Validating Causal Inference Methods

Harsh Parikh[1]   Carlos Vajao[2]   Louise Xu[2]   Eric Tchetgen Tchetgen[3]

Abstract

The fundamental challenge of drawing causal inference is that counterfactual outcomes are not fully observed for any unit. Furthermore, in observational studies, treatment assignment is likely to be confounded. Many statistical methods have emerged for causal inference under unconfoundedness conditions given pre-treatment covariates, including: propensity score-based methods, prognostic score-based methods, and doubly robust methods. Unfortunately for applied researchers, there is no 'one-size-fits-all' causal method that can perform optimally universally.In practice, causal methods are primarily evaluated quantitatively on handcrafted simulated data. Such data-generative procedures can be of limited value because they are typically stylized models of reality. They are simplified for tractability and lack the complexities of real-world data. For applied researchers, it is critical to understand how well a method performs for data at hand. Our work introduces a deep generative model-based framework, *Credence*, to validate causal inference methods. The framework's novelty stems from its ability to generate synthetic data anchored at the empirical distribution for the observed sample, and therefore virtually indistinguishable from the latter. The approach allows the user to specify ground truth for the form and magnitude of causal effects and confounding bias as functions of covariates. Thus simulated data sets are used to evaluate the potential performance of various causal estimation methods when applied to data similar to the observed sample. We demonstrate Credence's ability to accurately assess the relative performance of causal estimation techniques in an extensive simulation study and two real-world data applications from Lalonde and Project STAR studies.

[1]Duke University, Durham NC, USA [2]Amazon.com, Seattle WA, USA [3]The Wharton School, University of Pennsylvania, Philadelphia PA, USA. Correspondence to: Harsh Parikh <harsh.parikh@duke.edu>.

1. Introduction

Causal inference problems typically focus on estimating the effect of a point treatment (Z) on an outcome of interest (Y) (Angrist and Pischke, 2008) - one example would be determining if a drug will aid in a patient's recovery. In a binary treatment setting, the average causal effect (τ) of an intervention is defined as the contrast between the population average potential outcome if all units were given the active treatment ($\mathbf{E}[Y(1)]$) versus the average potential outcome if all units were given the control treatment ($\mathbf{E}[Y(0)]$). In this paper, the contrast of interest is defined on the additive scale, that is $\tau = \mathbf{E}[Y(1) - Y(0)]$, which hereafter is referred to as the average treatment effect or ATE.

We will proceed under the standard consistency assumption typically made in the causal inference literature, that for each unit with observed treatment value, say $Z = z$, its observed outcome matches the potential outcome $Y(z)$ while the counterfactual outcome $Y(1-z)$ remains unobserved or *missing*. Hence, knowing the true causal effect of intervention for each individual is impossible, however, it is possible under certain conditions to learn the population average causal effect τ. A standard approach to identify the ATE relies on an assumption that one can account for the possibility of confounding, on the basis of measure covariates, therefore ruling out the presence of unmeasured confounding. Confounding factors are typically common causes of the treatment and outcome variables, for instance, patients that were able to avail themselves of a treatment drug might be affluent and have better access to healthcare facilities compared to patients in the control group. Thus, the correlation between the treatment and the outcome might be primarily because of better access to healthcare. In order to evaluate the ATE, variables such as "access to healthcare" in this case, that simultaneously affect the treatment (Z) and the outcome (Y) are confounders (W) that must be accounted for. In addition to consistency and unconfoundedness, nonparametric identification of the ATE typically also requires a positivity assumption that for any covariate value observed in the population, there is at least one unit who would receive the active treatment and at least one unit

that would receive the control treatment.

Causal effect estimation. Under these identifying conditions, the causal inference literature has over the years produced several alternative approaches to explicitly account for observed covariates with the goal to adjust for confounding. This rich analytical arsenal includes three types of methods distinct in the manner in which covariate adjustment is achieved: (i) methods that directly model the dependence of the outcome variable on covariates, such as nearest neighbor matching or outcome regression adjustment; (ii) methods that model dependence of the treatment variable on covariates (also known as treatment propensity score), such as propensity score matching or inverse-probability-of-treatment-weighting; and (iii) so-called doubly robust methods that combine both approaches for improved robustness (Rosenbaum and Rubin, 1984; Imbens and Rubin, 2015; Hernán and Robins, 2010; Van der Laan et al., 2003; Van der Laan and Rose, 2018). Conditions under which methods (i)-(iii) can provide valid inferences about ATE are well understood, and typically require that bias incurred by the approach in estimating nuisance parameters is sufficiently small. Such bias will be small if nuisance functions of covariates, e.g the conditional mean for the outcome in the control group as a function of covariates or the propensity score, are smooth enough to be estimated at sufficiently fast rates. However, performance of these methods might vary significantly for any given finite size data-sample. For instance, for the well-known Lalonde's National Support Work Demonstration (NSW) temporary employment program and income dataset (LaLonde, 1986) where the treatment is randomly assigned, the average treatment effect (ATE) estimates using different methods are vastly different (see Table 1). Each causal inference estimation method has both strengths and limitations, and the optimal choice of estimation can depend on the problem at hand, in the sense that the best method in one application may be sub-optimal for the next application depending on the underlying data generation mechanism. For instance, parametric methods such as those based on linear regression work well even with limited data if the assumed model(s) are in congruence with the underlying true mechanism of dependency (Angrist and Pischke, 2008). However, under model misspecification, the performance of these methods significantly deteriorates. Matching-based approaches do not have to explicitly assume an outcome model, however, it has been shown that the result can be highly sensitive to the choice of distance metrics, whether the features being matched are scaled appropriately or other hyper-parameters (Stuart, 2010; Diamond and Sekhon, 2013; Parikh et al., 2018; Wang et al., 2021). Non-parametric methods such as double machine learning with gradient boosting trees and targeted maximum likelihood estimation (TMLE) with superlearner can require relatively large datasets and fine-

Table 1. Lalonde's NSW Sample ATE Estimated using few commonly used causal effect estimation methods.

Estimators	ATE Estimate	Std. Dev.
Difference of Means	886.30	277.37
Double Machine Learning	370.94	394.68
Causal BART	818.79	184.46
Propensity Score Matching	1079.13	158.59

tuning of hyper parameters to ensure optimal performance (Athey and Imbens, 2019; Cui and Tchetgen, 2019; Van der Laan et al., 2007; Zheng and van der Laan, 2011). Thus, there is clearly no '*one-size-fits-all*' estimation method for all scenarios. Validation of causal estimation methods for a given application is an important and challenging area of great consequence for causal inference practitioners, which to-date remains under-developed despite having received some attention.

