Efficient machine learning for causal effects

Paul Zivich

University of North Carolina at Chapel Hill

March 15, 2021

Outline

Quick Note on Efficiency

Efficiency refers to *statistical* efficiency of an estimator

- Low mean squared error

Cross-fitting procedures are not *computationally* efficient

Notation

Y_i: observed outcome of interest for individual i

X_i: observed exposure or treatment of interest

$Y_i(x)$: potential outcome under exposure set to x

Z_i: covariate(s)

UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

Motivating Problem

Our collaborators want us to help them address what the difference in the risk of atherosclerotic cardiovascular disease (Y) had been if everyone was given statins ($x = 1$) compared to if no one was given statins ($x = 0$)?

$$\psi = E[Y(x = 1)] - E[Y(x = 0)]$$

They are planning on collecting some observational (non-investigator randomized) data to address this.

Identification

Identifiability Assumptions

Need to determine under what assumptions the average causal effect is *identifiable*

One set of identifiability assumptions

- Causal consistency
- Conditional exchangeability
- Positivity

Causal Consistency

An assumption regarding the potential outcomes observed

- We are able to observe at least one for some individuals

For X can write as

$$Y_i = Y_i(x) \text{ if } x = X_i$$

or as the treatment-variation irrelevance analog

$$Y_i = Y_i(x, b) \text{ for all } b \in B$$

where b could be versions of statins (e.g., 10mg vs 100mg)

Conditional Exchangeability

Synonyms

- No unmeasured confounding
- Potential outcomes of ASCVD are conditionally independent of statin use
- No unmeasured common causes of statins and ASCVD

Requires substantive knowledge outside of the data

- Directed Acyclic Graph
- Single-World Intervention Graph

Positivity

Conditional exchangeability requires that the probability for all values of statins to be non-zero in all strata of the confounders

$$\text{if } \Pr(Z = z) > 0 \text{ then } \Pr(X = x | Z = z) > 0$$

Positivity violations

- Some individuals never have access to statins
- Example: statins for children under 18 without inherited hypercholesterolemia

UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

Estimation

Motivating Problem

After discussion with our collaborators, the following covariates are determined to be confounders:

- Age
- Low-density lipoprotein (LDL)
- American Heart Association's ASCVD risk scores
- Diabetes

Model-free Causal Inference

When Z is low-dimensional

- No model is required
- Directly apply the non-parametric g-formula
- Unlikely in practice

Our Z includes continuous covariates (LDL, ASCVD risk scores)

- Model-free is not an option

UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

High-Dimensional Data

High-dimensional data

- Continuous Z or more strata of z than n
  - LDL can range from less than 100 to more than 190
- Sparse data in comparison to data dimension

How can we make progress?

- Categorize
  - E.g., LDL as $<$140, 140-159, 160-189, 190$+$
- Model

Categorize

Model

Model is an *a priori* restriction on the distribution of the data

- Parametric model restricts to a parametric distribution
- Not necessary for identification
- Adds information not in the data

Two processes can choose to model (nuisance models)

- exposure-model: $\Pr(X|Z;\alpha) = \pi(Z;\alpha)$
- outcome-model: $\Pr(Y|X = x, Z;\beta) = m_x(Z;\beta)$

No Model Misspecification

Addition of information via a model is not free

- Requires the addition of assumption on model specification
- True density is within model's ($\mathcal{M}$) class of densities
- exposure-model: $\Pr(X|Z) \in \mathcal{M}_\alpha$
- outcome-model: $\Pr(Y|X,Z) \in \mathcal{M}_\beta$

Implications

- Suggests use of flexible models

Doubly Robust Estimators

Clever combination of $\pi(Z; \alpha)$ and $m_x(Z; \beta)$

- Consistent as long as one model is correct
- Two chances for parametric model

Common doubly-robust estimators

- Augmented inverse probability weighting
- Targeted maximum likelihood estimation

UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

Machine Learning

Data-adaptive estimators (machine learning)

- Less restrictive than parametric modeling
  - Capture wider class of densities
- May be more reasonable to believe flexible enough model
  - Or that model is sufficiently close

