Validating Causal Inference Methods

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

*Duke University  
+Amazon Science  
^The Wharton School, 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
Existing Approaches to Evaluating Causal Methods

- **Face-Validity Test**
- **Placebo/Negative Control Tests**
- **Handcrafted Synthetic Data Tests**
Objective: Evaluate Causal Methods using Synthetically Generated Data with (i) known Treatment Effects and (ii) is as complex as the Real Data of Interest
Credence
Notations

X, Y, Z : observed covariates, outcomes and treatment

Y(z) : potential outcome under treatment z
Notations

X, Y, Z : observed covariates, outcomes and treatment
Y(z) : potential outcome under treatment z

\[ Y = ZY(1) + (1 - Z)Y(0) \]
**Notations**

\(X, Y, Z\) : observed covariates, outcomes and treatment

\(Y(z)\) : potential outcome under treatment \(z\)

\[
Y = ZY(1) + (1 - Z)Y(0)
\]

\[
\mathbb{E}[Y|X, T = 1] - \mathbb{E}[Y|X, T = 0] = \mathbb{E}[Y(1) - Y(0)|X, T = 1]P(T = 1|X) \\
+ \mathbb{E}[Y(1) - Y(0)|X, T = 0]P(T = 0|X) \\
+ (\mathbb{E}[Y(0)|X, T = 1] - \mathbb{E}[Y(0)|X, T = 0]) P(T = 1|X) \\
+ (\mathbb{E}[Y(1)|X, T = 1] - \mathbb{E}[Y(1)|X, T = 0]) P(T = 0|X)
\]
**Notations**

X, Y, Z : observed covariates, outcomes and treatment

Y(z) : potential outcome under treatment z

\[ Y = ZY(1) + (1 - Z)Y(0) \]

\[ \mathbb{E}[Y|X, T = 1] - \mathbb{E}[Y|X, T = 0] = \mathbb{E}[Y(1) - Y(0)|X, T = 1]P(T = 1|X) + \mathbb{E}[Y(1) - Y(0)|X, T = 0]P(T = 0|X) + (\mathbb{E}[Y(0)|X, T = 1] - \mathbb{E}[Y(0)|X, T = 0])P(T = 1|X) + (\mathbb{E}[Y(1)|X, T = 1] - \mathbb{E}[Y(1)|X, T = 0])P(T = 0|X) \]

*Treatment effect*
Notations

$X, Y, Z$ : observed covariates, outcomes and treatment

$Y(z)$ : potential outcome under treatment $z$

$$Y = ZY(1) + (1 - Z)Y(0)$$

$$\mathbb{E}[Y|X, T = 1] - \mathbb{E}[Y|X, T = 0] = \mathbb{E}[Y(1) - Y(0)|X, T = 1]P(T = 1|X) + \mathbb{E}[Y(1) - Y(0)|X, T = 0]P(T = 0|X)$$

$$+ (\mathbb{E}[Y(0)|X, T = 1] - \mathbb{E}[Y(0)|X, T = 0]) P(T = 1|X)$$

$$+ (\mathbb{E}[Y(1)|X, T = 1] - \mathbb{E}[Y(1)|X, T = 0]) P(T = 0|X)$$
Notations

\[ \begin{align*}
  X, Y, Z & : \text{observed covariates, outcomes and treatment} \\
  Y(z) & : \text{potential outcome under treatment } z \\
  X' & : \text{simulated covariates} \\
  Y'(z) & : \text{simulated potential outcome} \\
  Z' & : \text{simulated treatment}
\end{align*} \]
Credence Framework

Our approach to generate synthetic data \((X', Y', Z')\) that satisfies two salient properties sought out in simulation studies:

- simulated samples that are **stochastically indistinguishable** from the **observed data sample**
- **User-specified causal treatment effects**, heterogeneity, and endogeneity.
Credence Framework

Our approach to generate synthetic data (X', Y', Z') that satisfies two salient properties sought out in simulation studies:

- simulated samples that are **stochastically indistinguishable** from the observed data sample
- **User-specified causal treatment effects**, heterogeneity, and endogeneity.
Learning a Candidate Data Generator under Constraints

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

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
Learning a Candidate Data Generator under Constraints

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

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
Learning a Candidate Data Generator under Constraints

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

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
Learning a Candidate Data Generator under Constraints

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

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
Learning a Candidate Data Generator under Constraints

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

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
Learning a Candidate Data Generator under Constraints
Variational Autoencoders

We leverage deep generative model such as Variational Autoencoders (VAE) trained on the data set of primary interest, which is the basis to operationalize the proposed framework.
How do Variational Autoencoders work?
(An Oversimplified Introduction)
How do Variational Autoencoders work? (An Oversimplified Introduction)
How do Variational Autoencoders work? (An Oversimplified Introduction)