Existing approaches to validating causal methods. Existing approaches in the literature to validate or evaluate the performance of causal methods on a specific dataset of interest can be divided into three main categories: (1) face-validity test, (2) placebo or negative control tests, and (3) synthetic data tests. Face-validity tests assess the estimated treatment effect against an expert's intuition. For instance, if the expected effect of a malaria medication according to a pharmacologist was a reduction in risk of death, but a causal method's estimate was contrary to the belief, then this fails the face-validity test. This test clearly is neither sufficient nor necessary - violation does not necessarily indicate that the causal estimation method is biased while agreement with expert's intuition does not imply that the causal estimation method is accurate. Placebo tests are the most widely used validation approach for causal models and they come in two varieties: "in-time" and "in-sample" placebo tests (Abadie et al., 2010; Ferman and Pinto, 2017; Athey and Imbens, 2017). An in-time-placebo test either restricts the analysis to a time-period whereby the causal effect is known to be null, or reassigns the treatment to a time period before the actual observed treatment and estimates the treatment effect to generate the distribution of treatment effects under null. In-sample-placebo tests, on the contrary, reassign the treatment to a control unit unaffected by the treatment, and use the causal inference method to estimate the effect. These tests are valid only under strong assumptions such as: (a) for in-time placebo tests, selection into treatment is assumed to be independent of pre-treatment outcomes which can lead to reverse causal-dependency bias, (b) in-sample placebo tests assumes that the placebo treatment assignment preserves key features of the true treatment assignment mechanism. Closely related to (a) and (b), negative control exposure or outcome variables are also routinely used for causal validation (Lipsitch et al., 2010; Shi et al., 2020) Such methods

require that negative control exposure and negative control outcome variables are available such that they satisfy key exclusion restrictions and U-comparability conditions (Lipsitch et al., 2010).

Lastly, for synthetic data tests, the design typically includes a known data-generating mechanism. This allows access to treatment effects ground truth used to evaluate the accuracy of causal methods on simulated data. However, synthetic data generating processes are commonly handcrafted, oversimplified, and thus unlikely to reflect the same level of complexity as real-world data. In addition, these overly-stylized data generating processes tend to a priori favor certain methods over others (Athey et al., 2021; Advani et al., 2019; Knaus et al., 2021; Gentzel et al., 2019; 2021). Hence, good performance of a given method on handcrafted synthetic data does not necessarily translate to equally good performance on real-world data. In this paper, we present an approach to generate synthetic data sets that satisfy two salient properties sought out in simulation studies: (i) user-specified causal treatment effects, heterogeneity, and endogeneity; (ii) simulated samples that are stochastically indistinguishable from the observed data sample of interest.

Contribution. Specifically, we introduce a general framework (*Credence*) to validate and evaluate the performance of various existing causal inference methods using synthetic data anchored at the empirical distribution of a given data set of interest (see Section 2.1). The approach relies on a deep generative model trained on a rich universe of data sets that share certain key features with the data set of primary interest, and is the basis to operationalize the proposed framework (see Section 2.2). *Credence* successfully fulfils requirements (i) and (ii), central to obtaining an objective evaluation of causal methods to a data set in view.

A recent work by Athey et al. (2021) similarly proposed to use of deep generative models to simulate synthetic data to validate causal methods. However, in contrast to our approach, Athey et al. (2021) assumed that the observed data satisfied conditional ignorability, a restriction which they also imposed in generated samples. Furthermore, their proposed framework does not appear to accommodate user-specified treatment effects and confounding bias. As result, their approach only fulfils requirement (ii) but not (i), a important gap we address in this work.

We demonstrate *Credence*'s utility with two synthetic data scenarios (read Section 3.1), and on two real data examples – the Lalonde and Project STAR studies (read Section 3.2). For synthetic data experiments, we compare the oracle performance ranking of various causal inference estimation methods in a Monte Carlo setting where the true data-generating process (DGP) is completely known and

compare it with the performance evaluation produced by *Credence* without a priori knowledge of the DGP. We find that *Credence* can successfully re-produce the oracle performance ranking of the various estimation methods based on the single observed sample. Furthermore, in each real-world data set for which both experimental and observational data are available, we compare the performance ranking for the same set of causal inference methods in the observational sample with that in the experimental sample. Finally, we discuss important implications of *Credence* and highlight both strengths and limitations of the proposed framework in Section 4.

2. Methodology

In this section, first we introduce the general framework (Section 2.1) to learn a data generating process from the observed data set of interest, with user-defined levels of treatment effects and confounding. Next, we introduce *Credence*, which operationalizes the proposed framework using an ensemble of variational autoencoders (Section 2.2).

2.1. Framework

Intuitively, our first objective is to learn a data generating mechanism, G_θ, such that the joint distribution of the variables in a generated sample $D' = (X', Y', Z') \sim G_\theta$ is nearly indistinguishable from the empirical distribution of the observed data $D = (X, Y, Z)$. Crucially, in addition to the above feature, we incorporate user-specified causal treatment effects, heterogeneity and both observed and hidden confounding bias as constraints that must be satisfied to the extent possible by the learned law G_θ and therefore can stand as a causal ground truth anchor for generated samples.

In this vein, let $\mathbf{G}_\Theta$ denote a model consisting of possible data generating laws for D. Let $G^\dagger \in \mathbf{G}_\Theta$ denote the true generating model from which the observed data $D = (X, Y, Z)$ was sampled. For $\{\epsilon_{X,i}, \epsilon_{Z,i}, \epsilon_{Y,i}\}_{i=1}^N \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{1})$ and $\{U_i\} \overset{iid}{\sim} \mathcal{N}(0, 1)$, throughout, we assume D follows the nonparametric structural equations model: (1) $X_i \sim \phi_X(\epsilon_{X,i})$, (2) $Z_i \sim \phi_Z(X_i, U_i, \epsilon_{Z,i})$, (3) $Y_i(1) \sim \phi_{Y(1)}(X_i, U_i, \epsilon_{Y,i})$, (4) $Y_i(0) \sim \phi_{Y(0)}(X_i, U_i, \epsilon_{Y,i})$, and (5) $Y_i = Z_i Y_i(1) + (1 - Z_i)Y_i(0)$ where, $G^\dagger = [\phi_X, \phi_Z, \phi_{Y(1)}, \phi_{Y(0)}]$. Further, let $\mathbf{G}_\Theta^\dagger \subseteq \mathbf{G}_\Theta$ be the subset of equally likely data generators for observed data D, and $\mathbf{G}_\Theta^* \subseteq \mathbf{G}_\Theta$ be the subset of feasible data generators that satisfy user-defined treatment effects and confounding bias functions. Informally, if $\mathbf{G}_\Theta^\dagger \cap \mathbf{G}_\Theta^* \neq \{\emptyset\}$ then we wish to find an arbitrary law $G_\theta^* \in \mathbf{G}_\Theta^\dagger \cap \mathbf{G}_\Theta^*$, otherwise, we wish to find a $G_\theta^* \in \mathbf{G}_\Theta^*$ which is "closest" to a $G_\theta \in \mathbf{G}_\Theta^\dagger$ (see Figure 1 for a graphical illustration). To formalize this idea, consider an optimization setup that takes in as input the observed data D; together with user-defined treatment effect
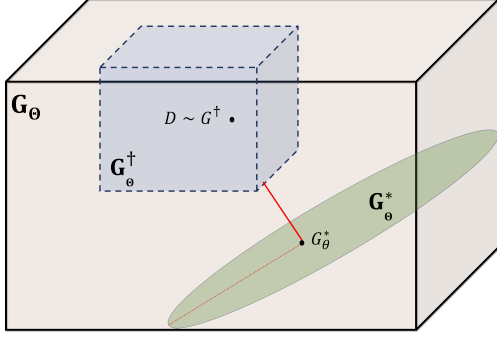
Figure 1. Schematic drawing of the optimization setup. Here $\mathbf{G}_\Theta$ is the space of all possible data generators, $G^\dagger$ is the true data generative process for the observed data D, $\mathbf{G}_\Theta^\dagger$ is the subspace of all equally likely data generators for observed data D, and $\mathbf{G}_\Theta^*$ is the subspace of potential data generators that can be learned using the assumed configuration. We want to find a $G_\theta^* \in \mathbf{G}_\Theta^*$ that is closest to a $G_\theta \in \mathbf{G}_\Theta^\dagger$.

