Learning Disentangled Representations for CounterFactual Regression

Negar Hassanpour
Department of Computing Science
University of Alberta
Edmonton, AB T6G2E8
hassanpo@ualberta.ca

Russell Greiner
Department of Computing Science
University of Alberta
Edmonton, AB T6G2E8
rgreiner@ualberta.ca

Abstract

We consider the challenge of estimating causal effects from observational data; and note that, in general, only some factors based on the observed covariates X contribute to selection of the treatment T, and only some to determining the outcomes Y. We model this by considering three underlying sources of $\{ X, T, Y \}$ and show that explicitly modeling these sources offers great insight to guide designing models that better handle selection bias. This paper is an attempt to conceptualize this line of thought and provide a path to explore it further.

In this work, we propose an algorithm to (1) identify disentangled representations of the above-mentioned underlying factors from any observational dataset $\mathcal{D}$ and (2) leverage this knowledge to reduce, as well as account for, the negative impact of selection bias on estimating the causal effects from $\mathcal{D}$. Our empirical results show that the proposed method (i) achieves state-of-the-art performance in both individual and population based evaluation measures and (ii) is highly robust under various data generating scenarios.

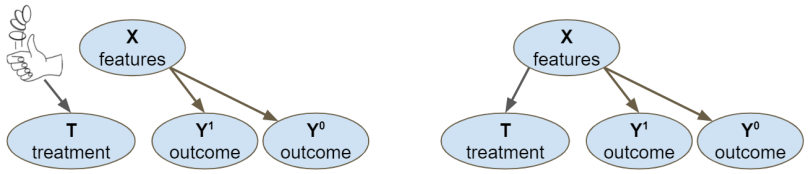
1 Introduction

As we rely more and more on artificial intelligence (AI) to automate the decision making processes, accurately estimating the causal effects of taking different actions gains an essential role. Precision medicine – *i.e.*, the customization of health-care tailored to each individual patient – is a prominent example, that attempts to identify which medical procedure $t \in \mathcal{T}$ will benefit a certain patient x the most, in terms of the treatment outcome $y \in \Re$. Learning such models requires answering counterfactual questions [1, 2] such as: *"Would this patient have lived longer [and by how much], had she received an alternative treatment?"*.

For notation: a dataset $\mathcal{D} = \{ [x_i, t_i, y_i] \}_{i=1}^{N}$ used for causal effect estimation has the following format: for the i^{th} instance (*e.g.*, patient), we have some context information $x_i \in \mathcal{X} \subseteq \Re^K$ (*e.g.*, age, BMI, blood work, etc.), the administered treatment t_i chosen from a set of treatment options $\mathcal{T}$ (*e.g.*, $\{0$: medication, 1: surgery$\}$), and the respective observed outcome $y_i \in \mathcal{Y}$ (*e.g.*, survival time; $\mathcal{Y} \subseteq \Re^+$) as a result of receiving treatment t_i. Note that $\mathcal{D}$ only contains the outcome of the administered treatment (aka *observed* outcome: y_i), but not the outcome(s) of the alternative treatment(s) (aka *counterfactual* outcome(s): y_i^t for $t \in \mathcal{T} \setminus \{t_i\}$), which are inherently unobservable. For the binary-treatment case, we denote the alternative treatment as $\neg t_i = 1 - t_i$.

Pearl [2] demonstrates that, in general, causal relationships can only be learned by experimentation (on-line exploration), or running a Randomized Controlled Trial (RCT), where the treatment assignment does not depend on the individual X – see Figure 1(a). In many cases, however, this is expensive, unethical, or even infeasible. Here, we are forced to approximate causal effects from

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.



(a) Randomized Controlled Trial (b) Observational Study

Figure 1: Belief net structure for randomized controlled trials and observational studies. Here, Y^0 (Y^1) is the outcome of applying $T =$ treatment#0 (#1) to the individual represented by X.

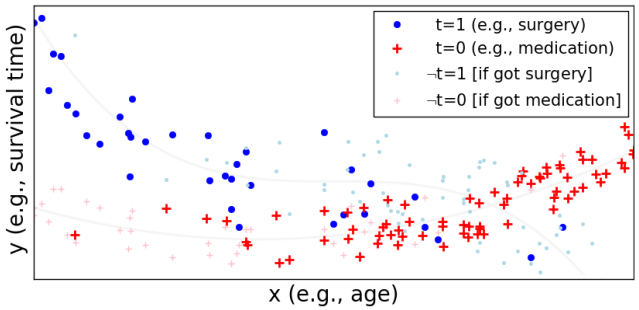


Figure 2: An example observational dataset (best viewed in color). Here, to treat heart disease, a doctor typically prescribes surgery $(t = 1)$ to younger patients (•) and medication $(t = 0)$ to older ones (+). Note that instances with larger (resp., smaller) x values have had a higher chance to be assigned to the $t = 0$ (resp., 1) treatment arm; hence we have selection bias. The counterfactual outcomes – only used for evaluation purpose – are illustrated by small • (+) for $\neg t = 1$ (0).

