LEARNING DISENTANGLED REPRESENTATIONS FOR COUNTERFACTUAL REGRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

We consider the challenge of estimating treatment effects from observational data; and point out 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 given 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 treatment 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.

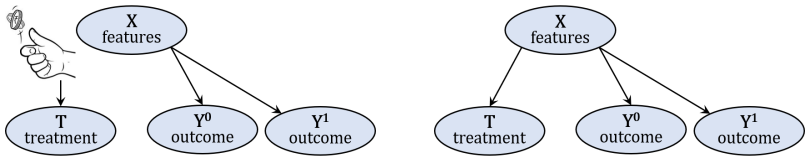
1 INTRODUCTION

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. A prominent example is precision medicine – *i.e.*, the customization of health-care tailored to each individual patient – which attempts to identify which medical procedure $t \in \mathcal{T}$ will benefit a certain patient x the most, in terms of the treatment outcome $y \in \mathbb{R}$. Learning such models requires answering counterfactual questions (Rubin, 1974; Pearl, 2009) 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 treatment effect estimation has the following format: for the ith instance (*e.g.*, patient), we have some context information $x_i \in \mathcal{X} \subseteq \mathbb{R}^K$ (*e.g.*, age, BMI, blood work, etc.), the administered treatment t_i chosen from a set of treatment options $\mathcal{T}$ (*e.g.*, $\{0\colon \text{medication}, 1\colon \text{surgery}\}$), and the respective observed outcome $y_i \in \mathcal{Y}$ (*e.g.*, survival time; $\mathcal{Y} \subseteq \mathbb{R}^+$) 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 (2009) 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 treatment effects from 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** (Imbens & Rubin, 2015). 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 $f : \mathcal{X} \times \mathcal{T} \to \mathcal{Y}$ 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



(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.

context information x_i for each individual. As mentioned earlier, there are two challenges associated with estimating treatment effects:

(i) The fact that counterfactual outcomes are unobservable (*i.e.*, not present in any training data) makes estimating treatment 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 addresses 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 treatment effects from $\mathcal{D}$. In this work, we borrow ideas from the representation learning literature (Bengio et al., 2013) in order to reduce selection bias and from the domain adaptation literature (Shimodaira, 2000) in order to account for the remainder selection bias that (might) still exist after its reduction.

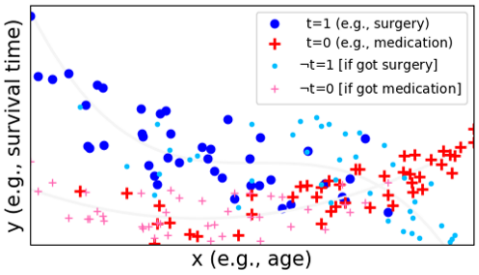


Figure 2: An example observational dataset. Here, to treat heart disease, a doctor typically prescribes surgery $(t=1)$ to younger patients (•) and medication $(t=0)$ to older ones (+). Note that instances with larger (resp., smaller) x values have a higher chance to be assigned to the $t=0$ (resp., 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).

Our analysis relies on the following assumptions:
Assumption 1: Unconfoundedness (Rosenbaum & Rubin, 1983) – 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 (Imbens, 2004) – Every individual x should have 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* (Rosenbaum & Rubin, 1983). Imbens & Wooldridge (2009) 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 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.

[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.

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.

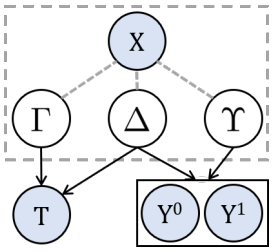


Figure 3: [R1: [New figure]] Underlying factors of X; Γ (Υ) are factors that partially determine only T (Y) but not the other random variable; and Δ are confounders; Selection bias is induced by factors Γ and Δ.

2 Related Works

Selection bias in observational datasets is equivalent to a domain adaptation scenario where a model is trained on a "source" (observed) data distribution, but should perform well on a "target" (counterfactual) one. [R2: Learning treatment effects from observational datasets is closely related to "off-policy learning from logged bandit feedback" – *cf.*, (Swaminathan & Joachims, 2015a), whose goal is learning an optimal policy that selects the best personalized treatment for each individual.] A common statistical solution is re-weighting certain data instances to balance the source and target distributions. The majority of re-weighting approaches belong to the Inverse Propensity Weighting (IPW) family of methods – *cf.*, (Austin, 2011; Bottou et al., 2013; Swaminathan & Joachims, 2015c). [R2: While IPW methods are unbiased, they suffer from high variance. Swaminathan & Joachims (2015b) proposed the Counterfactual Risk Minimization (CRM) principle to alleviate this issue.] In summary, re-weighting is an attempt to **account for** the selection bias.