Problems in Application

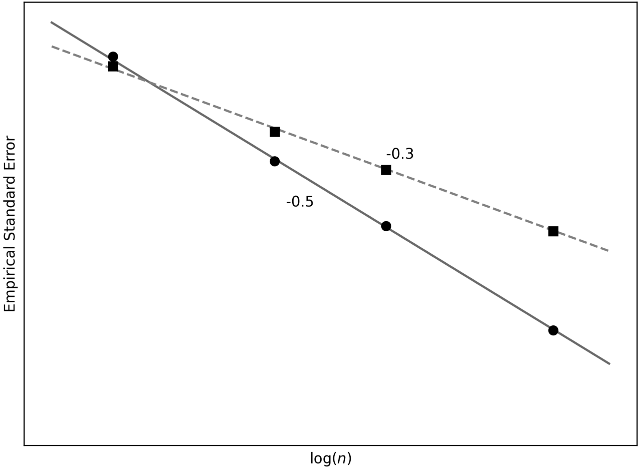
1. Convergence
2. Complexity

Convergence

Remainder term goes to zero as function of n

- *Not* computational complexity issue
  - **NOT** when software says 'did not converge'
  - Inherent feature of the estimators
- How fast standard error of estimators decreases with n
- For inference, root-n ($n^{-1/2}$) is desirable
  - Parametric models meet this criteria
  - Data-adaptive likely don't
- Slower than $n^{-1/2}$ may result in invalid inference
  - Estimated variance is too small
  - Implies greater certainty than true

Convergence



Convergence and AIPW

AIPW can allow for slower convergence

- Under the assumption that both models are correct
  - Second-order bias is a product of the approximation errors
  - Only need at least $n^{-1/4}$ for both models

- Wider range of models can be validly used with AIPW
  - Comes at cost of double-robustness
  - Becomes more akin to **double-susceptible**
  - Maybe okay if we use flexible models?

Complexity

Restrictions on the complexity of a model

- Allows borrowing of information across observations
- Restricted to Donsker class

Why restricting to Donsker is inappropriate

- Machine learning
  - Dimension allowed to increase with n
  - Exist in highly complex spaces

GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

Cross-fitting

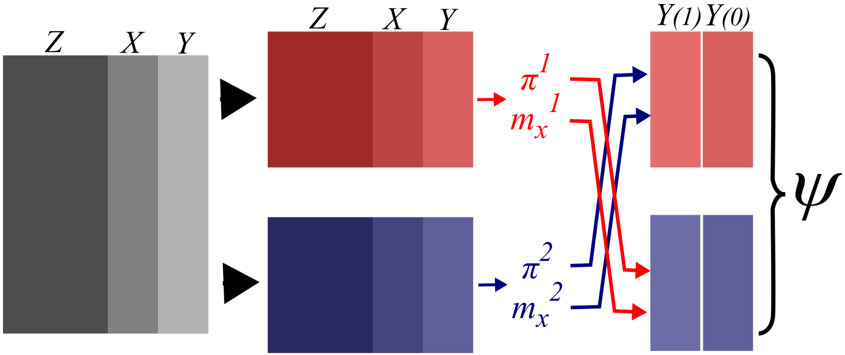
Approach to relax Donsker class restriction

- Estimate model in one split and predict in other split
- Since from different sample
  - Avoid the Donsker class restriction

Synonyms

- Sample splitting
- Double machine learning
- Cross-validated

Single Cross-fit



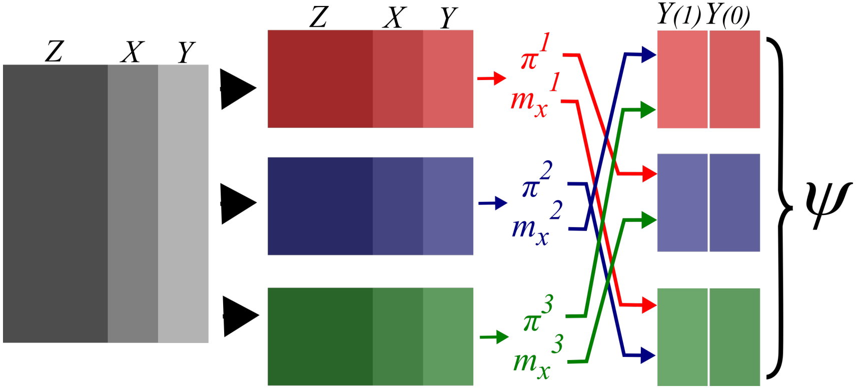
Single Cross-fit

Removes own-observation bias

- Best data-adaptive estimator memorizes the data
  - Highly predictive for the data
  - Poor out-of-sample performance
- Cross-fit prevents correlation between estimator and data
  - Without need for Donsker class

GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

Double Cross-fit



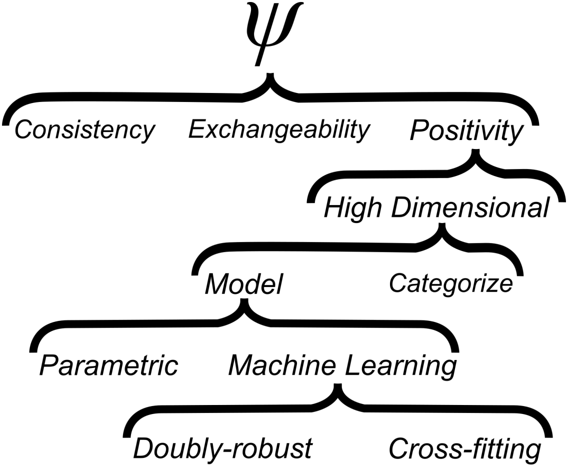
Double Cross-fit

Removes own-observation bias & non-linearity

- Cross-fit prevents correlation between estimator and data
  - Without need for Donsker class

- Further de-couples data and estimators
  - Compared to single cross-fit

Looking Back...