off-line datasets collected through Observational Studies. In such datasets, the administered treatment T depends on some or all attributes of individual X – see Figure 1(b). Here, as $\Pr(T \mid X) \neq \Pr(T)$, we say these datasets exhibit **selection bias** [3]. Figure 2 illustrates selection bias in an example (synthetic) observational dataset.

Here, we want to accurately estimate the Individual Treatment Effect (ITE) for each instance i – *i.e.*, to estimate $\mathrm{e}_i = y_i^1 - y_i^0$. We frame the solution as learning the function $f : \mathcal{X} \times \mathcal{T} \rightarrow \mathcal{Y}$ that can accurately predict the outcomes (both observed $\hat{y_i}^{t_i}$ as well as counterfactuals $\hat{y_i}^{\neg t_i}$) given the context information x_i for each individual. As mentioned earlier, there are two challenges associated with estimating causal effects:

(i) The fact that counterfactual outcomes are unobservable (*i.e.*, not present in any training data) makes estimating causal effects more difficult than the generalization problem in the supervised learning paradigm.

(ii) Selection bias in observational datasets implies having fewer instances within each treatment arm at specific regions of the domain. This sparsity, in turn, would decrease the accuracy and confidence of predicting counterfactuals at those regions.

This paper is aimed at addressing the second challenge by investigating the root causes of selection bias, by dissecting and identifying the underlying factors that can generate an observational dataset $\mathcal{D}$, and leveraging this knowledge to reduce as well as account for the negative impact of selection bias on estimating the causal effects from $\mathcal{D}$. In this work, we borrow ideas from the representation learning literature [4] in order to reduce selection bias and the domain adaptation literature [5] in order to account for the remainder selection bias that (might) still exist after its reduction.

Our analysis relies on the following assumptions:
Assumption 1: Unconfoundedness [6] – There are no unobserved confounders (*i.e.*, covariates that contribute to both treatment selection procedure as well as determination of outcomes). Formally, $\{Y^t\}_{t \in \mathcal{T}} \perp\!\!\!\perp T \mid X$.
Assumption 2: Overlap [7] – Every individual x has a non-zero chance of being assigned to any

treatment arm. That is, $\Pr(T = t \mid X = x) \neq 0 \quad \forall t \in \mathcal{T},\ \forall x \in \mathcal{X}$.
These two assumptions together are called *strong ignorability* [6]. Imbens and Wooldridge [8] showed that strong ignorability is sufficient for ITE to be identifiable.

Without loss of generality, we assume that the random variable X follows a(n unknown) joint probability distribution $\Pr(X \mid \Gamma, \Delta, \Upsilon)$, treatment T follows $\Pr(T \mid \Gamma, \Delta)$, and outcome Y^T follows $\Pr_T(Y^T \mid \Delta, \Upsilon)$, where Γ, Δ, and Υ represent the three underlying (unobserved) factors[1] that generate an observational dataset $\mathcal{D}$. The respective graphical model is illustrated in Figure 3. Conforming with the statements above, note that the graphical model also suggests that selection bias is induced by factors Γ and Δ, where Δ represents the confounding factors between T and Y.

Main contribution: We argue that explicit identification of the underlying factors $\{\Gamma, \Delta, \Upsilon\}$ in observational datasets offers great insight to guide designing models that better handle selection bias and consequently achieve better performance in terms of estimating ITEs. In this paper, we propose a model, named Disentangled Representations for CounterFactual Regression (DR-CFR), that is optimized to do exactly that. We also present experiments that demonstrate the advantages of this perspective; and show empirically that the proposed method outperforms state-of-the-art models in a variety of data generation scenarios with different dimensionality of factors; see below.

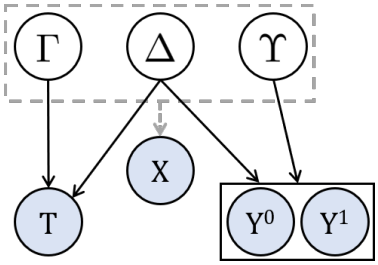


Figure 3: Underlying factors of X; Γ (Υ) are factors that partially determine only T (Y) but not the other random variable; and Δ are confounders; Selection bias is induced by factors Γ and Δ.