Johansson et al. (2016) is among the pioneer works that explored ways to use techniques from representation learning (Bengio et al., 2013) to **reduce** the selection bias. Shalit et al. (2017) present a refined version of (Johansson et al., 2016)'s method that learns a common representation space $\Phi(x) = \phi$ by minimizing the discrepancy (Mansour et al., 2009) (hereinafter "`disc`") between the conditional distributions of ϕ given $t{=}0$ versus ϕ, given $t{=}1$. That is,

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

which is (effectively) 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 & Greiner (2019) argued that the learned representation cannot and should not remove all the selection bias, as the confounders not only contribute to choosing a treatment but also to determining the respective outcomes.[c2] As a result, where there are confounders (which is a common situation), even ϕ would exhibit *some* selection bias, although less than that in the original domain x. They built on the work of (Shalit et al., 2017) by introducing *context-aware* importance sampling weights, that attempt to account for the above-mentioned remainder selection bias. These weights

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

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

[c2 R3: While Hassanpour & Greiner (2019) presented a graphical model similar to our Figure 3, they only used it to investigate the nature of selection bias. *N.b.*, they did not implement the idea of learning disentangled representations for counterfactual regression; instead, their method [like (Shalit et al., 2017)] learns a common representation ϕ that can represent *only* the confounders, but not the other factors. Our approach extends theirs by providing an algorithm that can learn disentangled representations of the underlying factors from observational datasets.]

Note that both (Shalit et al., 2017) and (Hassanpour & Greiner, 2019) use Φ 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, the Δ factor 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 lead to problematic results as it discards some of the confounders.

Yao et al. (2018) proposed the Similarity preserved Individual Treatment Effect (SITE) method, which extends Shalit et al. (2017)'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 $\Pr(\, t{=}1 \mid x\,)$ and $\Pr(\, t{=}1 \mid \phi\,)$. This, however, results in learning sub-optimal representations when $\Gamma \neq \emptyset$ as SITE tries to keep instances whose Γs are far apart, 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.

RK: Our work has many similarities to (Kuang et al., 2017), who decomposed X into two subsets: confounding and adjustment variables, which are similar to our Δ and Υ factors respectively. They then used an optimization algorithm for identifying these variables, to ultimately find an unbiased estimate of the Average Treatment Effect (ATE). We extend their work in three ways: (i) In addition to confounders and adjustment variables, we also identify the factors that determine the treatment and have no effect on the outcome (*i.e.*, Γ). (ii) Unlike Kuang et al. (2017) that takes a linear approach by tagging the raw features as either confounders or adjustment variables, our proposed method has the capacity to learn [non-linear] representations of the underlying factors. (iii) Our method facilitates estimating both ATE as well ITE, whereas (Kuang et al., 2017) cannot provide estimates of ITEs.

3 Learning Disentangled Representations

We assume, without loss of generality, that any dataset of the form $\{\, X,\, T,\, Y \,\}$ is generated from three underlying factors $\{\, \Gamma, \Delta, \Upsilon \,\}$, as illustrated in Figure 3. [c1] 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: $\pi_0\big(t \mid \Gamma(x), \Delta(x)\big)$ to model the logging policy – aka behaviour policy in Reinforcement Learning; *cf.*, (Sutton & Barto, 1998) – and $\pi\big(t \mid \Delta(x)\big)$ to design weights that account for the confounders' impact.

We therefore try to minimize the following objective function:

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

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

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

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

[c1] *RJ:* Note that the assumption of unconfoundedness still holds; here is why:
Short: Observing either X or Δ blocks the path from T to Y, which supports the unconfoundedness assumption.
Long: Once the representation networks are learned from the observational data, we can compute the latent factors $\{\, \Gamma, \Delta, \Upsilon \,\}$ from X only. Therefore, although these factors are not explicitly observed, they are effectively observed, in that they are derived directly from the observed X, and so should not be categorized as "unobserved confounders". For example, the latent factor for "zip code" in X is "socio-economic status" (perhaps in Δ). In other words, "socio-economic status" can be inferred from "zip code" which can be viewed as a proxy for it.

