It was developed for the 2018 Atlantic Causal Inference Conference competition (ACIC)[3], which include 30 different data generating process settings with subject sample sizes from 1,000 to 50,000.

To train CEVIB, we use the architecture in Figure 2. For both IHDP and ACIC experiments, we randomly split each file data into test/validation/train with proportion 63/27/10 and report the In Sample[4] and Out of Sample[5] errors, and repeat the procedure for 25 times. For Twins experiments, we randomly split the data into test/validation/train with proportion 56/24/20 and report the In Sample and Out of Sample errors, and repeat the procedure for 50 times.

To evaluate the performance, we report the results of the following metrics for each data set, which are the absolute error in average treatment effect

$$\epsilon_{\text{ATE}} = |\frac{1}{N} \sum_{i=1}^{N} (y_1^{(i)} - y_0^{(i)}) - \frac{1}{N} \sum_{i=1}^{n} (\hat{y}_1^{(i)} - \hat{y}_0^{(i)})|,$$

the Precision in Estimation of Heterogeneous Effect (PEHE)

$$\epsilon_{\text{PEHE}} = \frac{1}{N} \sum_{i=1}^{N} ((y_1^{(i)} - y_0^{(i)}) - (\hat{y}_1^{(i)} - \hat{y}_0^{(i)}))^2,$$

[1] The IHDP data set is available at https://github.com/Osier-Yi/SITE/tree/master/data

[2] The Twins data set is available at https://github.com/jsyoon0823/GANITE/tree/master/data or https://github.com/AMLab-Amsterdam/CEVAE/tree/master/ datasets/TWINS.

[3] We use the scaling folder in ACIC to evaluate our methods. The ACIC data set is available at https://github.com/IBM-HRL-MLHLS/IBM-Causal-Inference-Benchmarking-Framework/tree/master/data/LBIDD.

[4] In Sample uses all observational data for both training and prediction.

[5] Out of Sample uses the training set to train and the test set for prediction.

and the relative PEHE:

$$PEHE_{REL} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{(y_1^{(i)} - y_0^{(i)}) - (\hat{y}_1^{(i)} - \hat{y}_0^{(i)})}{(y_1^{(i)} - y_0^{(i)})} \right)^2.$$

Note that the relative PEHE could be able to remove the scaling effect on PEHE.

We compared CEVIB with Dragonnet [19], the regularized Dragonnet (Dragon-tarreg) [19] and causal effect variational auto-encoder(CEVAE) [13]. We also compare to Dragonnet-tarreg which is a Dragonnet using a targeted regularization method based on the augmented inverse probability weighted(AIPW) estimator. In addition, we also compared CEVIB with least squares regression using treatment as a feature (OLS/LR1), separate least squares regressions for each treatment (OLS/LR2), BLR [10], BART [6], CForest [21], BNN [10], CFR_{wass} and $TARNet$ [18], and GANITE [23].

B. Results

For comparison purpose, various methods described in the previous subsection are applied to the data sets IHDP and Twins. For the error metrics ATE and PEHE, both the within samples and out of samples were computed. The results are shown in Table I which shows that our CEVIB outperforms all of other methods on the data IHDP. On the Twins data set, CEVIB is among the best results. Due to computational limits, on the ACIC data, we compared CEVIB to CEVAE, Dragonnet, and Dragon-tarreg. The results are shown in the Table II. The results clearly indicate that our CEVIB outperforms all the other methods across all the error metrics. To have a more clearer comparison, we randomly chose 10 data files from IHDP and ACIC to plot the distributions for the replicated results using CEVIB, Dragonnet, and Dragonnet-tarreg. The results are plotted in the Figure V-B. We showed the logarithmic values for relative PEHE. It shows that accross all the data files, CEVIB outperformed the other methods.