2 Related Works

Closely related to the problem of "causal effects estimation from observational studies" is "off-policy learning in contextual bandits" – *cf.*, [9–11]. The goal there is to learn an optimal policy $\pi(t \mid x)$ that decides which action (*i.e.*, treatment) is best for each context (*i.e.*, individual). One way to address this task is through"outcome prediction" – *i.e.*, predicting $y(x, t)\ \ \forall t \in \mathcal{T}$ for each x. This is equivalent to what is done for ITE estimation. Having the estimated outcomes of all possible treatment, the policy is then set to select the one that promises the best outcome. Another strategy bypasses the outcome prediction step and directly obtains the optimal policy by maximizing a utility function – similar to "expected return" in Reinforcement Learning [12]. The majority of approaches under this category belong to the Inverse Propensity Weighting (IPW) family of methods – *cf.*, [13–15], which attempt to balance the source and target distributions by re-weighting certain data instances according to their propensity score $\Pr(t=1 \mid x)$.

Johansson et al. [16] is among the pioneer works that suggested employing representation learning [4] to reduce selection bias – see Figure 4. Shalit et al. [17] present a refined version of [16] in which a common representation space $\Phi(x) = \phi$ is learned by minimizing the discrepancy [18] (hereinafter `disc`) between the conditional distributions of ϕ given $t=0$ versus ϕ given $t=1$. That is,

$$\texttt{disc}\Big(\big\{\Phi(x_i)\big\}_{i:\, t_i=0}, \big\{\Phi(x_i)\big\}_{i:\, t_i=1}\Big) \tag{1}$$

This is a regularization term that attempts to reduce selection bias in the learned representation. On top of this representation learning network, they trained two regression networks $h^t(\phi)$ – one for each treatment arm ($t \in \{0, 1\}$) – that predict the outcomes.

Hassanpour and Greiner [19] argued that the learned representation cannot and should not remove all the selection bias; because the confounders not only contribute to selecting a treatment but also

[1] Examples for: (Γ) rich patients receiving the expensive treatment while poor patients receiving the cheap one – although outcomes of the possible treatments are not particularly dependent on patients' wealth status; (Δ) younger patients receiving surgery while older patients receiving medication; and (Υ) genetic information that determines the efficacy of a medication, however, such relationship is unknown to the attending physician.

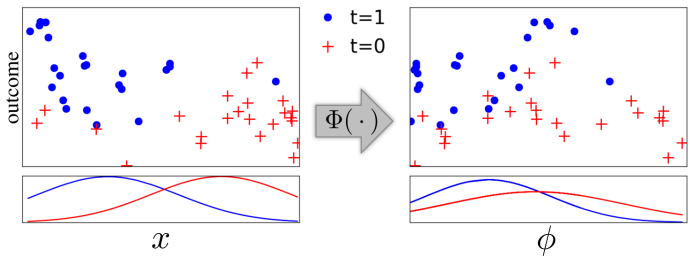


Figure 4: The learned representation has reduced the selection bias. That is, the $t = 1$ and $t = 0$ distributions of the transformed instances $\Phi(\, x\,)$ – here, the distribution of + versus • on the x-axis – are much closer to each other compared to those distributions in the original x space. Also note that the observed outcomes y (on the y-axis) remain unchanged through this transformation.

to determining the respective outcomes. As a result, in case confounding factors do exist (which is a quite viable scenario), even ϕ would exhibit *some* selection bias, although less than that in the original domain x. They built on the work of [17] by introducing *context-aware* importance sampling weights, that attempt to account for the above-mentioned remainder selection bias. These weights are designed to enhance performance of estimating both factual as well as counterfacual outcomes (by the 1 and $\frac{\Pr(\,\phi \,|\, \neg t\,)}{\Pr(\,\phi \,|\, t\,)}$ terms respectively):

$$\omega_i \;=\; 1 + \frac{\Pr(\,\phi_i \,|\, \neg t_i\,)}{\Pr(\,\phi_i \,|\, t_i\,)} \;=\; 1 + \frac{\Pr(\,t_i\,)}{1 - \Pr(\,t_i\,)} \cdot \frac{1 - \pi\big(t_i \,|\, \phi_i\big)}{\pi\big(t_i \,|\, \phi_i\big)} \tag{2}$$

where $\pi\big(t_i \,|\, \phi_i\big)$ is the probability of assigning the observed t_i conditioned on the learned context ϕ_i.

Note that in both [17] and [19], Φ is in fact trying to model the concatenation of factors Δ and Υ (see Figure 3). Although it does make sense that there should be no discrepancy between conditional distributions of Υ, Δ should model the confounding factors, which by definition, must embed some information about treatment assignment. This would result in a positive discrepancy between conditional distributions of Δ that should not be minimized. Thus, minimizing Equation (1) with respect to Φ can do some harm by discarding some of the confounders.

