Validating Causal Inference Methods

Harsh Parikh* Carlos Varjao⁺ Louise Xu⁺ Eric Tchetgen Tchetgen^

*Duke University
⁺Amazon.com
^University of Pennsylvania

Duke UNIVERSITY

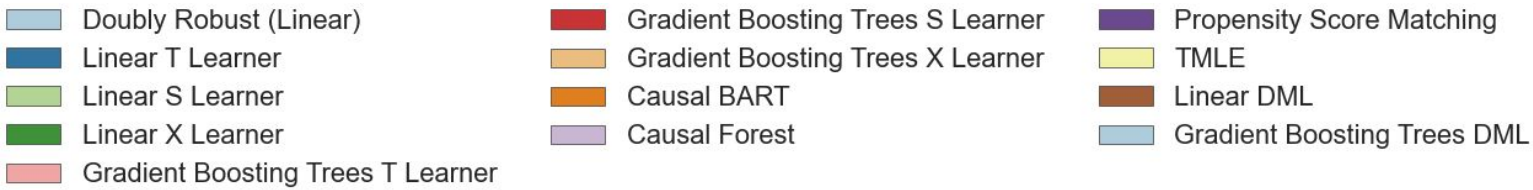
amazon

Wharton University of Pennsylvania

The Zoo of Causal Methods

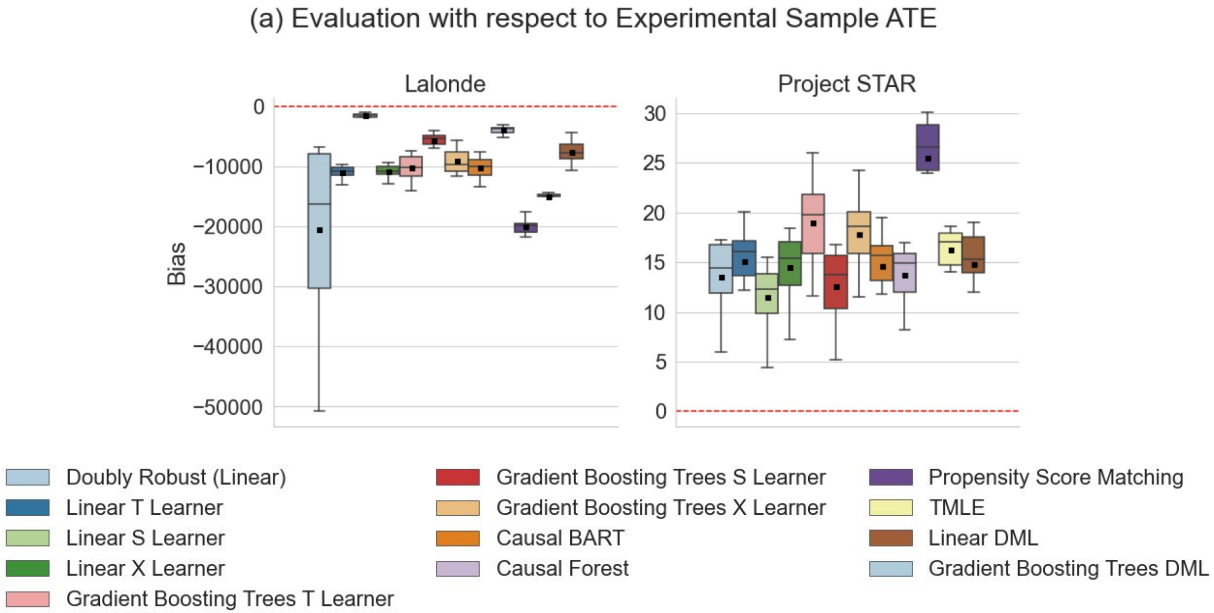
Many statistical methods have emerged for causal inference under unconfoundedness conditions given pre-treatment covariates, including:

- propensity score-based methods,
- prognostic score-based methods,
- doubly robust methods.