function (f) and unmeasured confounding-bias function (g), both of which are defined as followed on the mean additive scale:

$$
\begin{aligned}
f(x) &= \mathbf{E}[Y(1) - Y(0)|X = x] \\
g(x,z) &= \mathbf{E}[Y(z)|X = x, Z = z] \\
&\quad - \mathbf{E}[Y(z)|X = x, Z = 1 - z]
\end{aligned}
$$

The setup also accommodates specification of the degree to which the user wishes the data generating mechanism be faithful to specified functions f and g, regulated by the pair of penalties, referred throughout as *rigidness parameters* α for f and β for g, respectively. We design the optimization objective to finding the parameter θ of the generative model G_θ that minimizes a certain distributional distance metric between the empirical distribution of D and the simulated data under specified treatment effect and confounding bias constraints.

Specifically, one may consider the Wasserstein distance as choice of distributional distance metric; in which case, the treatment effect constraint shrinks the conditional average treatment effect towards $f(X)$ and the corresponding confounding bias constraint shrinks the amount and nature of confouding bias towards $g(X, Z)$ in the fitted DGP. In case there is a trade-off between minimizing each of the three objectives, the rigidness coefficient is used to encode specific user's preferences for one over the other. More broadly, let d denote the distance metric of choice between empirical and estimated laws; then we wish to learn a data generating
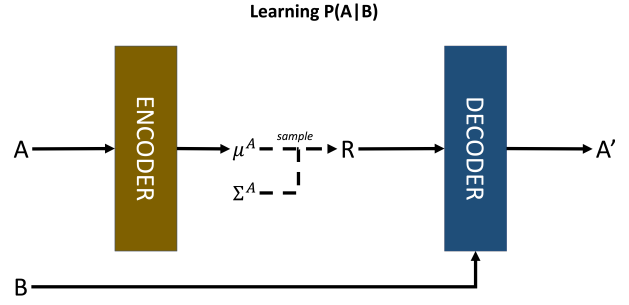

Figure 2. Schematic diagram showing conditional VAE (conVAE) used to learn a conditional distribution $P(A|B)$ for any two random variables A and B.

mechanism G_θ that satisfies:

$$
\begin{aligned}
\min_\theta \quad & \mathbf{E}\left[d((X, Y, Z), (X', Y', Z'))\right] \\
& + \alpha \left\| \mathbf{E}[Y'(1) - Y'(0)|X' = x'] - f(x') \right\| \quad (1) \\
& + \beta \left\|
\begin{array}{c}
\mathbf{E}[Y'(z')|X' = x', Z' = z'] \\
- \mathbf{E}[Y'(z')|X' = x', Z' = 1 - z'] \\
- g(x', z')
\end{array}
\right\|
\end{aligned}
$$

where $D' = (X', Y', Z') \sim G_\theta$.

Next, we discuss our procedure to operationalize this optimization using deep generative models.

2.2. Credence

Credence uses this framework to generate data for comparing and validating causal inference methods. Credence learns the joint distribution $P(X, Z, Y)$ as an ensemble of three conditional distributions: $P(Y|X, Z)$, $P(X|Z)$ and $P(Z)$ using a variational autoencoder neural networks. A brief background on variational autoencoders (VAE) is discussed in Appendix A. Our work uses the VAE framework to learn the above mentioned conditional distributions by allowing the decoder to not only be a function of sampled latent vectors but also of variables we are conditioning on (see Figure 2). In machine learning literature, VAEs have been shown to work extremely well for modeling numerical as well as vision and text data (Kingma and Welling, 2019; Kusner et al., 2017). However, to the best of our knowledge, deep generative models such as VAEs haven't been used in the specific way described for validating and evaluating causal inference methods.

In this work, we have assumed Z to be a binary indicator. Thus, learning a generative model for $P(Z)$ boils down to estimating $p_z = P(Z = 1)$, the parameter of a Bernoulli distribution. Thus, the sample proportion $\widehat{p}_z = \frac{\sum_i Z_i}{N}$ is an unbiased and consistent estimate of p_z. We use the conditional VAEs to then learn probability distributions

$P(X|Z)$ and $P(Y|X, Z)$ separately.

Consider the VAE setup $H^{X|Z}_{(\phi_{X|Z}, \theta_{X|Z})} = (F^{X|Z}_{\phi_{X|Z}}, G^{X|Z}_{\theta_{X|Z}})$ to learn $P(X|Z)$. Here, $F^{X|Z}_{\phi_{X|Z}}$ is the encoder that maps the observed data to a lower dimensional latent space, and $G^{X|Z}_{\theta_{X|Z}}$ is the decoder that maps back points to the original dimensions. Thus, the VAE encoder maps X_i to latent space as $[\mu_i^X, \Sigma_i^X] = F^{X|Z}_{\phi_{X|Z}}(X_i)$, then sample a point R_i^X in latent space in the vicinity of μ_i^X by drawing it from the distribution $R_i^X \sim \mathcal{N}(\mu_i^X, \Sigma_i^X)$ and R_i^X is mapped back to the space of X using the decoder, $X_i' = G^{X|Z}_{\theta_{X|Z}}(R_i^X, Z_i)$. We use the ELBO loss function for the setup which is extremely common in the VAE literature. It is given as follows,

$$\mathcal{L}_{X|Z}(\phi_{X|Z}, \theta_{X|Z}; \{(X_i, Z_i)\}_{i=1}^N)$$
$$= \sum_{i=1}^N \|X_i - X_i'\|_2 + D_{\mathrm{KL}}(\mathcal{N}(\mu_i^X, \Sigma_i^X)\|\mathcal{N}(0, 1)).$$