Encoder

Sample from neighborhood of $\pi$

Latent space
How do Variational Autoencoders work? (An Oversimplified Introduction)
Conditional Variational Autoencoders

Learning $P(A \mid B)$

Sample from neighborhood of $\pi$

Encoder

Decoder

Latent space

$A$ $\rightarrow$ $\pi$ $\rightarrow$ $A'$

$B$
Credence with Conditional VAE

We learn 3 distributions:

- $P(Z)$
- $P(X | Z)$
- $P(Y(1), Y(0) | X, Z)$

$$\min_\theta \left( \frac{\mathbb{E} \left[ d((X, Y, Z), (X', Y', Z')) \right]}{\alpha \| \mathbb{E}[Y'(1) - Y'(0) | X' = x'] - f(x') \|} + \beta \| \mathbb{E}[Y'(z') | X' = x', Z' = z'] - \mathbb{E}[Y'(z') | X' = x', Z' = 1 - z'] - g(x', z') \| \right)$$
Credence with Conditional VAE

We learn 3 distributions:

- $P(Z)$
  - Binary $Z$: just learn the proportion of treated units
- $P(X | Z)$
- $P(Y(1), Y(0) | X, Z)$

$$\min_\theta \left( \frac{\mathbb{E} \left[ d((X, Y, Z), (X', Y', Z')) \right]}{\alpha \mathbb{E}[Y'(1) - Y'(0)|X' = x'] - f(x')} \right) + \beta \left( \mathbb{E}[Y'(z')|X' = x', Z' = z'] - \mathbb{E}[Y'(z')|X' = x', Z' = 1 - z'] - g(x', z') \right)$$
Credence with Conditional VAE

We learn 3 distributions:

- $P(Z)$
  - Binary $Z$: just learn the proportion of treated units; No VAE needed
- $P(X | Z)$
  - Fit a conditional VAE; minimize $d(X,X')$
- $P(Y(1), Y(0) | X, Z)$

$$\min_{\theta} \left( \frac{E \left[ d((X, Y, Z), (X', Y', Z')) \right]}{\alpha \| E[Y'(1) - Y'(0) | X' = x'] - f(x') \|} + \beta \| E[Y'(z') | X' = x', Z' = z'] - E[Y'(z') | X' = x', Z' = 1 - z'] - g(x', z') \|} \right)$$
Credence with Conditional VAE

We learn 3 distributions:

- **P( Z )**
  - Binary Z: just learn the proportion of treated units; No VAE needed

- **P( X | Z )**
  - Fit a conditional VAE; minimize \( d(X, X') \)

- **P( Y(1), Y(0) | X, Z )**
  - Fit a conditional VAE;
  - minimize \( d(Y, Y') \) + constraints for treatment effects and selection bias

\[
\min_\theta \left( \begin{array}{c}
\mathbb{E} [d((X, Y, Z), (X', Y', Z'))] \\
+ \alpha \| \mathbb{E}[Y'(1) - Y'(0)|X' = x'] - f(x') \|
\end{array} \right)
\]

\[
+ \beta \| \mathbb{E}[Y'(z')|X' = x', Z' = z'] - \mathbb{E}[Y'(z')|X' = x', Z' = 1 - z'] - g(x', z') \|
\]
Review: Credence Framework

Generate synthetic data (X', Y', Z') that satisfies two salient properties sought out in simulation studies:

- simulated samples that are **stochastically indistinguishable** from the observed data sample
- **User-specified** causal treatment effects, heterogeneity, and endogeneity.
- Use **conditional VAEs** to with constraints to learn the data generative process
True DGP* vs Credence learned DGP?

* only possible for synthetic data

(a) Evaluation / Validation using True DGP

Quadratic DGP

Friedman's DGP
True DGP* vs Credence learned DGP?

* only possible for synthetic data

(a) Evaluation / Validation using True DGP

- Quadratic DGP
- Friedman's DGP

(b) Evaluation / Validation using Credence

- Quadratic DGP
  \[ f(X) = (1^T X)^2 \]
  \[ g(X, T) = 0 \]

- Friedman's DGP
  \[ f(X) = X_{i,3} \cos(n X_{i,1} X_{i,2}) \]
  \[ g(X, T) = 0 \]
True DGP* vs Credence learned DGP?

* only possible for synthetic data

(a) Evaluation / Validation using True DGP

(c) Evaluation / Validation using Credence

- Quadratic DGP
- Friedman's DGP

- Propensity Score Matching
- TMLE
- Linear DML
- Gradient Boosting Trees DML
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
Experimental ATE* vs Credence learned DGP?

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