No ‘One-Size Fits All’ Method

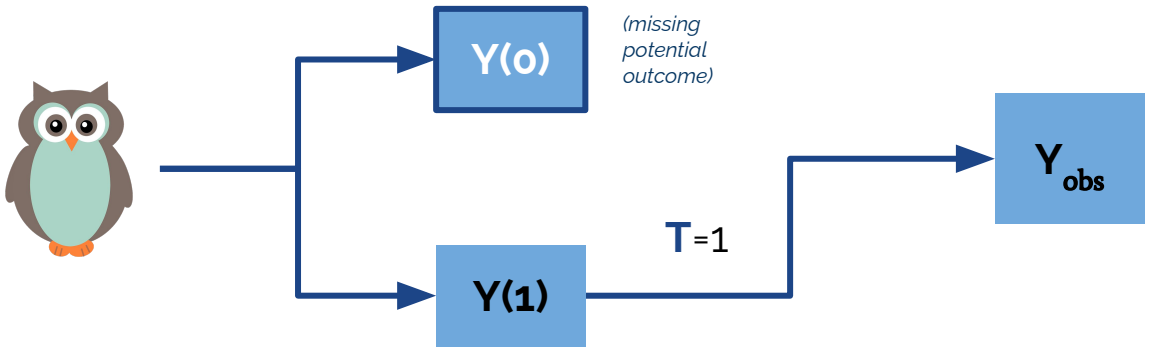
Unfortunately for applied researchers, there is *no ‘one-size-fits-all’* causal method that can perform optimally universally



The Difficulty on Estimating and Validating Causal Effects

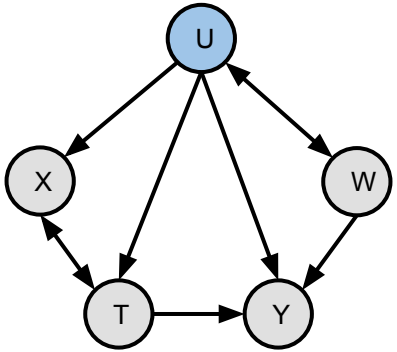
The fundamental challenge of drawing causal inference is that

- The counterfactual outcomes are not fully observed for any unit.
- Furthermore, in observational studies, treatment assignment is likely to be confounded.
- Thus, almost all causal inference methods depend on some untestable assumption(s).

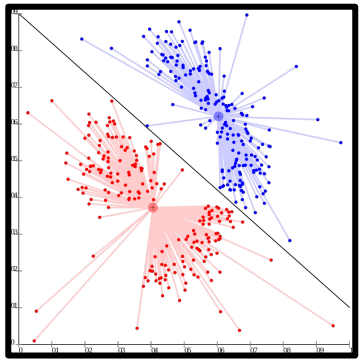


Existing Approaches to Evaluating Causal Methods

Face-Validity Test



Placebo/Negative Control Tests



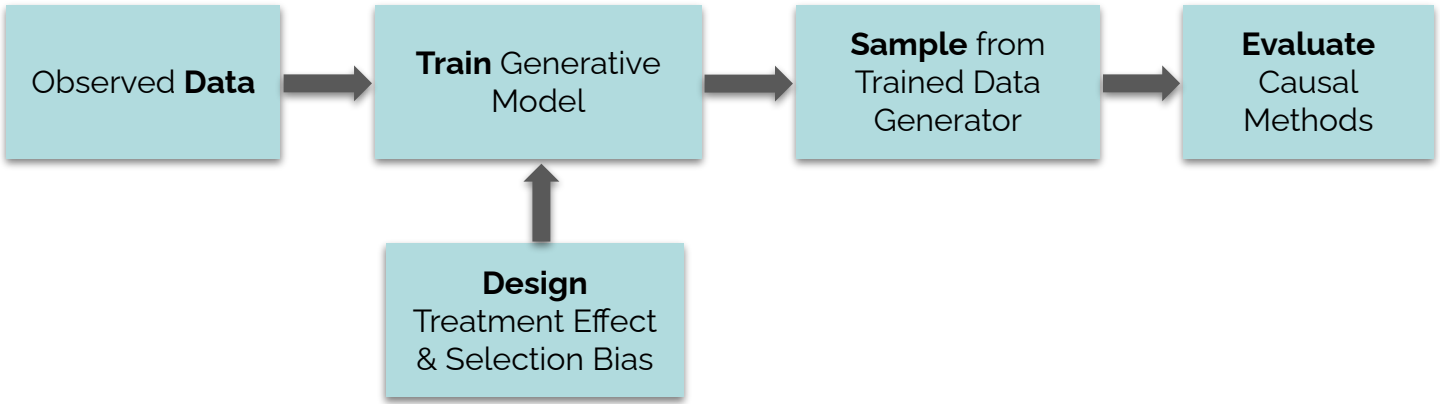
Handcrafted Synthetic Data Tests

Credence Framework

Our approach to generate synthetic data satisfies two salient properties sought out in simulation studies:

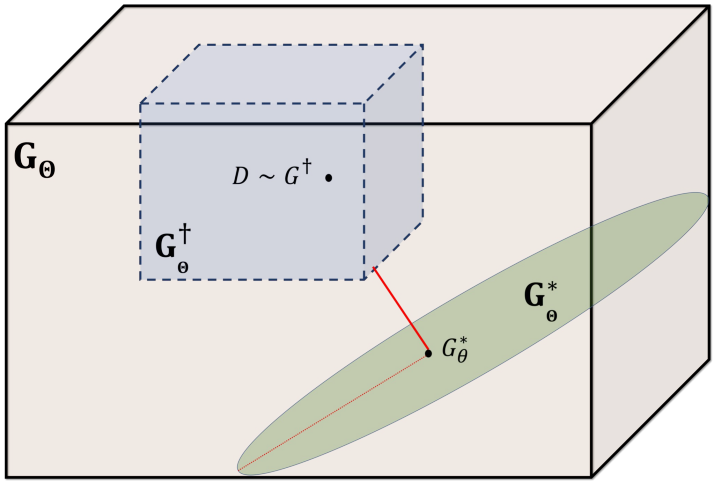
(i) user-specified causal treatment effects, heterogeneity, and endogeneity;

(ii) simulated samples that are stochastically indistinguishable from the observed data sample of interest.



Learning a Candidate Data Generator under Constraints

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



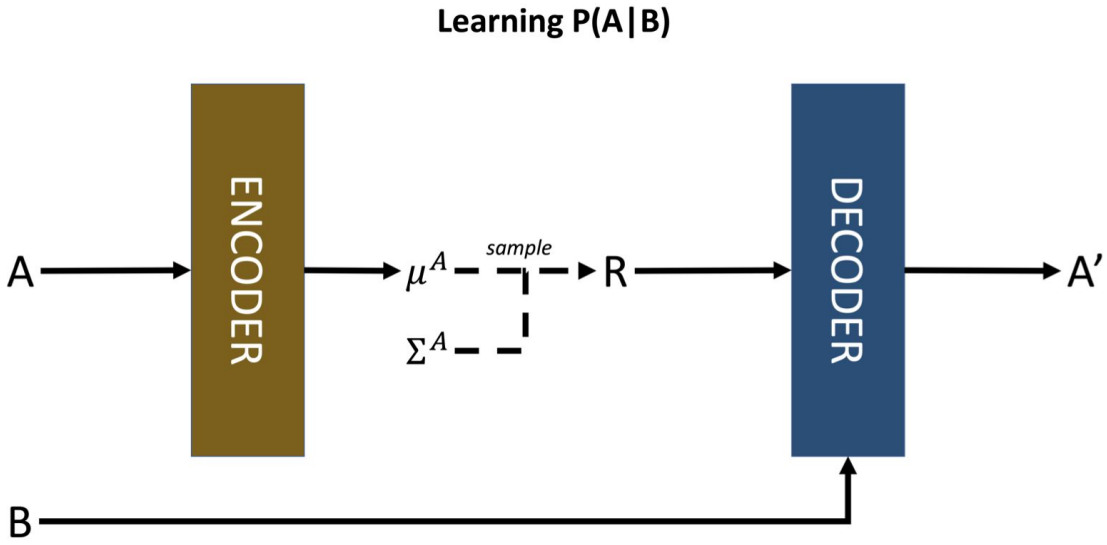
Validate and evaluate the performance using learned DGP anchored at