Yao et al. [20] proposed the Similarity preserved Individual Treatment Effect (SITE) method, which is an extension of Shalit et al. [17]'s framework by adding a local similarity preserving component. This component acts as a regularization term, that attempts to retain the same neighbourhood relationships in the learned representation space as exhibited in the original space, by matching the propensity scores $\Pr(\,t{=}1 \,|\, x\,)$ and $\Pr(\,t{=}1 \,|\, \phi\,)$. This, however, results in learning sub-optimal representations when $\Gamma \neq \emptyset$ as SITE enforces to keep dissimilar samples in terms of Γ also far apart in ϕ. In other words, this component penalizes reducing selection bias in ϕ by not discarding the irrelevant information present in Γ even when it does not hurt the outcome estimation at all.

3 Learning Disentangled Representations

Without loss of generality, we assume that any dataset in form of $\{\, X,\, T,\, Y\,\}$ is generated from three underlying (unobserved) factors $\{\Gamma, \Delta, \Upsilon\}$, as illustrated in Figure 3. We argue that explicit identification of these factors offers great insight to guide the design of models that better handle selection bias and consequently achieve better performance in terms of estimating ITEs. Observe that the factor Γ (resp., Υ) partially determines only T (resp., Y), but not the other variables; and Δ includes the confounding factors between T and Y. This graphical model suggests that selection bias is induced by factors Γ and Δ. It also shows that the outcome depends on the factors Δ and Υ. Inspired by this graphical model, our model architecture incorporates the following components:

- Three representation learning networks; one for each underlying factor: $\Gamma(x)$, $\Delta(x)$, and $\Upsilon(x)$.
- Two regression networks; one for each treatment arm: $h^0(\,\Delta(x), \Upsilon(x)\,)$ and $h^1(\,\Delta(x), \Upsilon(x)\,)$.
- Two logistic networks: $\pi_0\big(t \,|\, \Gamma(x), \Delta(x)\big)$ to model the logging policy (aka behaviour policy in the Reinforcement Learning literature – *cf.*, [12]) and $\pi\big(t \,|\, \Delta(x)\big)$ to design weights that account for the confounders' impact.

We therefore try to minimize the following objective function:

$$J(\Gamma, \Delta, \Upsilon, h) = \frac{1}{N}\sum_{i=1}^{N} \omega\big(t_i, \Delta(x_i)\big) \cdot \mathcal{L}\big[\, y_i,\, h^{t_i}\big(\Delta(x_i), \Upsilon(x_i)\big)\,\big] \tag{3}$$

$$+ \alpha \cdot \texttt{disc}\big(\{\Upsilon(x_i)\}_{i:t_i=0}, \{\Upsilon(x_i)\}_{i:t_i=1}\big) \tag{4}$$

$$+ \beta \cdot \frac{1}{N}\sum_{i=1}^{N} -\log\big[\,\pi_0\big(t_i \,|\, \Gamma(x_i), \Delta(x_i)\big)\,\big] \tag{5}$$

$$+ \lambda \cdot \mathfrak{Reg}(h^0, h^1, \pi_0) \tag{6}$$

where $\omega\big(t_i, \Delta(x_i)\big)$ is the re-weighting function; $\mathcal{L}\big[\, y_i,\, h^{t_i}\big(\Delta(x_i), \Upsilon(x_i)\big)\,\big]$ is the prediction loss for observed outcomes (aka factual loss); $\texttt{disc}\big(\{\Upsilon(x)\}_{i:t_i=0}, \{\Upsilon(x)\}_{i:t_i=1}\big)$ calculates the discrepancy between conditional distributions of Υ given $t=0$ versus $t=1$; $-\log \pi_0(\,\cdot\,)$ is the cross entropy loss of predicting the assigned treatments given the learned context; and $\mathfrak{Reg}(\,\cdot\,)$ is the regularization term for penalizing model complexity.

The following sections elaborate on each of these terms.

3.1 Factual Loss: $\mathcal{L}\big[\, y,\, h^t\big(\Delta(x), \Upsilon(x)\big)\,\big]$

Similar to [16, 17, 19, 20], we train two regression networks h^0 and h^1, one for each treatment arm. As guided by the graphical model in Figure 3, the inputs to these networks are the outputs of the $\Delta(x)$ and $\Upsilon(x)$ representation networks. The outputs of these regression networks are the predicted outcomes for their respective treatments.

