Does Machine Learning Work?: Designing Better Simulation Studies for Inference

Paul Zivich

Institute of Global Health and Infectious Diseases
University of North Carolina at Chapel Hill

June 12, 2023

Problem

How can we *know* our tools work as intended?

Causal Inference

Problem: estimate the average causal effect (ACE)

$$\psi_{1-0} = E[Y^1] - E[Y^0]$$

where Y^a is the potential outcome under action a

Data: action (A), outcome (Y), and covariates (W) for n units

Identification of Causal Effects

Identification

- Express ψ_{1-0} in terms of W, A, Y
- Causal consistency, conditional exchangeability, positivity

$$
\begin{aligned}
E[Y^a] &= \sum_w E[Y^a|W=w]\Pr(W=w) && \text{Total Exp} \\
&= \sum_w E[Y^a|A=a,W=w]\Pr(W=w) && \text{Exch \& Pos} \\
&= \sum_w E[Y|A=a,W=w]\Pr(W=w) && \text{Consistency}
\end{aligned}
$$

Hereafter assume identification assumptions are true[1]

[1] This is the best case, but could also be consider when not met

Identification is *Not* Enough[3]

Suppose W is high-dimensional

- Continuous variable or many categorical variables
- $\Pr(A = a|W = w)$ and $E[Y|A = a, W = w]$
  - Cannot nonparametrically estimate
  - Informally,[2] there will always be a w with $A = 0$ but no units with $A = 1$ even as $n \to \infty$
  - Common is all but the simplest applications

To make progress, we use models

[2] This is the case asymptotically when W is continuous

[3] Maclaren & Nicholson (2019). 'What can be estimated? Identifiability, estimability, causal inference and ill-posed inverse problems' *arXiv*

Causal Effect Estimation with Models

Using models to estimate

- Define a narrower set of distributions; $\mathcal{M}_\alpha$, $\mathcal{M}_\beta$

$$\Pr(A = a | W = w; \alpha) = \pi_a(W; \alpha)$$

$$E[Y | A = a, W = w; \beta] = m_a(W; \beta)$$

- Allows us to interpolate or extrapolate over sparsity

AIPW

Outcome process

$$\hat{\psi}_a = \frac{1}{n} \sum_{i=1}^{n} m_a(W_i; \hat{\beta}) + \frac{Y_i - m_a(W_i; \hat{\beta})}{\pi_a(W_i; \hat{\alpha})}$$