(i) the empirical distribution of a given data set of interest

(ii) user defined treatment effect/selection bias functions

Conditional Variational Autoencoders

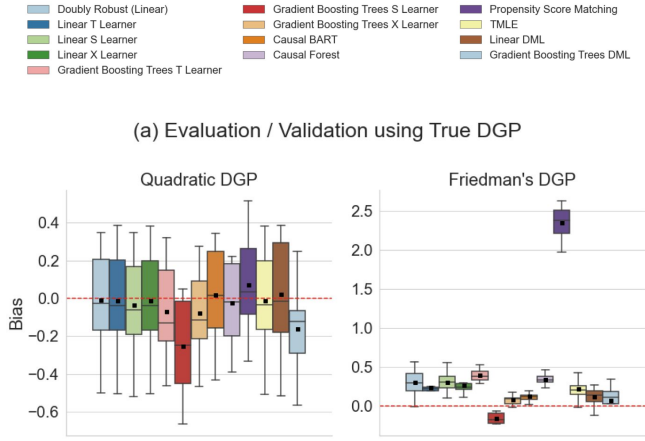
We leverage deep generative model trained on the data set of primary interest, which is the basis to operationalize the proposed framework.



True DGP* vs Credence learned DGP?

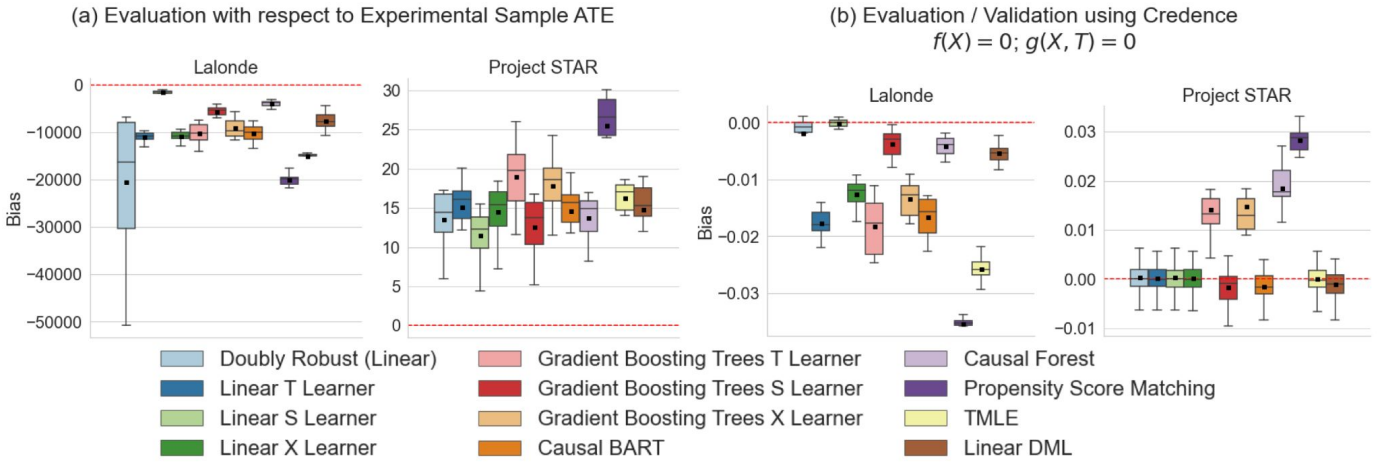
* only possible for synthetic data

- 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.



Experimental ATE* vs Credence learned DGP?

* only possible for where we have access to both experimental as well as observational data



- For *Lalonde's data*, rankings based on comparing observational ATE with experimental ATE are largely similar to rankings produced using Credence learned DGP except with respect to estimated variance of estimators.
- 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 except GBT T-learner, GBT X-learner, Causal Forest and PSM

Limitations

- Generative models are sensitive to hyper-parameters
- Evaluations as good as the assumptions user makes

Future Directions

- Use Credence as a deep-bootstrap for *inference*
- Extension to scenarios with interference/homophily
- Theoretical guarantees on Credence based ranking

Thank you so much!

Duke UNIVERSITY

amazon

Wharton University of Pennsylvania

Discussion Questions

- How do you choose f() and g()?
  - Min-Max strategy: method that performs best for the worst choice of f and g
  - Using observed data to estimate largest feasible OVB using observed data
- Why do doubly robust methods not perform optimally always?
  - Finite sample
  - Quadratic rate of bias
- VAE vs GAN?
  - VAE allows user to find the latent space location for every point in observed data
    - This allows user to sample from an interesting subspace if they are interested in doing that
  - GANs can be finicky and training them is more of an art sometimes
  - BTW, Credence can also be used with GANs or any other generative model of user's choice