C. Robustness

We also evaluated the robustness of CEVIB on a synthetic data set via varying selection bias. We assume the control and treatment subjects are from two different distributions. The distance between the two distributions can be evaluated by using Kullback-Leibler (KL) divergence. If the KL distance is larger, and so is the selection bias. Similar to [22], the synthetic data were generated with the following procedure. 5,000 control subject samples were randomly generated from $N(0^{10 \times 1}, 0.5(\Sigma + \Sigma^T))$ and 2500 treatment samples from $N(\mu^{10 \times 1}, 0.5(\Sigma + \Sigma^T))$, where Σ was generated using Uniform distributions $U((-1, 1)^{10 \times 10})$. By changing the value of μ, data with different selection bias can be generated. The outcome is generated by $Y = W^T X + \epsilon$, where $W \sim U((-1, 1)^{10 \times 2})$ and $\epsilon \sim N(0^{2 \times 1}, 0.1 \times I^{2 \times 2})$. In this experiment, we compared CEVIB against Dragonnet and Dragonnet-tarreg. We set $\mu \in [1, 2, 3, 4]$, all the algorithms ran 50 replications using the generated data. The results are shown in figure 4. We observed that when KL values get larger,

Table I

THE RESULTS OF APPLYING VARIOUS MODELS TO IHDP AND TWINS DATA. BEST RESULTS ARE DENOTED IN BOLD. NOTE THAT $within-s$ MEANS THAT TRAINING DATA WERE USED FOR BOTH TRAINING AND PREDICTION AND $out-of-s$ MEANS THAT THE MODELS WERE TRAINED ON TRAINING DATA AND TESTED ON THE TEST DATA.

	Datasets(Mean +- std)							
	IHDP				Twins			
Method	$\sqrt{\epsilon_{PEHE}^{within-s}}$	$\epsilon_{ATE}^{within-s}$	$\sqrt{\epsilon_{PEHE}^{out-of-s}}$	$\epsilon_{ATE}^{out-of-s}$	$\sqrt{\epsilon_{PEHE}^{within-s}}$	$\epsilon_{ATE}^{within-s}$	$\sqrt{\epsilon_{PEHE}^{out-of-s}}$	$\epsilon_{ATE}^{out-of-s}$
OLS-1	5.8±.3	.73±.04	5.8±.3	.94±.06	.319±.001	.0038±.0025	.318±.007	.0069±.0056
OLS-2	2.4±.1	.14±.01	2.5±.1	.31±.02	.320±.002	.0039±.0025	.320±.003	.0070±.0059
BLR	5.8±.3	.72±.04	5.8±.3	.93±.05	.312±.003	.0057±.0036	.323±.018	.0334±.0092
k-NN	2.1±.1	.14±.01	4.1±.2	.79±.05	.333±.001	.0028±.0021	.345±.007	**.0051±.0039**
BART	2.1±.1	.23±.01	2.3±.1	.34±.02	.347±.009	.1206±.0236	.338±.016	.1265±.0234
RF	4.2±.2	.73±.05	6.6±.3	.96±.06	.306±.002	.0049±.0034	.321±.005	.0080±.0051
CF	3.8±.2	.18±.01	3.8±.2	.40±.03	.366±.003	.0286±.0035	.316±.011	.0335±.0083
BNN	2.2±.1	.37±.03	2.1±.1	.42±.03	.325±.003	.0056±.0032	.321±.018	.0203±.0071
CFRW	**.71±.0**	.25±.01	**.76±.0**	.27±.01	.315±.007	.0112±.0016	.313±.008	.0284±.0032
CEVAE	2.7±.1	.34±.01	2.6±.1	.46±.02	.341±.006	.0065±.0040	.373±.012	.0679±.0212
TARNet	.88±.0	.26±.01	.95±.0	.28±.01	.317±.005	.0108±.0017	.315±.003	.0151±.0018
GANITE	1.9±.4	.43±.05	2.4±.4	.49±.05	**.289±.005**	.0058±.0017	**.297±.016**	.0089±.0075
Dragonnet	1.31±.4	.14±.02	1.32±.5	.21±.04	.322±.001	.0092±.0078	.317±.001	.0074±.0092
Dragon-tarreg	1.22±.3	.14±.01	1.30±.3	.20±.05	.322±.0017	.0060±.0088	.318±.002	.0060±.0101
CEVIB	.85±.1	**.12±.01**	.92±.2	**.15±.02**	.320±.0003	**.0016±.0009**	.315±.001	.0055±.0018

Table II

THE RESULTS FOR THE METHODS APPLIED TO ACIC DATA.