Propensity score

Causal Effect Estimation with Models

Use of models *is not free*

- No model misspecification

$$\Pr(A = a|W = w) \in \mathcal{M}_\alpha$$

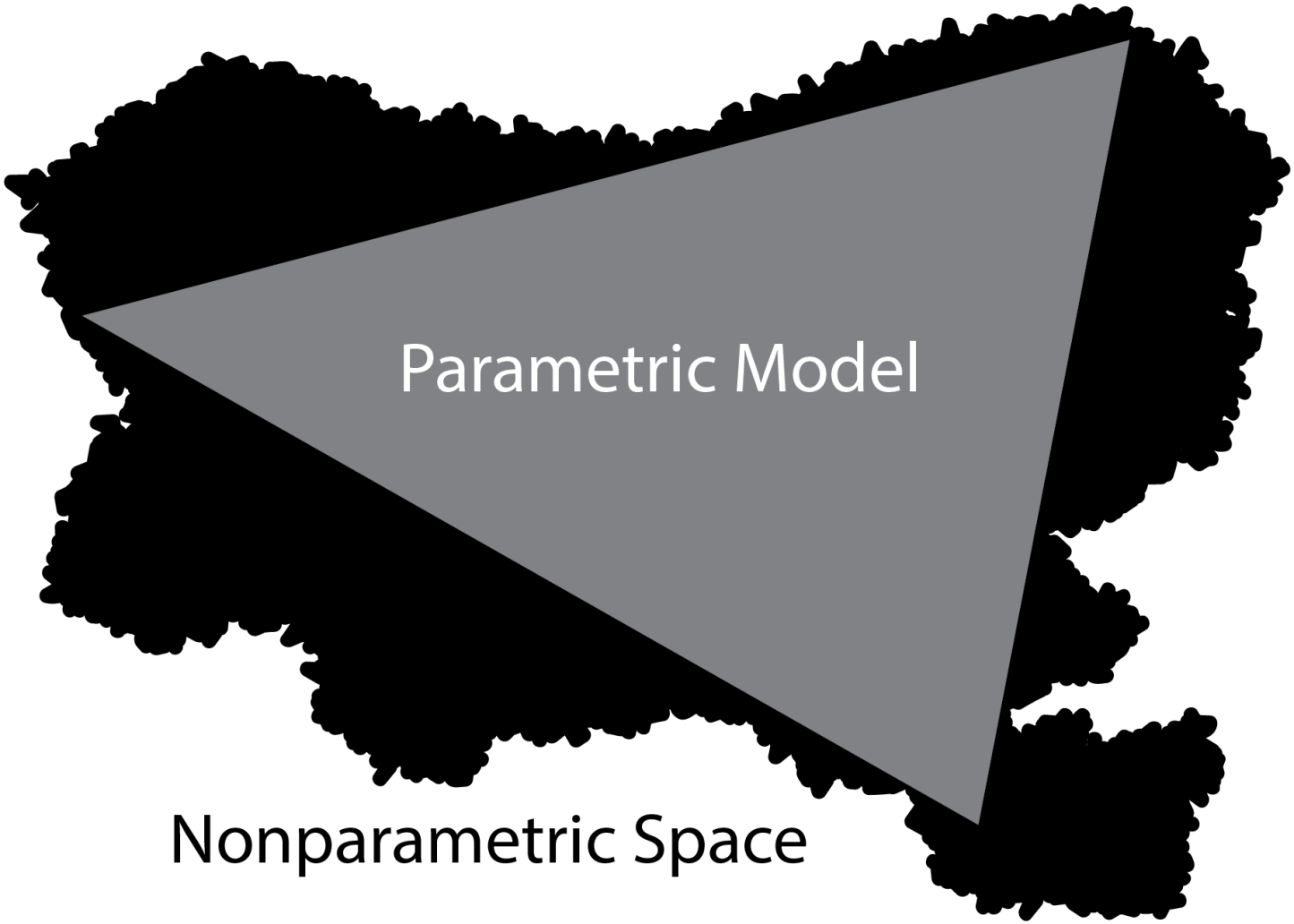
$$E[Y|A = a, W = w] \in \mathcal{M}_\beta$$

- Models include correct functional forms
  - Interaction terms, variable transformations, etc.

Epidemiologists commonly use *parametric* models

- Quite restrictive assumptions on functional forms

Causal Effect Estimation with Models[4]



[4] Coverage of parametric models not to scale

The Promise of Machine Learning

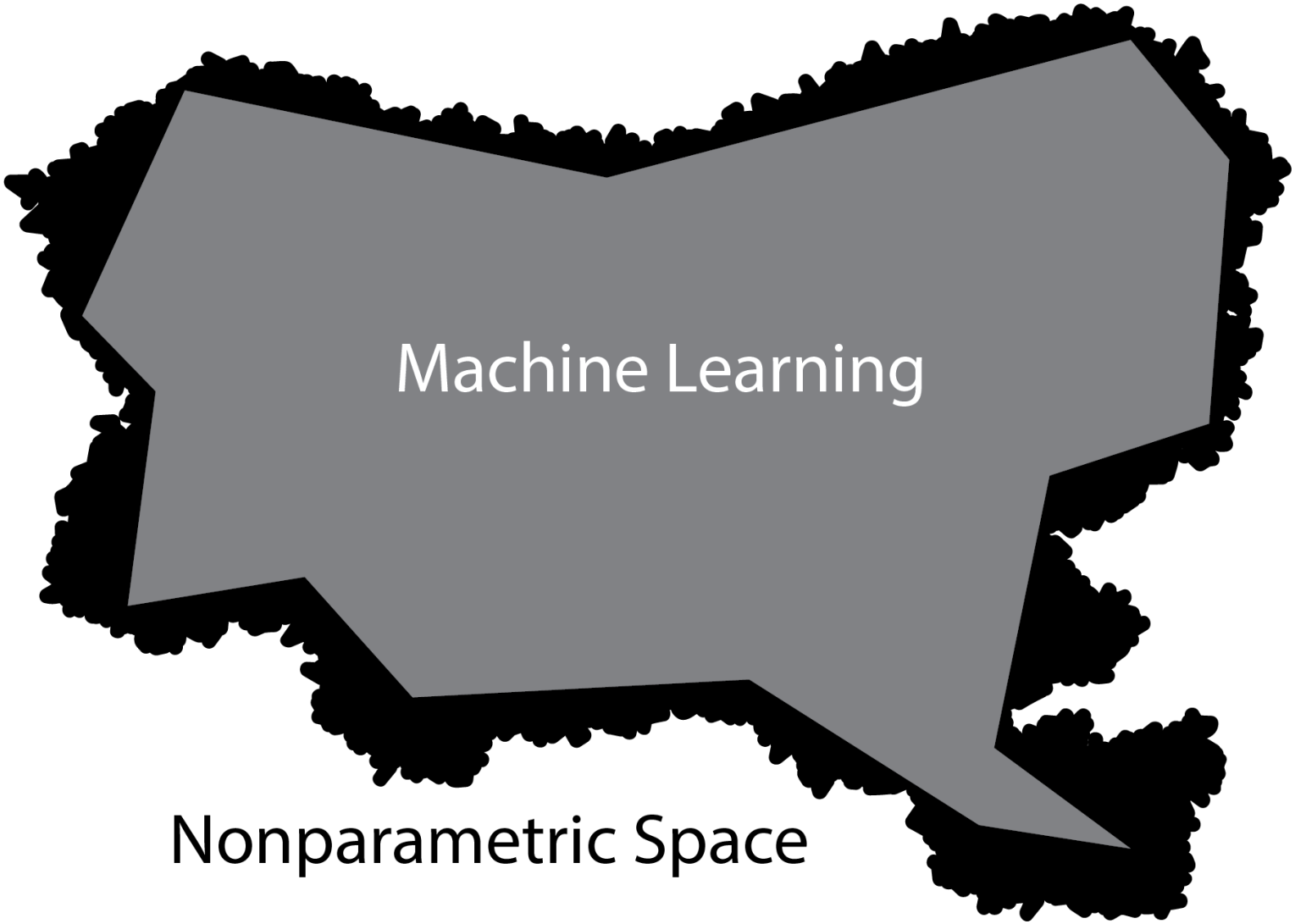
The assumption of no model misspecification is worrisome