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 given $t=1$; $-\log \pi_0(\,\cdot\,)$ 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 (Johansson et al., 2016; Shalit et al., 2017; Hassanpour & Greiner, 2019; Yao et al., 2018), 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 regression networks are the outputs of the $\Delta(x)$ and $\Upsilon(x)$ representation networks and their outputs 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 Figure 3, Υ should be independent of T due to the collider structure at Y. Therefore,

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

We used Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) 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. (2017)'s approach in that they do not distinguish between the independent factors Δ and Υ; and minimizing 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 & Greiner, 2019)'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 design, we anticipate our weights should perform better in practice than those in (Hassanpour & Greiner, 2019) as: (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 have not contributed to treatment selection (*i.e.*, Υ).

Notably, the weights design in Equation (2) is different from the common practice in re-weighting techniques (*e.g.*, IPW) 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 yield a wrong objective function to be optimized.

4 EXPERIMENTS

4.1 BENCHMARKS

Evaluating treatment effect estimation methods is problematic on real-world datasets since, as mentioned earlier, their counterfactual outcomes are inherently unobservable. A common solution is to synthesize datasets where the outcomes of all possible treatments are available, then discard some outcomes to create a proper observational dataset with characteristics (such as selection bias) similar to a real-world one – *cf.*, (Beygelzimer & Langford, 2009). In this work, we use two such benchmarks: our synthetic series of datasets as well as a publicly available benchmark.

4.1.1 SYNTHETIC DATASETS

We generated our synthetic datasets according to the following process, which takes an input the sample size N; dimensionalities $[m_\Gamma, m_\Delta, m_\Upsilon] \in \mathcal{Z}^{+(3)}$; for each factor $L \in \{\Gamma, \Delta, \Upsilon\}$, means and covariance matrices (μ_L, Σ_L); and a scalar ζ that determines the slope of the logistic curve.

- For each latent factor $L \in \{\Gamma, \Delta, \Upsilon\}$
  - Sample L from $\mathcal{N}(\mu_L, \Sigma_L)$,
  - Concatenate Γ, Δ, and Υ to make the covariates matrix X [of size $N \times (m_\Gamma + m_\Delta + m_\Upsilon)$]
  - 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 $m_\Gamma + m_\Delta$ tuple of coefficients θ from $\mathcal{N}(0,1)^{m_\Gamma + m_\Delta}$
  - Define the logging policy as $\pi_0(t=1|z) = \frac{1}{1+\exp(-\zeta z)}$, where $z = \Psi \cdot \theta$.
  - Sample treatment t_i for instance x_i from a Bernoulli distribution with parameter $\pi_0(t=1|z_i)$.
- For outcomes Y^0 and Y^1:
  - Sample $m_\Delta + m_\Upsilon$ tuple of coefficients ϑ^0 and ϑ^1 from $\mathcal{N}(0,1)^{m_\Delta + m_\Upsilon}$
  - 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 considered all the viable datasets in a mesh generated by $m_\Gamma, m_\Delta, m_\Upsilon \in \{0, 4, 8\}$. This creates 24 scenarios[c0] that consider all possible situations in terms of the relative sizes of the factors Γ, Δ, and Υ. We synthesized five datasets with various initial random seeds for each scenario.

4.1.2 INFANT HEALTH AND DEVELOPMENT PROGRAM (IHDP)

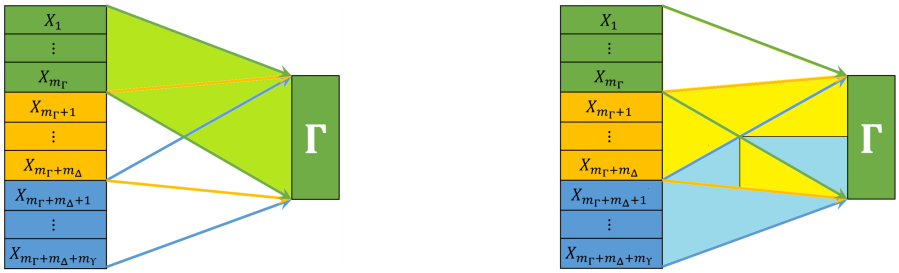
The original RCT data was designed to evaluate the effect of specialist home visits on future cognitive test scores of premature infants. Hill (2011) induced selection bias by removing a non-random subset of the treated population to create a realistic observational dataset. The resulting dataset contains 747 instances (608 control, 139 treated) with 25 covariates. We run our experiments on the same benchmark (100 realizations of outcomes) provided by and used in (Johansson et al., 2016; Shalit et al., 2017). Outcomes of this semi-synthetic benchmark were simulated according to response surfaces provided in the Non-Parametric Causal Inference (NPCI) package (Dorie, 2016).