	ACIC(Mean +- std)			
Method	$\sqrt{\epsilon_{PEHE}^{within-s}}$	$\epsilon_{ATE}^{within-s}$	$\sqrt{\epsilon_{PEHE}^{out-of-s}}$	$\epsilon_{ATE}^{out-of-s}$
CEVAE	131±1.8	50.2±1.3	118±2.2	51.9±1.4
Dragonnet	67.8±2.0	25.8±2.1	68.2±2.6	26.3±2.1
Dragon-tarreg	62.5±2.0	24.8±3.1	62.2±2.4	24.5±3.0
CEVIB	**43.9±0.8**	**14.2±1.7**	**44.3±1.8**	**14.7±1.9**

the errors are also larger. It shows that Dragonnet was more sensitive to the bias comparing to CEVIB and Dragonnet-tarreg. For both error metrics, it shows that both CEVIB and Dragonnet-tarreg were similar in terms of sensitivity with respect to the bias. This may indicate that CEVIB and Dragonnet-tarreg could have a better ability to extract the latent confounders from observational data.

VI. CONCLUSION

In this paper, we have proposed a variational information bottleneck (VIB) approach to estimate causal effects. The interesting point of using VIB is that VIB is able to distill compact information representing the latent confounders from data. Representing confounding variables is important because the effect of confounding variables on outcome can then be integrated out so that the effect of treatment could be purified. We showed that causal inference could be represented as a prediction problem. We have used VIB to train a model to predict both factual and counterfactual outcomes and so the causal effect can be calculated. The proposed algorithm CEVIB was applied to three data sets and compared to other methods showing that our method outperformed other approaches. We also demonstrated the robustness of our method.

REFERENCES

[1] A. M. Alaa and M. van der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes. *arXiv preprint arXiv:1704.02801*, 2017.

[2] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy. Deep variational information bottleneck. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[3] S. Athey, G. Imbens, T. Pham, and S. Wager. Estimating average treatment effects: Supplementary analyses and remaining challenges. *American Economic Review*, 107(5):278–81, 2017.

[4] D. Benkeser, M. Carone, M. V. D. Laan, and P. Gilbert. Doubly robust nonparametric inference on the average treatment effect. *Biometrika*, 104(4):863–880, 2017.

[5] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

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

[7] A. Gelman and J. Hill. *Data analysis using regression and multi-level/hierarchical models.* Cambridge university press, 2006.

[8] G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences.* Cambridge University Press, Cambridge, 2015.

[9] F. Johansson, U. Shalit, and D. Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pages 3020–3029. PMLR, 2016.

[10] F. D. Johansson, U. Shalit, and D. Sontag. Learning representations for counterfactual inference. page 4407–4418, 2016.

[11] N. Kallus. Recursive partitioning for personalization using observational data. In *International Conference on Machine Learning*, pages 1789–1798. PMLR, 2017.

[12] K. Kuang, P. Cui, B. Li, M. Jiang, and S. Yang. Estimating treatment effect in the wild via differentiated confounder balancing. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 265–274, 2017.

[13] C. Louizos, U. Shalit, J. Mooij, D. Sontag, R. Zemel, and M. Welling. Causal effect inference with deep latent-variable models. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6449–6459, 2017.

[14] C. Louizos, U. Shalit, J. Mooij, D. Sontag, R. Zemel, and M. Welling.
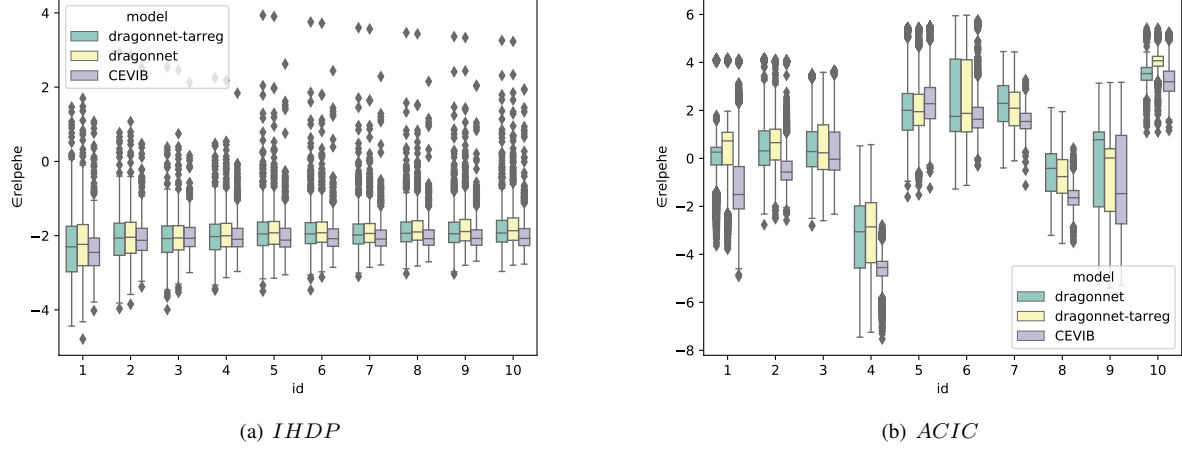
(a) $IHDP$