- Are parametric models flexible enough?
- Do epidemiologists specify them to that level of flexibility?

Machine learning

- More flexible than standard parametric approaches
  - Captures a wider set of distributions
- Removes some burden from the researchers
  - May not need to specify interactions or non-linearities

Causal Effect Estimation with Models[5]



[5] Coverage of ML models not to scale

But Does Machine Learning Work?

Does machine learning allow us to estimate causal effects we couldn't otherwise *in practice*?

- Would a flexible penalized parametric model work similarly?

Does machine learning give us a 'better' estimator for the ACE?

- Is the computational complexity worth it?
- Can I reasonably trust these black box algorithms?

Secondary questions: best practices for application

Forms of Evidence to Consider

Mathematical Proof (deductive)

Simulation (inductive)

Mathematical Proof

Given a set of assumptions[6]

- Does the tool work?

Population inference

- Random sample of population
- Asymptotic results
  - Behavior as $n \to \infty$
  - Given large amounts of data, our method *should* work
  - If it doesn't, suspect for any realistic n

[6] Learn deductively rather than inductively

Mathematical Proofs: Machine Learning

Key results[7]

- Machine learning can capture a broader range of distributions
- Two concerns for machine learning application
  - Statistical convergence rates
    - Flexibility / convergence trade-off
    - Solution: AIPW / TMLE
  - Complexity
    - Limit complexity as to prevent over-fitting
    - Solution: sample-splitting / cross-fitting

But does this additional flexibility matters *in practice*?

[7] Reviewed in Zivich, Breskin, Kennedy (2023). 'Machine Learning and Causal Inference'. In *Wiley StatsRef*