Similarly, consider the VAE setup $H^{Y|X,Z}_{(\psi_{Y|X,Z}, \theta_{Y|X,Z})} = (F^{Y|X,Z}_{\psi_{Y|X,Z}}, G^{Y|X,Z}_{\theta_{Y|X,Z}})$ for learning $P((Y(1), Y(0))|X, Z)$. As noted before, for each unit in the original dataset we only observe one of $Y(0)$ or $Y(1)$. Thus, we update the loss function to inform the learning of $P((Y(1), Y(0))|X, Z)$ using the user-provided treatment effect function f and confounding bias function g. That is, the constraints imposed by choices of f and g together with their corresponding penalties implicitly induce correlation between $Y(1)$ and $Y(0)$. Let $[\mu_i^Y, \Sigma_i^Y] = F^{Y|X,Z}_{\psi_{Y|X,Z}}(Y_i)$, $R_i^Y \sim \mathcal{N}(\mu_i^Y, \Sigma_i^Y)$, $(Y_i'(1), Y_i'(0)) = G^{Y|X,Z}_{\theta_{Y|X,Z}}(R_i^Y, X_i, T_i)$ and $(Y_i''(1), Y_i''(0)) = G^{Y|X,Z}_{\theta_{Y|X,Z}}(R_i^Y, X_i, 1 - T_i)$. Then, the loss functions training $H^{Y|X,Z}_{(\psi_{Y|X,Z}, \theta_{Y|X,Z})}$ is given by

$$\mathcal{L}_{Y|X,Z}(\psi_{Y|X,Z}, \theta_{Y|X,Z}; \{(Y_i, X_i, Z_i)\}_{i=1}^N)$$
$$= \frac{1}{N} \sum_{i=1}^N \|Y_i - (Z_i Y_i'(1) + (1 - Z_i)Y_i'(0))\|_2$$
$$+ D_{\mathrm{KL}}(\mathcal{N}(\mu_i^Y, \Sigma_i^Y)\|\mathcal{N}(0, 1))$$
$$+ \alpha \|Y_i'(1) - Y_i'(0) - f(X_i)\|$$
$$+ \beta \|Y_i'(1 - Z_i) - Y_i''(1 - Z_i) - g(X_i, Z_i)\|$$

where α and β are the user-defined hyper-parameters. The expression $\mathcal{L}_{Y|X,Z}(\phi_{Y|X,Z}, \theta_{Y|X,Z}; \{(Y_i, X_i, Z_i)\}_{i=1}^N)$ can be intuitively decoded – the first and second term ensure distributional similarity of observed and generated data, while the second term also forces the distribution in the latent space to be standard normal. The third term shrinks the individualized treatment effects to the one defined by f, and the fourth term shrinks the magnitude and nature of confounding bias towards g. Note that the conditions enforced by the third and fourth terms of the loss

function are tighter than the one described in equation 1, thus narrowing the space of feasible generative models. This choice is to primarily ensures efficient computation and training. Also, note that if $\alpha = 0$ then the treatment effects are free to assume any value, while if $\beta = 0$ then the unmeasured confounding can be arbitrary.

This approach approximates the population-level optimization setup described in Section 2.1 and Equation 1. Further, this setup allows each of the two VAEs to be trained independently. Hence, the procedure can be performed in parallel which reduces the overall time of analysis.

Credence-based Validation. Let G_θ^* be an optimal data generator that minimizes the objective described in Equation 1. Then, we use the ensemble of decoders from the trained VAEs $\widehat{G}_\theta = \{\widehat{G}_{\theta_{Y|X,Z}}, \widehat{G}_{\theta_{X|Z}}\}$ as an estimate of G_θ^*. Further, $\widehat{G}_\theta$ is used to generate data $(X', Z', Y'(1), Y'(0))$ with known treatment effects such that $\forall j$, $R_j^{X'} \sim \mathcal{N}(0, 1)$, $R_j^{Y'} \sim \mathcal{N}(0, 1)$, $Z_j' \sim Bernoulli\left(\frac{\sum_i Z_i}{|D|}\right)$, $X_j' \sim \widehat{G}_{\theta_{X|Z}}(R_j^{X'}, Z_j')$ and $(Y_j'(1), Y_j'(0)) \sim \widehat{G}_{\theta_{Y|X,Z}}(R_j^{Y'}, X_j', Z_j')$. The generated data is then used to assess the performance of various causal methods by comparing the estimated ATE to the true ATE, $E\{Y_j'(1) - Y_j'(0)\}$. In the remainder of the paper, we demonstrate Credence in two settings: (i) using synthetic data where ground truth is known; (ii) two real data applications each of which having available both observational and experimental samples. In settings (i) and (ii), Credence is trained on a universe of data sets generated via the nonparametric bootstrap from a single observed data sample. Credence is then used to conduct a Monte Carlo study anchored at a single realization constituting the observed sample, in order to compare the performance of the following set of canonical causal inference methods taken from published literature as representative of the landscape of available methods: (1) meta-learners (Künzel et al., 2019), (2) nonparametric double machine learning (DML) (Chernozhukov et al., 2018) (3) propensity score matching estimator (Rosenbaum, 2002), (4) causal forest (Wager and Athey, 2018), (5) targeted maximum-likelihood estimator (TMLE) (Van Der Laan and Rubin, 2006), (6) causal BART (Chipman et al., 2010), and (7) doubly robust estimator (Bang and Robins, 2005).

3. Experiments

In this section, we study *Credence*-generated recommendations and rankings of causal inference methods for various datasets under different conditions. We divide this section into two parts. The first part focuses on synthetically generated data with known true data generating process (DGP) . Here, we compare the performance of various causal infer-

ence methods on data sampled from the known true DGP. We compare these performances with the methods' performance on data sampled from three trained Credence models under different constraints. Ideally, we aim to evaluate the extent to which Credence can replicate the oracle's ranking under various sets of assumptions about treatment effects and confounding. The second part focuses on two real data case studies–(1) Lalonde data sets (Dehejia and Wahba, 1999), and (2) Project STAR dataset (Achilles et al., 2008). Both data sets have an experimental component (in which treatment is randomized) and an observational component (in which units self-select into treatment). We rank the performance of causal methods on the observational part of the data, using the difference of means estimate of ATE on experimental part as an unbiased estimator of ground truth. Further, we compare this ranking with the ranking produced using *Credence* trained on the observational component of each study.

3.1. Synthetic Data Experiment

We use two DGPs for our analysis. The first DGP generates normally distributed covariates with a quadratic treatment effect and selection into treatment as a function of pre-treatment covariates. The second one is the so-called Friedman's DGP that which admits highly non-linear outcome and treatment selection functions of uniformly distributed covariates. Both the DGP's are described in details in Appendix B. We analyze and rank the performance of different causal inference methods using samples (of size N=2500 and 10-dimensional pretreatment covariate space) drawn from (1) the true DGPs and (2) the DGPs learned by training *Credence* on a single sample data set from true DGP with only observed covariates. The goal is to understand if the ranking produced by both approaches are comparable.

In this study, we train three Credence models for each of the two datasets with different configurations of the treatment effect function ($f(X)$) and selection bias function ($g(X,T)$). (1) For the first one, we shrink $g(X,T)$ towards zero and $f(X)$ towards $(\mathbf{1}^T X_i)^2$ for quadratic DGP and $X_{i,3} \cos(\pi X_{i,1} X_{i,2})$ for Friedman's DGP (Figure 3(b)). (2) For the second one, we constraint both $f(X)$ and $g(X,T)$ to be equal to zero for all X and T. (3) Lastly, for the third one, we shrink both $f(X)$ towards zero but constraint $g(X,T) = 0.15(2T - 1)$ to understand the sensitivity of different methods to unobserved confounding. For each of the cases, the potential outcomes will choose the values that will minimize Credence's training loss while respecting the above mentioned constraints.