Note that the prediction loss $\mathcal{L}$ can only be calculated on the observed outcomes (hence the name *factual loss*), as counterfactual outcomes are not available in any training set. This would be an L2-loss for real-valued outcomes and a log-loss for binary outcomes. By minimizing the factual loss, we ensure that the union of the learned representations $\Delta(x)$ and $\Upsilon(x)$ retain enough information needed for accurate estimation of the observed outcomes.

3.2 Cross Entropy Loss: $-\log\big[\,\pi_0\big(t \,|\, \Gamma(x), \Delta(x)\big)\,\big]$

We model the logging policy as a logistic regression network parameterized by $[\,W_0, b_0\,]$ as follows: $\pi_0(\,t \,|\, \psi\,) = \big[1 + e^{-(\,2t-1\,)(\,\psi \cdot W_0 + b_0\,)}\big]^{-1}$, where ψ is the concatenation of matrices Γ and Δ. Minimizing the cross entropy loss enforces learning Γ and Δ in a way that allows $\pi_0(\,\cdot\,)$ to predict the assigned treatments. In other words, the union of the learned representations of Γ and Δ retain enough information to recover the logging policy that guided the treatment assignments.

3.3 Imbalance Loss: $\texttt{disc}\big(\{\Upsilon(x_i)\}_{i:t_i=0}, \{\Upsilon(x_i)\}_{i:t_i=1}\big)$

According to the graphical model in Figure 3, Υ should be independent of T due to the collider structure at Y. Therefore, we should have:

$$\Upsilon \perp\!\!\!\perp T \quad \Longrightarrow \quad \Pr(\,\Upsilon \,|\, T\,) = \Pr(\,\Upsilon\,) \quad \Longrightarrow \quad \Pr(\,\Upsilon \,|\, T=0\,) = \Pr(\,\Upsilon \,|\, T=1\,) \tag{7}$$

We used Maximum Mean Discrepancy (MMD) [21] to calculate dissimilarity between the two conditional distributions of Υ given $t=0$ versus $t=1$.

By minimizing the imbalance loss, we ensure that the learned factor Υ embeds no information about T and all the confounding factors are retained in Δ. Capturing all the confounders in Δ and only in Δ is the hallmark of the proposed method, as we will use it for optimal re-weighting of the factual loss term (next section). Note that this differs from Shalit et al. [17]'s approach in that they do not distinguish between the independent factors Δ and Υ; and minimize the loss defined on only one factor Φ which might erroneously suggest discarding some of the confounders in Δ.

3.4 Re-Weighting Function: $\omega\big(t, \Delta(x)\big)$

We follow Hassanpour and Greiner [19]'s design for weights as re-stated in Equation (2) with the modification that we employ Δ to calculate the weights instead of Φ. Although following the same

model, our weights should perform better in practice than those in [19] due to the following two reasons: (i) no confounders are discarded due to minimizing the imbalance loss (because our `disc` is defined based on Υ, not Φ); and (ii) only the legitimate confounders are used to derive the weights (*i.e.*, Δ), not the ones that we know have not contributed to treatment selection (*i.e.*, Υ).

Also note that this is different from the common practice in re-weighting techniques (*e.g.*, Inverse Propensity Weighting) in that the weights are calculated based on all factors that determine T (*i.e.*, Γ as well as Δ). However, we argue that incorporation of Γ in the weights might result in emphasizing the wrong instances. In other words, since the factual loss $\mathcal{L}$ is only sensitive to factors Δ and Υ, and not Γ, re-weighting $\mathcal{L}$ according to Γ would only confuse the optimization.

4 Experiments

As mentioned earlier, an inherent characteristic of causal effect estimation datasets is that counterfactual outcomes are unobservable, which makes it difficult to evaluate any proposed algorithm. The common solution in the literature is to synthesize datasets where the outcomes of all possible treatments are available. Some entries are then discarded in order to create a proper observational dataset with characteristics (such as selection bias) similar to a real-world one – see for example [22] and [23]. In this work, we use two benchmarks; our synthetic series of datasets as well as a publicly available benchmark. But prior to discussing the benchmarks and results, we introduce the performance measures that we used for evaluation.

4.1 Evaluation Criteria

There are two categories of performance measures for evaluating causal effect estimation algorithms: individual-based and population-based. For individual-based measure, we look at "Precision in Estimation of Heterogeneous Effect": $\text{PEHE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\text{e}}_i - \text{e}_i\right)^2}$ where $\hat{\text{e}}_i = \hat{y}_i^1 - \hat{y}_i^0$ is the predicted effect and $\text{e}_i = y_i^1 - y_i^0$ is the true effect. For population-based measure, we look at bias of the "Average Treatment Effect (ATE)": $\epsilon_{\text{ATE}} = \left|\text{ATE} - \widehat{\text{ATE}}\right|$ where $\text{ATE} = \frac{1}{N}\sum_{i=1}^{N} y_i^1 - \frac{1}{N}\sum_{j=1}^{N} y_j^0$ in which y_i^1 and y_j^0 are the true outcomes for the treatment and control arms respectively[2] and $\widehat{\text{ATE}}$ is calculated based on the estimated outcomes.