Limits of Mathematical Proofs

Asymptotic results

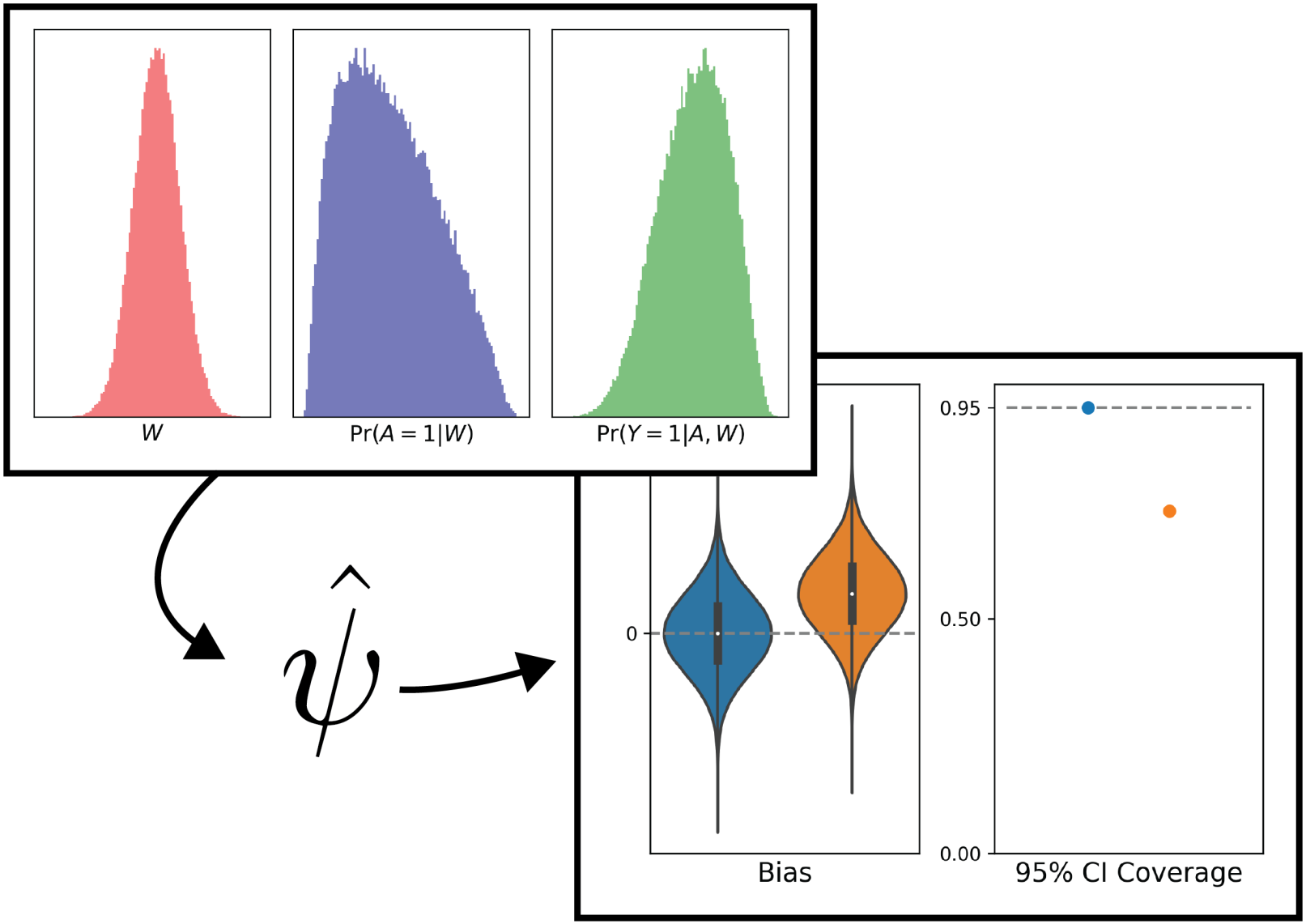
- What can this tell us about practical application?
  - Good asymptotic properties $\neq$ good in practice
  - Finite-sample bias can be too large
- Therefore not the full picture

Limits of Mathematical Proofs

"Such considerations reinforce the notion that (even if supporting theorems are available) great caution is needed in using a rule or method without extensive simulations to investigate the conditions under which it might be reasonable for practice. [...]
[E]pidemiologic inference [...] may come to rely on computations and simulations tailored to the specifics of the study context, rather than rely solely on general results or methods" [8]

[8] Greenland (2012) 'Commentary: Intuitions, Simulations, Theorems: The Role and Limits of Methodology' *Epidemiology*

Simulation



Simulation: Parametric

$$W \sim \text{Normal}(\mu, \sigma)$$

$$A \sim \text{Bernoulli}(\text{expit}(\alpha_0 + \alpha_1 W))$$

$$Y \sim \text{Bernoulli}(\text{expit}(\beta_0 + \beta_1 A + \beta_2 W))$$

Limitations

- Idealized example
  - No reason to presuppose the world adheres to parametric models
  - Too abstracted to be relevant to practice
- An unfair comparison?
  - May set up parametric models to succeed
  - Machine learning looks less impressive

Simulation: Parametric[9]

eAppendix 1; http://links.lww.com/EDE/B782. The incidence of statin use (X) was chosen to be similar to reported empirical trends in US adults,[21] and generated from the following model inspired by the 2018 primary prevention guidelines:

$$\begin{aligned}\Pr(X=1\mid Z) = \text{Bernoulli}\big(\text{expit}\big(&-3.471+1.390\,D_i+0.112\,L_i\\&+0.973\,I\big(L_i>\ln(60)\big)-0.046\,(A_i-30)\\&+0.003\big(A_i-30\big)^2+0.273\,I\big(0.05\le R_i<0.075\big)\\&+1.592\,I\big(0.075\le R_i<0.2\big)+2.641\,I\big(R_i\ge 0.2\big)\big)\end{aligned}$$