For *Quadratic DGP*, we observe that while most of the methods have comparable accuracy, gradient boosting trees (GBT) S learner's estimates are the most biased on average (Figure 3(a)). Further, linear S learner and Causal BART

and linear DML tend to perform the best. An analogous conclusion can be drawn from the performance assessed using Credence's framework (see Figure 3(b) and (c)) where GBT S learner has the largest bias on average. In Figure 3(d), we study the case with no treatment effect but high selection bias that selects units with higher $Y(1)$ into treated group and units with higher $Y(0)$ into control group. There, as expected, we observe that almost all methods are biased. However, the true treatment effect is in 95 percent confidence interval for GBT S learner and GBT DML. Further, comparing the performance of causal methods on highly non-linear *Friedman's DGP*, we observe that propensity score matching is highly biased while Causal BART, GBT DML and GBT S learners tend to be the least biased, as indicated using the true DGP (Figure 3(a)). Similar performance is observed when evaluated using Credence-learned-DGPs (see Figure 3(b), (c) and (d)).

The main takeaway from this analysis is that Credence is able to reproduce rankings obtained by an oracle with access to the true DGP in cases where the constraints broadly align with the structure of true DGP. This highlights that the performances evaluated using Credence can provide reliable inferences in such a setting.

3.2. Real Data Case Studies

The case studies described in this section aim at (1) showcasing the applicability of *Credence* on real world data, and (2) highlight the strength of *Credence* in evaluating causal methods. To address the second objective, we selected data sets that include both experimental and observational arms. We use the difference of means between treatment and control arms estimated in experimental sample as ground truth ATE and evaluate estimation of ATE in the observational arm against the former. Next, we train *Credence* models only using the observational sample for each of the data sets and evaluate the methods on samples drawn from the trained models. We compare these rankings to determine the extent to which *Credence* performance evaluation of various causal inference methods can be reliable.

3.2.1. DATA DESCRIPTION

Lalonde Temporary Employment Program. The National Support Work Demonstration (NSW) temporary employment program is a randomized controlled trial aimed at studying the effect of a temporary employment program in the US on post intervention income level of the participants (LaLonde, 1986). Because this data set is from an RCT, treatment assignment is random and there is no unobserved confounding. An observational counterpart of this data set is the Population Survey of Income Dynamics (PSID) control sample (described in (Dehejia and Wahba, 99)). Treated units from NSW combined with control units from PSID

are frequently used as benchmarks to compare the performance of causal inference methods on real data (REFS). The outcome variable for both NSW and PSID samples is an individual's income in 1978. The observed pre-treatment covariates include the age, education, race, degree and 1975 income of participants. The unbiased point estimate of average treatment effect for NSW sample using difference of means estimator is $886.30. Hence, comparing the estimated treatment effect on the combined data set with the ATE from NSW sample benchmarks the performance of causal inference methods.

Project STAR. Project STAR (Student-Teacher Achievement Ratio) is an experiment primarily designed to study the effect of class-size in the earlier grades on short and long-term student performance (Achilles et al., 2008; Mosteller, 2014). A total of 79 schools in the US were chosen across the state of Tennessee in inner-city (17), urban (8), suburban (16) and rural areas (38). The experiment has a single cohort that was studied for four years. This included the students entering kindergarten in 1985 and students that initiated their public schooling in first grade in 1986. For each school of the 79 STAR schools, the students and teachers were randomly assigned to one of three treatment arms: a small class (13 to 17 students), a regular-sized class (22 to 25 students) and a regular-sized class with a full-time teacher aide. For our analysis, we consider the students in small class room as treated while all the other students as control. The outcome was measured using a standardized achievement tests in the spring of each year from 1986 to 1989. Apart from the STAR school data, the information about the non-randomized comparison group was collected concurrently. The comparison group includes information on 1780 students across grades 1 to 3 from 21 schools. The comparison schools were also selected from the same 13 districts as STAR schools. The unbiased point estimate of average treatment effect on Project STAR's experimental sample is 7.24.

3.2.2. ANALYSIS

Analyzing performance with respect to the experimental ATE. First, we estimated the causal effects of the interventions for Lalonde data and Project STAR's observational data using the various candidate causal inference methods. We compared these estimates with a "difference of means" estimate of causal effect in experimental data sets for both studies. As treatment is randomly assigned to units in each of the experimental data sets, the difference of means is an unbiased and consistent estimator. Thus, we treat it as a gold-standard. We compare and rank the performance of ATE estimators (see Figures 4(a)). For both data sets we find that the bias of each of the estimators on average are very similar. Furthermore, the variance of estimated bias was obtained from 50 bootstrap samples. We observed that

the non-parametric DML had significantly larger standard errors on Lalonde data compared to all the other methods, conditional on the observed sample. For Project STAR data, the propensity score method had slightly higher bias relative to other methods.

Analyzing performance on Credence generated data. Next, we trained *Credence* only using the observational component of each study setting $f(X) = g(X, T) = 0, \forall X, T$. We sampled 50 data sets using trained Credence models for Lalonde and Project STAR observational data, respectively. For each sampled data set, we evaluated bias and variance of each candidate ATE estimator. For the Lalonde study, we observed that S-learners, causal forest, linear DML and doubly robust estimators gave the lowest average bias (Figure 4(b)). For the STAR study, most of the estimators (except GBT T learner, GBT X learner, causal forest and propensity score matching) had negligible bias according to evaluation using Credence generated data (Figure 4(b)).

For Lalonde's data, rankings based on comparing observational ATE with experimental ATE are largely similar to rankings produced using Credence except with respect to estimated variance of estimators . However, Figure 4(b) shows that, for Project STAR data, the estimated treatment effect based on observational data is significantly different from experimental data which possibly indicates that the experimental sample lacks external validity (von Hippel and Wagner, 2018; Justman, 2018). Acknowledging this caveat, most methods perform similarly as shown in Figure 4(a)&(b). The congruence between both rankings further highlights the validity of *Credence* as an evaluation method anchored at a single observational sample.

4. Discussion and Conclusion

Implications. In this paper, we demonstrate how one can use deep generative models (in this case variational autoencoders) to generate synthetic data sets that are virtually stochastically indistinguishable from a real data set of interest. Importantly, we establish that synthetic data generated via Credence with user-specified treatment effects and hidden confounding bias functions can accurately recover an oracle's evaluation of various causal effect estimators with knowledge of true data generating mechanism. Researchers continuously develop new causal inference methods with the aim of improving validity of causal effect estimates; however, the task of determining which methods may perform well for data at hand remains complex and time consuming. Therefore, Credence has the potential to become a useful tool to assist the applied researcher in selecting an adequate method for their observed data.