4.2 Results and Discussions

In this paper, we compare these ITE estimation methods:

- **CFR**: CounterFactual Regression [17].
- **CFR-ISW**: CounterFactual Regression with Importance Sampling Weights [19].
- **SITE**: Similarity preserved Individual Treatment Effect [20].
- **DR-CFR**: Disentangled Representations for CounterFactual Regression – our proposed method.

Note that all four methods share the same core code-base – that of CFR; developed by Johansson et al. [16] and Shalit et al. [17]. Therefore, they share a lot in terms of model architecture. In order to allow for fair comparison, we searched the hyperparameter space such that different methods enjoy the same model complexity. For example, if DR-CFR gets to train 3 representation networks each with k hidden units, the other methods would get $3\times k$ hidden units for their 1 representation network.

Below, we explain the characteristics of the two benchmarks used for evaluation. We also discuss the performance of the proposed method and compare it with its contenders.

4.2.1 Synthetic Datasets

The 2019 Atlantic Causal Inference Conference held a related data challenge and graciously published their code-base for generating the datasets after the challenge was over [24]. Upon reviewing their

[2] We can calculate ATE here since the data is synthetic and so both observed and counterfactual outcomes are available. In RCTs, the Sample Average Treatment Effect (SATE) $= \frac{1}{N_1}\sum_{i=1}^{N_1} y_i^1 - \frac{1}{N_0}\sum_{j=1}^{N_0} y_j^0$ is used as a proxy for the true ATE, where N_1 (N_0) is the number of treated (controlled) subjects and y_i^1 (y_j^0) is the outcome of subject i (j) upon being treated (controlled).

code, we noticed that there was no guarantee that it would synthesize any datasets with significant presence of either Γ or Υ. Therefore, we decided to generate our own synthetic datasets according to the following process:

- Select sample size N
- Select dimensionality m_L of each latent factor $L \in \{\Gamma, \Delta, \Upsilon, \Xi\}$, where Ξ is the noise factor (*i.e.*, determines neither T nor Y).
- Sample each L from $\mathcal{N}(\mu, \Sigma)$, where μ and Σ are of size $m_L \times 1$ and $m_L \times m_L$ respectively.
  - Concatenate Γ, Δ, Υ, and Ξ to make X [of size $N \times (m_\Gamma + m_\Delta + m_\Upsilon + m_\Xi)$]
  - Concatenate Γ and Δ to make Ψ [of size $N \times (m_\Gamma + m_\Delta)$]
  - Concatenate Δ and Υ to make Φ [of size $N \times (m_\Delta + m_\Upsilon)$]
- For treatment t:
  - Sample $m_\Gamma + m_\Delta$ coefficients θ from $\mathcal{N}(0, 1)$
  - Define the logging policy as $\pi_0(t=1|z) = \frac{1}{1+\exp(-\zeta z)}$, where $z = \Psi \cdot \theta$ and ζ is a scalar that determines the slope of the logistic curve.
  - Sample treatment t_i for instance x_i from the Bernoulli distribution with parameter $\pi_0(t=1|z_i)$
- For outcomes y^0 and y^1:
  - Sample $m_\Delta + m_\Upsilon$ coefficients ϑ^0 and ϑ^1 from $\mathcal{N}(0, 1)$
  - Define $y^0 = \Phi \cdot \vartheta^0 + \varepsilon$ and $y^1 = (\Phi \circ \Phi) \cdot \vartheta^1 + \varepsilon$, where ε is a white noise sampled from $\mathcal{N}(0, 0.1)$ and $\circ$ is symbol for element-wise (Hadamard/Schur) product.

We set the dimension of noise $m_\Xi = 1$, and considered all the viable datasets in a mesh generated by $m_\Gamma, m_\Delta, m_\Upsilon \in \{0, 4, 8\}$. This creates 24 scenarios[3] that exhaust all possible situations in terms of dominance of each factor Γ, Δ, and Υ over the others. We synthesized five datasets with various initial random seeds for each scenario in order to allow for significance testing.