The ASCVD potential outcomes under each potential value of X were generated from the following model:

$$\begin{aligned}\Pr\big(Y^x=1\mid Z\big) = \text{Bernoulli}\big(\text{expit}\big(&-6.25-0.75x+0.35x\big(5-L_i\big)I\big(L_i<\ln(130)\big)\\&+0.45\big(A_i-39\big)^{0.5}+1.75D_i+0.29\exp\big(R_i+1\big)\\&+0.14\,I\big(L_i>\ln(120)\big)L_i^2\big)\big)\end{aligned}$$

[9] Zivich & Breskin (2021) 'Machine Learning for Causal Inference: On the Use of Cross-fit Estimators' *Epidemiology*

Simulation: Parametric[10]

$$P(X = 1 \mid C) = \text{expit}\{-1 + \log(1.75)C_1 + \log(1.75)C_2 + \log(1.75)C_3 + \log(1.75)C_4\}, \quad (7)$$

A continuous outcome was generated as:

$$Y = 120 + 6X + 3C_1 + 3C_2 + 3C_3 + 3C_4 + \epsilon, \quad (8)$$

where the true average treatment effect $\psi = 6$, with ϵ drawn from a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 6$.

Data Generating Mechanism: Model Misspecification

To induce model misspecification, we followed previous research[30] and transformed each of the continuous confounders as follows:

$$Z_1 = \exp(C_1/2)$$

$$Z_2 = C_2/(1 + \exp(C_1)) + 10$$

$$Z_3 = (C_1 C_3/25 + 0.6)^3$$

$$Z_4 = (C_2 + C_4 + 20)^2$$

[10] Naimi, Mishler & Kennedy (2021) 'Challenges in Obtaining Valid Causal Effect Estimates with Machine Learning Algorithms' *Am J Epidemiol*

Simulation: Plasmode

Plasmode simulations use a "dataset that is created from natural processes but has some aspect of the data-generating model known"[11]

- Tailored to a specific application
- Reference (truth) can be easily computed using known model
- Improvement over previous approach

Limitations

- Use of a known model to generate the data
  - Places us back in the same criticism

[11] Quote from Franklin et al. (2014) 'Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases' *Comput Stat Data Anal*

Simulation: Plasmode[12]

2. Fit the PS model using the data. Use the estimated coefficients to re-sample treatment variable, but modifying the intercept of the treatment variable to preserve observed treatment prevalence.
3. Estimate coefficients based on the OM from the whole data. Manually set the main coefficient for the treatment variable to the desired ATE (e.g. 6.6 a plausible increase in child weight due to maternal obesity status). Interaction terms, if involved, remain intact.
4. Generate the outcome using the OM with modified treatment coefficients and add error terms by randomly sampling the residuals of the OM with replacement.

[12] Meng & Huang (2021) 'REFINE2: A tool to evaluate real-world performance of machine-learning based effect estimators for molecular and clinical studies' *arXiv*

Beyond Plasmode Simulations

Generate simulation data where

- Avoid simple parametric distributions
- But still easily compute the reference (true) ACE

Beyond Plasmode Simulations

Credence[13]

- Variational autoencoder neural networks

Wasserstein Generative Adversarial Neural Networks[14]

- Pair of neural networks to mimic data

[13] Parikh et al. (2022) 'Validating causal inference methods. In International Conference on Machine Learning' *PMLR*

[14] Athey et al. (2021) 'Using Wasserstein generative adversarial networks for the design of Monte Carlo simulations' *Journal of Econometrics*

A Closing Thought

Machine learning for causal inference cannot cover every model

- In general[15]
- Specific mechanisms[16]

Some lingering concerns

- Need to *a priori* rule out certain mechanisms
- Using the tools to prove the tools works
  - Do we need something more general than what we seek to prove?

[15] Maclaren & Nicholson (2019). 'What can be estimated? Identifiability, estimability, causal inference and ill-posed inverse problems' *arXiv*

[16] Aronow et al. (2021). 'Nonparametric identification is not enough, but randomized controlled trials are' *arXiv*

Acknowledgements

Supported by NIH T32-AI007001

pzivich@unc.edu pzivich