Selecting f and g. For evaluating causal methods, the best one can hope is to understand the performance of these methods on the synthetic data generated by assuming one or
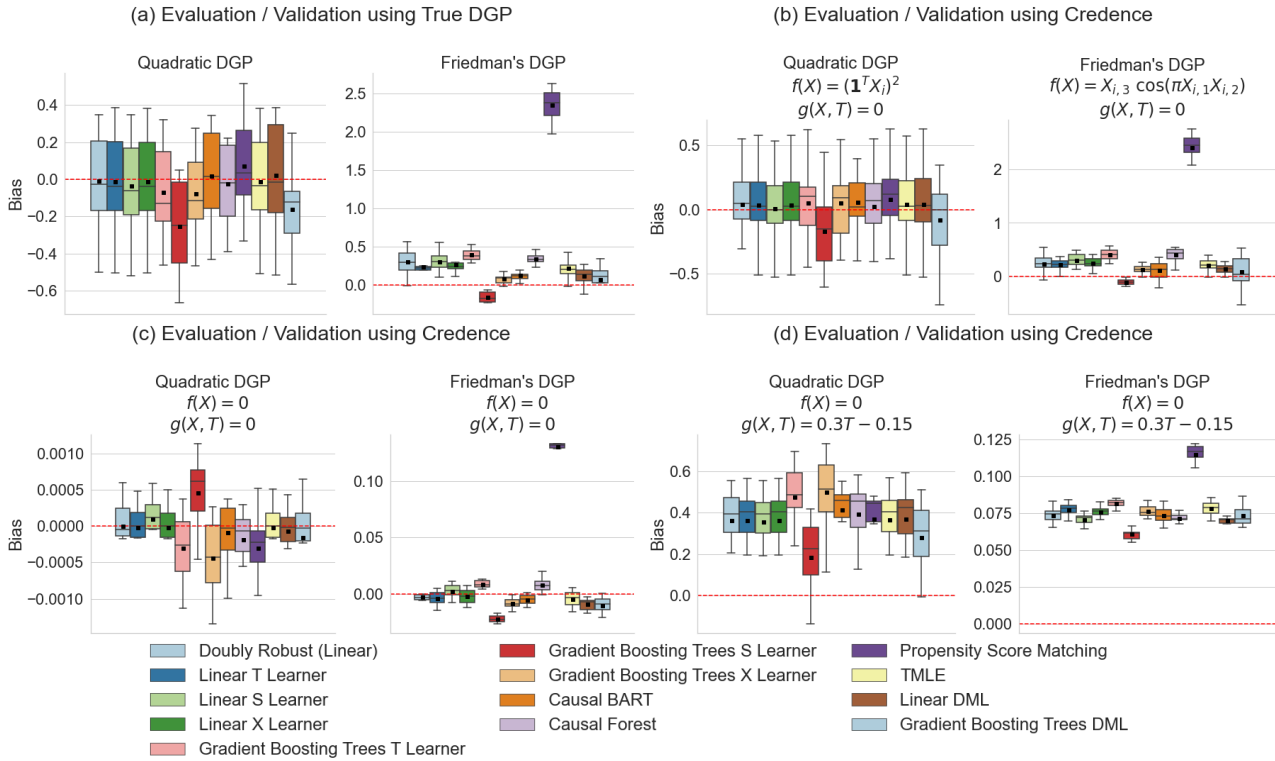
Figure 3. Comparing the performance of various causal inference methods on data generated using (a) *ground truth* quadratic and Friedman's DGPs; (b) Credence trained on a sample of quadratic DGP with $f(X) = (\beta^T X_i)^2 - \beta^T X_i + \alpha^T X_i$ and Credence trained on a sample of Friedman's DGP with $f(X) = X_{i,3} \cos(\pi X_{i,1} X_{i,2})$ and $g(X,T) = 0$ for both cases; (c) Credence trained on a sample of quadratic DGP and Credence trained on a sample of Friedman's DGP with $f(X) = 0$, $g(X,T) = 0$; (d) Credence trained on a sample of quadratic DGP and Credence trained on a sample of Friedman's DGP with $f(X) = 0$, $g(X,T) = 0.3T - 0.15$ for both cases.
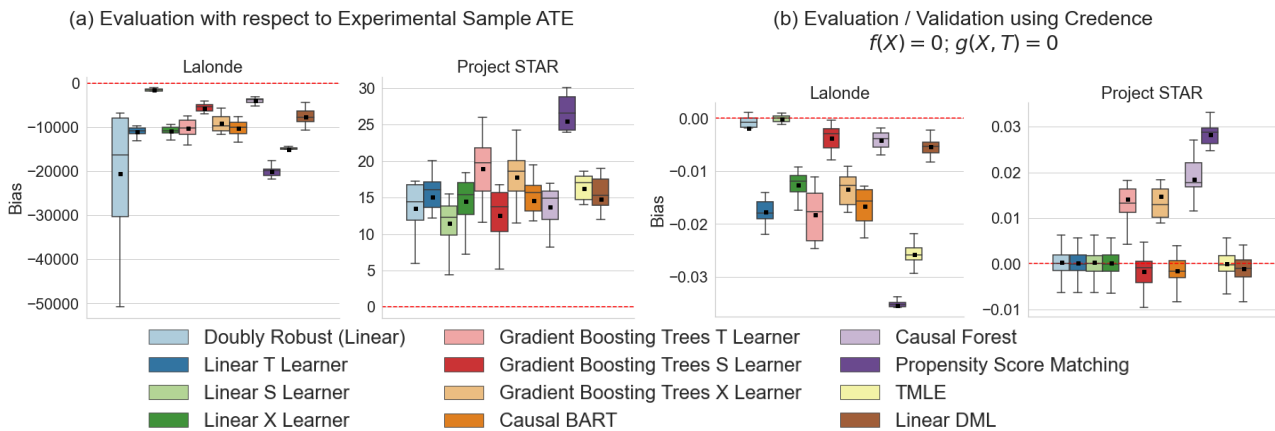

Figure 4. Comparing the performance of various causal inference methods in estimating observation ATE with respect to difference in means experimental ATE estimate for (a) Lalonde NSW+PSID data (b) Project STAR's comparison group data, and with respect to true ATE for data generated using (c) Credence trained on NSW+PSID data (d) Credence trained on Project STAR comparison group data.

both of the treatment effect function, $f(\cdot)$ and/or selection bias function, $g(\cdot)$. Here, the primary purpose of Credence is not to infer the underlying causal effect but rather assess these methods under certain assumptions about the DGP and rank these methods. As we show in the synthetic data experiments, the suggested approach is to evaluate the methods using *Credence* for a suite of different $f(\cdot)$'s and $g(\cdot)$'s. We recommend two primary strategies for selecting f and g. The first approach follows from using the observed data to estimate the larger feasible omitted variable bias possible if one of the observed covariates was omitted. This estimate can help a research understand the potential scale of selection bias (g) they can choose. Thus, by doing this one can fixing g to the largest observed omitted variable bias and $\alpha = 0$ – this will let the model learn flexible treatment effects based on the observed data. The second approach is to choose a class of f and g, e.g. polynomials of degree 3 or less. One perform a *min-max* search i.e. to find the most adversarial f and g from the class, and choose a causal inference method performs best in that setting.