Figure 5 visualizes the PEHE measures in radar charts for the CFR, CFR-ISW, SITE, and DR-CFR methods, trained with datasets of size $N = 2{,}500$ (left) and $N = 10{,}000$ (right). As expected, all methods perform better with observing more training data; however, DR-CFR took the most advantage by reducing PEHE the most (by 0.026, going down from 0.101 to 0.075), while CFR, CFR-ISW, and SITE reduced PEHE by 0.021, 0.22, and 0.013 respectively. We also observe that, under all scenarios, DR-CFR's performance is smoother than the other methods. For example, considering all the seven scenarios with $m_\Gamma + m_\Delta + m_\Upsilon = 12$, the standard deviation of PEHE is smallest for DR-CFR (0.007), while it is double or more for CFR, CFR-ISW, and SITE. This robust behaviour of DR-CFR is of great value because the dimensionality of the underlying factors is never known.

Table 1 summarizes the PEHE and ϵ_{ATE} measures (lower is better) for all scenarios, in terms of mean and standard deviation of all the 24×5 datasets, in order to give a unified view on the performance. DR-CFR achieves the best performance among the contending methods. These results are statistically significant based on the Welch's unpaired t-test with $\alpha = 0.05$).

4.2.2 Infant Health and Development Program (IHDP)

The original RCT data was designed to evaluate the effect of specialist home visits on future cognitive test scores of premature infants. Hill [25] induced selection bias by removing a non-random subset of the treated population to create a realistic observational dataset. The resulting dataset contains 747 instances (608 control, 139 treated) with 25 covariates that measure different attributes of infants and their mothers. We run our experiments on the same benchmark (100 realizations of outcomes) provided by and used in [16, 17]. The outcomes of this semi-synthetic benchmark are simulated according to the response surface "A" of the Non-Parametric Causal Inference (NPCI) package [26]. Similar to the literature, the noiseless outcomes (available for evaluation purpose only) are used to compute the true ITEs.

Table 2 summarizes the PEHE and ϵ_{ATE} measures (lower is better) on the IHDP benchmark. The results are reported in terms of mean and standard deviation over the 100 datasets with various realizations of outcomes. Again, DR-CFR achieves the best performance among the contending methods.

[3] The following three tuples are not viable: $(0, 0, 0)$, $(4, 0, 0)$, and $(8, 0, 0)$. The reason for the first is trivial, and the reason for the second and third is that the outcomes would be pure noise.

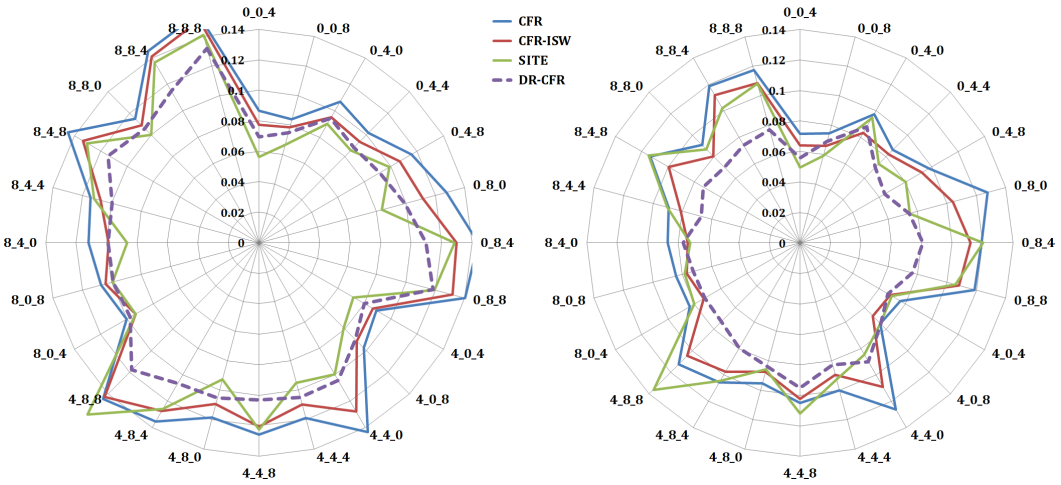


Figure 5: Radar charts for visualizing the PEHE performance results on the synthetic datasets. Training sample size on the left chart is 2,500 and on the right chart is 10,000. Each vertex on the polygons is named after the factors' dimension sequence ($m_\Gamma\_m_\Delta\_m_\Upsilon$) of the respective group of datasets. The polygons' radii are scaled between $0:0.14$ to quantify the PEHE values (*i.e.*, the closer to the centre, the smaller the PEHE). The dashed purple curve illustrates the results of the proposed method.