(b) $ACIC$

Figure 3. The result of relative PEHE (logarithm) when CEVIB, Dragonnet, and Dragonnet-tarreg were applied to IHDP and ACIC. Ten data files were randomly chosen to plot their distributions using the replicated results.

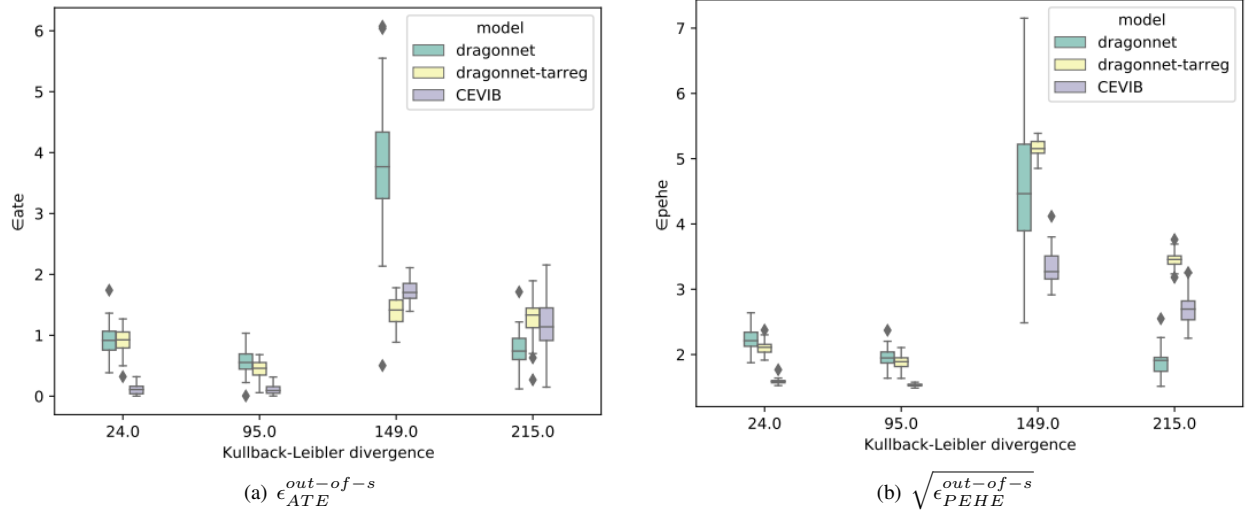(a) $\epsilon_{ATE}^{out-of-s}$

(b) $\sqrt{\epsilon_{PEHE}^{out-of-s}}$

Figure 4. The sensitivity of CEVIB, Dragonnet, and Dragonnet-tarreg to the subject sample bias. The larger KL divergence indicates more difference between the distributions of the control and the treatment samples.

Causal effect inference with deep latent-variable models. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6449–6459, 2017.

[15] J. Pearl. *Causality*. Cambridge University Press, Cambridge, 2009.

[16] J. Peters, D. Janzing, and B. Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[17] P. R. Rosenbaum. Model-based direct adjustment. *Journal of the American Statistical Association*, 82(398):387–394, 1987.

[18] U. Shalit, F. D. Johansson, and D. Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR, 2017.

[19] C. Shi, D. M. Blei, and V. Veitch. Adapting neural networks for the estimation of treatment effects. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 2507–2517, 2019.

[20] Y. Shimoni, C. Yanover, E. Karavani, and Y. Goldschmidt. Benchmarking framework for performance-evaluation of causal inference analysis. *arXiv e-prints*, pages arXiv–1802, 2018.

[21] S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

[22] L. Yao, S. Li, Y. Li, M. Huai, J. Gao, and A. Zhang. Representation learning for treatment effect estimation from observational data. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 2638–2648, 2018.

[23] J. Yoon, J. Jordon, and M. Van Der Schaar. Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*, 2018.