Limitations and Future Direction. Researchers face a wide variety of causal measurement problems with different data structures (e.g., small/large number of observations, i.i.d. or time series, small/large number of features). Even when problems share a common structure, underlying data sets can be very unique: some may be sparse, some may be long-tailed, some may have very little feature overlap between treatment and control arms, etc. Although Credence leverages a deep generative model which can flexibly adapt to most of these settings, it still requires careful hyper-parameter tuning for each data set for optimal performance. Hence, a future important direction would involve further developing algorithms to automatically tune hyper-parameters in order to produce synthetic data of the best quality requiring minimal discretion from the applied researcher.

Potential future use for Credence is to perform inference in a manner analogous to the Bootstrap by directly leveraging the learned DGP to obtain repeated samples from Credence. For validity, this would require the researcher to make a non-testable assumption that Credence has successfully uncovered the true underlying DGP for the observed sample. This assumption is more restrictive than required to use Credence to benchmark different causal models. In this latter case, Credence only needs to learn a DGP that can generate data sufficiently similar to the single observed sample, and can reproduce to the extent possible, performance rankings of candidate causal inference estimators similar to those of an oracle with knowledge of the DGP.

Finally, Credence's evaluation is conditional on the assumptions encoded by the users in learning the DGP, and therefore its diagnostics are as good as these assumptions. For instance, Credence assumes no interference, and no mea-

surement error; if this were not true, the expected performance of candidate estimators of the ATE provided by Credence may be invalid. However, Credence is sufficiently flexible that such evaluation can in principle be conducted under identifying conditions that may be varied in a form of sensitivity analysis.

References

Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American statistical Association*, 105(490):493–505.

Achilles, C., Bain, H. P., Bellott, F., Boyd-Zaharias, J., Finn, J., Folger, J., Johnston, J., and Word, E. (2008). Tennessee's Student Teacher Achievement Ratio (STAR) project.

Advani, A., Kitagawa, T., and Słoczyński, T. (2019). Mostly harmless simulations? using monte carlo studies for estimator selection. *Journal of Applied Econometrics*, 34(6):893–910.

Angrist, J. D. and Pischke, J.-S. (2008). *Mostly harmless econometrics*. Princeton university press.

Athey, S. and Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2):3–32.

Athey, S. and Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, 11:685–725.

Athey, S., Imbens, G. W., Metzger, J., and Munro, E. (2021). Using wasserstein generative adversarial networks for the design of monte carlo simulations. *Journal of Econometrics*.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters.

Chipman, H. A., George, E. I., and Mcculloch, R. E. (2010). BART: Bayesian additive regression trees. *Annals of Applied Statistics*, pages 266–298.

Cui, Y. and Tchetgen, E. T. (2019). Selective machine learning of doubly robust functionals. *arXiv preprint arXiv:1911.02029*.

Dehejia, R. H. and Wahba, S. (1999–). Lalonde Data. [Accessed: 2018-11-10].

Dehejia, R. H. and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, 94(448):1053–1062.

Diamond, A. and Sekhon, J. S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*, 95(3):932–945.

Dorie, V., Chipman, H., McCulloch, R., Dadgar, A., Team, R. C., Draheim, G. U., Bosmans, M., Tournayre, C., Petch, M., de Lucena Valle, R., et al. (2019). Package 'dbarts'.

EconML, M. (2019). Econml: A python package for ml-based heterogeneous treatment effects estimation.

Ferman, B. and Pinto, C. (2017). Placebo tests for synthetic controls.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, pages 1–67.

Gentzel, A., Garant, D., and Jensen, D. (2019). The case for evaluating causal models using interventional measures and empirical data. *arXiv preprint arXiv:1910.05387*.

Gentzel, A. M., Pruthi, P., and Jensen, D. (2021). How and why to use experimental data to evaluate methods for observational causal inference. In *International Conference on Machine Learning*, pages 3660–3671. PMLR.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Hernán, M. A. and Robins, J. M. (2010). Causal inference.

Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, 42(8):1–28.

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

Justman, M. (2018). Randomized controlled trials informing public policy: Lessons from project star and class size reduction. *European Journal of Political Economy*, 54:167–174.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational bayes.

Kingma, D. P. and Welling, M. (2019). An introduction to variational autoencoders.

Knaus, M. C., Lechner, M., and Strittmatter, A. (2021). Machine learning estimation of heterogeneous causal effects: Empirical monte carlo evidence. *The Econometrics Journal*, 24(1):134–161.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165.

Kusner, M. J., Paige, B., and Hernández-Lobato, J. M. (2017). Grammar variational autoencoder.

LaLonde, R. J. (1986). Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *American Economic Review*, 76(4):604–620.

Lipsitch, M., Tchetgen, E. T., and Cohen, T. (2010). Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology (Cambridge, Mass.)*, 21(3):383.

Mosteller, F. (2014). *The Tennessee study of class size in the early school grades*. Princeton University Press.

Parikh, H., Rudin, C., and Volfovsky, A. (2018). Malts: Matching after learning to stretch. *arXiv preprint arXiv:1811.07415*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Rosenbaum, P. R. (2002). Observational studies. In *Observational Studies*, pages 1–17. Springer.

Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association*, 79(387):516–524.

Shi, X., Miao, W., and Tchetgen, E. T. (2020). A selective review of negative control methods in epidemiology. *Current epidemiology reports*, pages 1–13.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science*, 25(1):1.

Van der Laan, M. J., Laan, M., and Robins, J. M. (2003). *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media.

Van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super learner. *Statistical applications in genetics and molecular biology*, 6(1).

Van der Laan, M. J. and Rose, S. (2018). *Targeted learning in data science: causal inference for complex longitudinal studies*. Springer.

Van Der Laan, M. J. and Rubin, D. (2006). Targeted maximum likelihood learning. *The international journal of biostatistics*, 2(1).

von Hippel, P. and Wagner, C. (2018). Does a successful randomized experiment lead to successful policy? project challenge and what happened in tennessee after project star. *Project Challenge and What Happened in Tennessee After Project STAR (March 31, 2018)*.

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

Wang, T., Morucci, M., Awan, M. U., Liu, Y., Roy, S., Rudin, C., and Volfovsky, A. (2021). Flame: A fast large-scale almost matching exactly approach to causal inference. *J. Mach. Learn. Res.*, 22:31–1.

Zheng, W. and van der Laan, M. J. (2011). Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning*, pages 459–474. Springer.

Zivich, P. (2020). zepid.