4.2 RESULTS AND DISCUSSIONS

4.2.1 EVALUATING IDENTIFICATION OF FACTORS $\{\Gamma, \Delta, \Upsilon\}$

First, we want to determine if the proposed method is able to identify the variables that belong to each underlying factor. To do so, we look at the weight matrix in the first layer of each representation network, which is of size $(m_\Gamma + m_\Delta + m_\Upsilon) \times K$, where K is the number of neurons in the first hidden layer of the respective representation network. For example, to check if Γ is identified properly, we partition the weights matrix into two slices, as shown in Figure 4, and calculate the average of each slice. The first slice [referred to as S_Γ; highlighted in Figure 4(a)] pertains to "Γ's ground truth

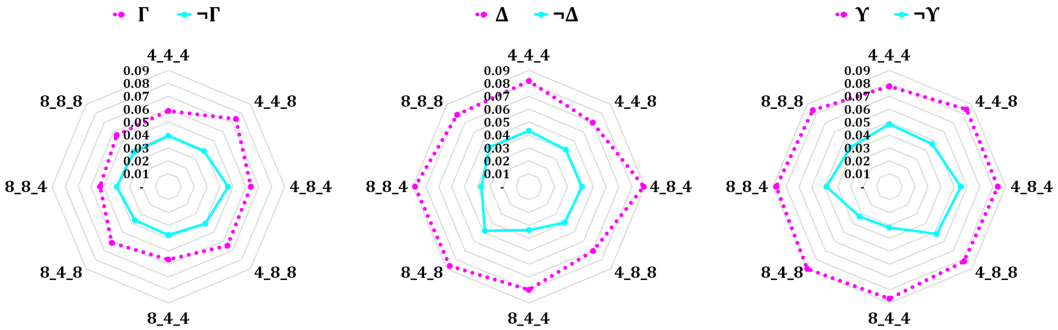
[c0] There are not $2^3 = 27$ scenarios because we removed the three tuples: $(0,0,0)$, $(4,0,0)$, and $(8,0,0)$, as any scenario with $\Delta = \Upsilon = \emptyset$ would generate outcomes that are pure noise.



(a) Slice of the weights matrix that connects {the variables in X belonging to Γ} to {the first layer of the representation network that attempts to identify Γ}. The size of this slice is $m_\Gamma \times K$.

(b) Slice of the weights matrix that connects {the variables in X *not* belonging to Γ} to {the first layer of the representation network that attempts to identify Γ}. The size of this slice is $(m_\Delta + m_\Upsilon) \times K$.

Figure 4: Visualization of slicing the learned weights matrix in the first layer of the representation network (number of neurons: K) for identifying Γ (best viewed in color).



(a) Identification of Γ (b) Identification of Δ (c) Identification of Υ

Figure 5: Radar charts that visualize the capability of DR-CFR in identifying the underlying factors Γ, Δ, and Υ. Each vertex on the polygons is named after the factors' dimension sequence (m_Γ_m_Δ_m_Υ) of the respective synthetic dataset. The polygons' radii are scaled between $0:0.09$ and quantify the average weights of the first slice (in dotted magenta) and the second slice (in cyan).

variables in X" and the second slice [$\mathrm{S}_{\neg\Gamma}$; Figure 4(b)] pertains to "variables in X that do not belong to Γ". Constructing S_Δ, $\mathrm{S}_{\neg\Delta}$, S_Υ, and $\mathrm{S}_{\neg\Upsilon}$ follow a similar procedure.

If the proposed method achieves a good identification, then we expect the average of the absolute values of weights in S_Γ should be higher than that of $\mathrm{S}_{\neg\Gamma}$; this same claim should hold for $(\mathrm{S}_\Delta, \mathrm{S}_{\neg\Delta})$ and $(\mathrm{S}_\Upsilon, \mathrm{S}_{\neg\Upsilon})$ as well. Note that only the relative relationships between the average weights in either of the slices matter; since this analysis is aimed at checking whether, for example, for identifying Γ, its respective representation network has indeed learned to emphasize on "Γ's ground truth variables in X" more than the other variables in X. Figure 5 illustrates the identification performance of DR-CFR according to this analysis; showing empirically that the proposed method successfully identifies all the three underlying factors, for all synthetic datasets.