Table 1: Synthetic datasets
(24×5 with $N=10{,}000$)

Methods	PEHE	ϵ_{ATE}
CFR	0.099 (0.017)	0.012 (0.007)
CFR-ISW	0.089 (0.016)	0.009 (0.004)
SITE	0.090 (0.020)	0.013 (0.008)
DR-CFR	**0.075 (0.009)**	**0.005 (0.003)**

Table 2: IHDP datasets
(100 with $N=747$)

Methods	PEHE	ϵ_{ATE}
CFR	0.81 (0.30)	0.13 (0.12)
CFR-ISW	0.73 (0.28)	0.11 (0.10)
SITE	0.73 (0.33)	0.10 (0.09)
DR-CFR	**0.65 (0.37)**	**0.03 (0.04)**

PEHE and ϵ_{ATE} measures (lower is better) represented in the form of "mean (standard deviation)".

5 Future Works and Conclusion

The majority of methods proposed to estimate causal effects – including this work – fall under the category of discriminative approaches. A promising direction is to consider developing generative models, in an attempt to shed light on the true underlying data generating mechanism. Perhaps, this could also facilitate generating new, virtual, yet realistic data instances – similar to what is done in computer vision. Louizos et al. [27]'s method is a notable generative approach, which uses Variational Auto-Encoder (VAE) to extract latent confounders from their observed proxies. While that work is an interesting step in that direction, it is not yet capable of addressing the problem of selection bias. We believe that our proposed perspective on the problem can be helpful to solve this open question. This is left to future work.

In this paper, we studied the problem of estimating causal effect from observational studies. We argued that not all factors in the observed covariates X might contribute to the procedure of selecting treatment T, or more importantly, determining the outcomes Y. We modeled this using three underlying sources of X, T, and Y, and showed that explicit identification of these sources offers great insight to help us design models that better handle selection bias in observational datasets. We proposed an algorithm, Disentangled Representations for CounterFactual Regression (DR-CFR), that can (1) identify disentangled representations of the above-mentioned underlying sources and (2) leverage this knowledge to reduce as well as account for the negative impact of selection bias on estimating the causal effects from observational data. Our empirical results showed that the proposed method (i) achieves state-of-the-art performance in both individual and population based evaluation measures and (ii) is highly robust under various data generating scenarios.

References

[1] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[2] Judea Pearl. *Causality*. Cambridge University Press, 2009.

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

[4] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE TPAMI*, 35(8):1798–1828, 2013.

[5] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2), 2000.

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

[7] Guido W Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29, 2004.

[8] Guido W Imbens and Jeffrey M Wooldridge. Recent developments in the econometrics of program evaluation. *Journal of economic literature*, 47(1):5–86, 2009.

[9] Alex Strehl, John Langford, Lihong Li, and Sham M Kakade. Learning from logged implicit exploration data. In *NeurIPS*, pages 2217–2225. 2010.

[10] Léon Bottou, Jonas Peters, Joaquin Quinonero Candela, Denis Xavier Charles, Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Y Simard, and Ed Snelson. Counterfactual reasoning and learning systems: the example of computational advertising. *JMLR*, 14(1), 2013.

[11] Adith Swaminathan and Thorsten Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. *JMLR*, 16, 2015.

[12] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*, volume 1. MIT Press Cambridge, 1998.

[13] Peter C Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424, 2011.

[14] Adith Swaminathan and Thorsten Joachims. The self-normalized estimator for counterfactual learning. In *NeurIPS*, 2015.

[15] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *Proceedings of the 33rd ICML - Volume 48*, 2016.

[16] Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *ICML*, pages 3020–3029, 2016.

[17] Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *ICML*, pages 3076–3085, 2017.

[18] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009.

[19] Negar Hassanpour and Russell Greiner. Counterfactual regression with importance sampling weights. In *IJCAI*, pages 5880–5887, 7 2019.

[20] Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. In *Advances in Neural Information Processing Systems*, pages 2633–2643, 2018.

[21] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *JMLR*, 13(Mar):723–773, 2012.

[22] Negar Hassanpour and Russell Greiner. A novel evaluation methodology for assessing off-policy learning methods in contextual bandits. In *Canadian AI*, pages 31–44, 2018.

[23] Alina Beygelzimer and John Langford. The offset tree for learning with partial labels. In *Proceedings of the 15th ACM SIGKDD*. ACM, 2009.

[24] Susan Gruber, Geneviève Lefebvre, Tibor Schuster, and Alexandre Piché. Atlantic Causal Inference Conference (ACIC) Data Challenge, 2019. `https://sites.google.com/view/ACIC2019DataChallenge/data-challenge`.

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

[26] Vincent Dorie. NPCI: Non-parametrics for causal inference, 2016. `https://github.com/vdorie/npci`.

[27] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *NeurIPS*, pages 6446–6456. 2017.