A. Background on Variational Autoencoders (VAE)

Autoencoders refer to a particular machine learning estimator that aims to learn a lower dimensional representation of the data which provides a one-to-one mapping between the original data and the lower-dimensional representation (Goodfellow et al., 2016). Variational autoencoders (VAE) extend this idea to learn a representation of a high-dimensional complex data set as a standard normal distribution in a lower dimensional latent space. This allows sampling from standard normal distributions in the latent space and projecting it to the original high-dimensional space while ensuring that the distributional properties of the projection and the original data are virtually identical (Kingma and Welling, 2014). Typically, a VAE constitutes two parts – an encoder and a decoder. The encoder takes in the data in the original high dimensional space (S) and maps it to π, a lower dimensional latent space. A decoder performs the opposite operation to that of the encoder, mapping a vector R in lower dimensional latent space to vector S' in higher dimensional space of the original data. The learning algorithm includes (1) encoding the data S as a vector in low-dimensional latent space $\pi = [\mu, \Sigma]$, (2) sampling random vector R in the latent space from a multivariate normal distribution with mean μ and covariance matrix Σ derived from the encoded vector, and (3) decoding the sample vector R from the latent space by projecting it into the space of original data. The VAE loss function has two parts: (i) reconstruction loss (similarity of the input and the decoded output), and (ii) KL divergence between the normal distribution $\mathcal{N}(\mu, \Sigma)$ and standard normal distribution $\mathcal{N}(0, 1)$. The reconstruction loss enforces similarity between the empirical distributions of the original and generated samples, while KL divergence ensures that the distribution of latent vectors is as close to a standard normal distribution as possible. This ensures that sampling from standard normal distribution and decoding will have distributional congruence with the original data.

B. Data Generative Procedure for Synthetic Data Experiments

Quadratic DGP: The pre-treatment covariates X are sampled from a multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$, and the potential outcome function and treatment selection function are defined as follows:

$$
\begin{aligned}
Y_i(0) &= \beta^T X_i + \epsilon_{0,i} \quad \text{where } \epsilon_{0,i} \sim \mathcal{N}(0, 1) \\
Y_i(1) &= Y_i^2(0) + \alpha^T X_i + \epsilon_{1,i} \quad \text{where } \epsilon_{1,i} \sim \mathcal{N}(0, 1) \\
\pi_i &= \text{expit}(\gamma \times (\mathbf{1}^T X_i)) \\
T_i &\sim \text{Bernoulli}(\pi_i)
\end{aligned}
$$

Friedman's DGP: This DGP (Friedman, 1991) was first proposed to assess the performance of prediction methods. We augment Friedman's simulation setup to evaluate causal inference methods. The pre-treatment covariates are sampled from the standard uniform distribution. The potential outcome $Y_i(0)$ is defined by Friedman's function (Friedman, 1991; Chipman et al., 2010). The expected treatment effect we study is equal to the cosine of the product of the first two covariates scaled by the third covariate.

$$
\begin{aligned}
Y_i(0) &= 10 \, \sin(\pi X_{i,1} X_{i,2}) + 20 \, (X_{i,3} - 0.5)^2 \\
&\quad + 10 \, X_{i,4} + 5 \, X_{i,5} + \epsilon_{i,0} \\
Y_i(1) &= Y_i(0) + X_{i,3} \, \cos(\pi X_{i,1} X_{i,2}) \\
\pi_i &= \text{expit}(X_{i,0} + X_{i,1} - 0.5) \\
T_i &\sim \text{Bernoulli}(\pi_i)
\end{aligned}
$$

C. Credence's Goodness of Fit

In this section, we discuss if the Credence data generated using the learned DGP for Lalonde data and Project STAR data is comparable to the observed real data. We have argued in the paper that the comparing first and second moment is not a sufficient metric of similarity. However, in this section, we compare the correlation matrix of the generated data with the real data because it is easy to visualize and communicate. As we standardize the data removing the mean from each covariate and scaling it by the variance of the same, the vector of means for covariates in the observed and generated data is anchored at 0. In Figure 5(a) and (b), we show that the correlation matrices (a proxy measure for second moment of the distribution) of the Credence generated data under constraints $f(X) = 0$ and $g(X, T) = 0$ is visually extremely similar to the correlation matrices of the observables in the real data. This provides a convincing evidence that Credence's generated data has similar distributional properties.

(a)Correlation Matrix for real and Credence generated Lalonde data

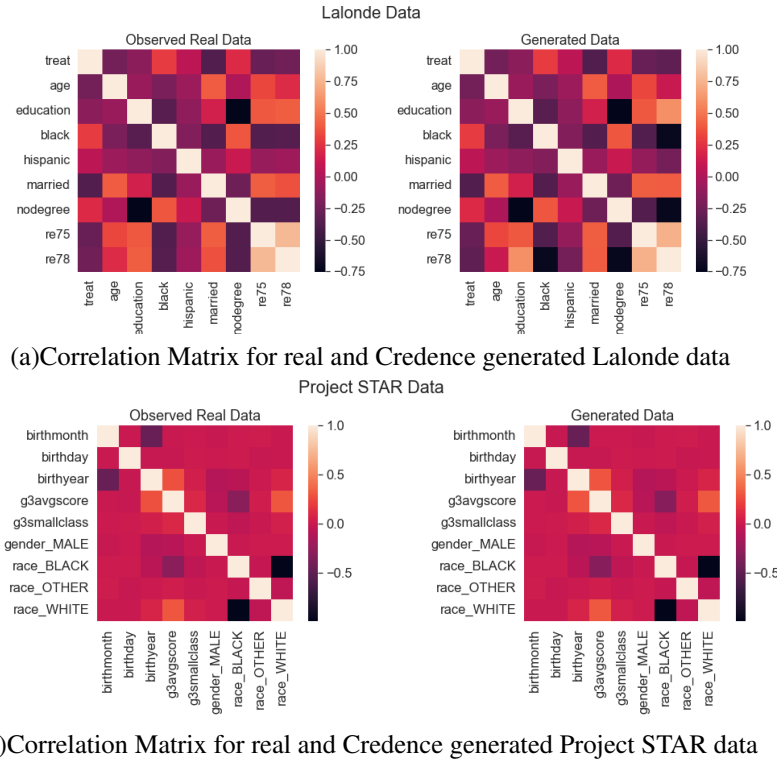(b)Correlation Matrix for real and Credence generated Project STAR data

Figure 5. Correlation matrices across covariates for comparing the second moments of distributions for the observed real data and Credence generated data for (a) Lalonde and (b) Project STAR. We compare them to understand the distributional similarities between both the DGPs

D. Implementation of Causal Methods

In this section, we discuss our implementation of causal inference methods studied in Section 3 such as double machine learning (DML), doubly robust estimation, propensity score matching, causal BART, causal forest, metalearners and TMLE. We used the existing libraries and packages

- We used the commonly used MatchIt's implementation of propensity score matching (Ho et al., 2011). We chose the method to match with replacement for estimating ATE.

- Our paper uses causal forest implementation from the 'grf' R package (Athey et al., 2019). We used the default setting designed by the developer with 2000 number of trees and $\sqrt{p} + 20$ variables tried per split.

- For causal BART, we use R implementation of BART by Vincent Dorie (Dorie et al., 2019). We only use the method with default hyperparameters.

- Our implementation of GBT DML and linear DML used EconML's implementation of these methods (EconML, 2019). We used the scikit-learn's machine learning API for the same (Pedregosa et al., 2011). For GBT DML, we used the method with 100 trees, and the linear DML used ridge regression.

- Further, we used EconML's implementation of metalearners such as S learner, T learner and X learner (EconML, 2019). Similar to DML, we used scikit-learn's ML API to for gradient boosting trees and ridge regression (Pedregosa et al., 2011).

- We used Paul Zivich's implementation of TMLE (Zivich, 2020).