4.2.2 Evaluating Estimation of Treatment Effects

Given a synthetic dataset (that include both factual as well as counterfactual outcomes), one can evaluate treatment effect estimation methods with two types of performance measures:

- Individual-based: "Precision in Estimation of Heterogeneous Effect" $\mathrm{PEHE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{\mathrm{e}}_i - \mathrm{e}_i)^2}$ where $\hat{\mathrm{e}}_i = \hat{y}_i^1 - \hat{y}_i^0$ is the predicted effect and $\mathrm{e}_i = y_i^1 - y_i^0$ is the true effect.

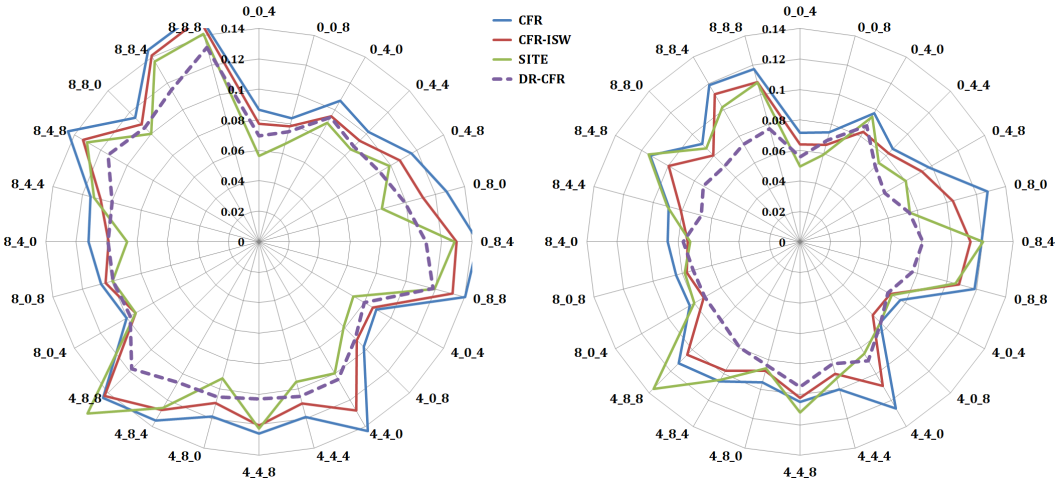


Figure 6: Radar charts for visualizing the PEHE performance results on the synthetic datasets. Training sample size on the left chart is $2{,}500$ and on the right chart is $10{,}000$. Each vertex on the polygons is named after the factors' dimension sequence (m_Γ_m_Δ_m_Υ) of the respective group of 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.

- Population-based: "Bias of the Average Treatment Effect" $\epsilon_{\text{ATE}} = \left|\text{ATE} - \widehat{\text{ATE}}\right|$ 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 respective treatments and $\widehat{\text{ATE}}$ is calculated based on the estimated outcomes.

In this paper, we compare performances of the following treatment effect estimation methods: [c0]

- **CFR**: CounterFactual Regression (Shalit et al., 2017).
- **CFR-ISW**: CFR with Importance Sampling Weights (Hassanpour & Greiner, 2019).
- **SITE**: Similarity preserved Individual Treatment Effect (Yao et al., 2018).
- **DR-CFR**: Disentangled Representations for CFR – our proposed method.

Figure 6 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 $\alpha = 0.05$. Table 2 summarizes the PEHE and ϵ_{ATE} measures 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 (statistically significant for ϵ_{ATE}) among the contending methods.

[c0] Note that all four methods share the same core code-base: that of CFR; developed by Johansson et al. (2016) and Shalit et al. (2017). 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 three representation networks each with K hidden units, the other methods would get $3 \times K$ hidden units for their one representation network.

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 treatment 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. (2017)'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 treatment 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 treatment 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

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.

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

Alina Beygelzimer and John Langford. The offset tree for learning with partial labels. In *ACM SIGKDD*. ACM, 2009.

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.

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

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

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

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

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

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

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

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

Kun Kuang, Peng Cui, Bo Li, Meng Jiang, Shiqiang Yang, and Fei Wang. Treatment effect estimation with data-driven variable decomposition. In *AAAI*, 2017.

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

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

Judea Pearl. *Causality*. Cambridge University Press, 2009.

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

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

Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: Generalization bounds and algorithms. In *ICML*, pp. 3076–3085, 2017.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning And Inference*, 90(2), 2000.

Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*, volume 1. MIT Press Cambridge, 1998.

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

Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *ICML*, 2015b.

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

Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. In *NeurIPS*, pp. 2633–2643, 2018.