Simulations

Data Generation

Average Causal Effect of Statins on ASCVD

- Age, low-density lipoprotein, ASCVD risk scores, diabetes
- 2000 reps with $n = 3000$

Metrics

- Bias
- 95% Confidence interval coverage

Approaches

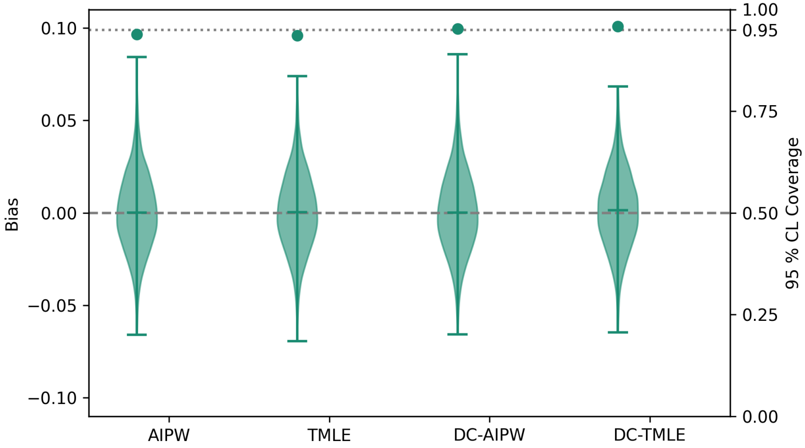
ACE estimators

- AIPW
- TMLE
- Double Cross-fit (DC-)AIPW
- DC-TMLE

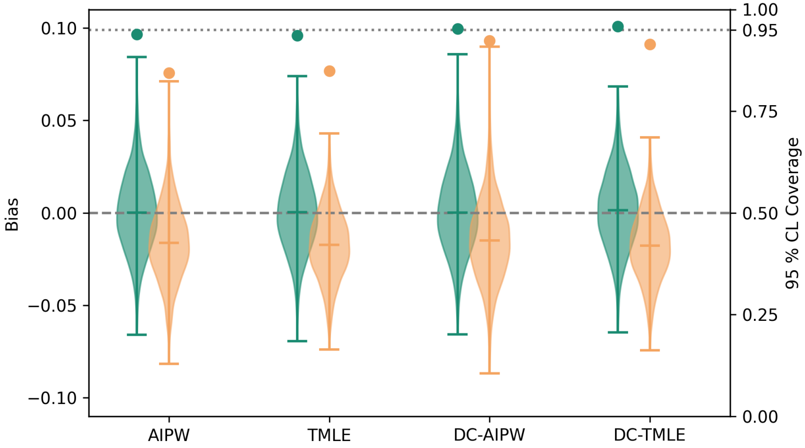
Nuisance model estimators

- Correct parametric model
- Main-effect parametric model
- Machine learning

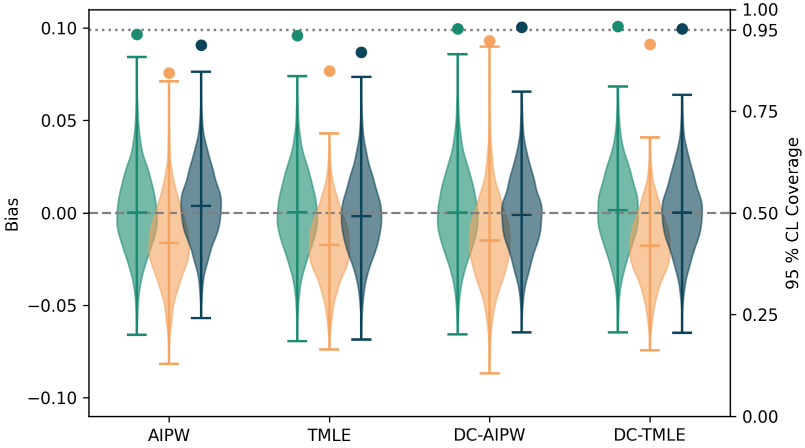
Simulation Results



Simulation Results



Simulation Results



GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

Implementation

GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

Partition versus Split

Split

- Observations grouped into non-overlapping, equal-sized groups
- Let s indicate the number of splits for a data set
  - $s = 2$ randomly splits observations into two groups

Partition

- A particular division of splits
- E.g., IDs 1,2,4 are in split 1 and IDs 3,5,6 are in split 2

Algorithm Pseudo-Code

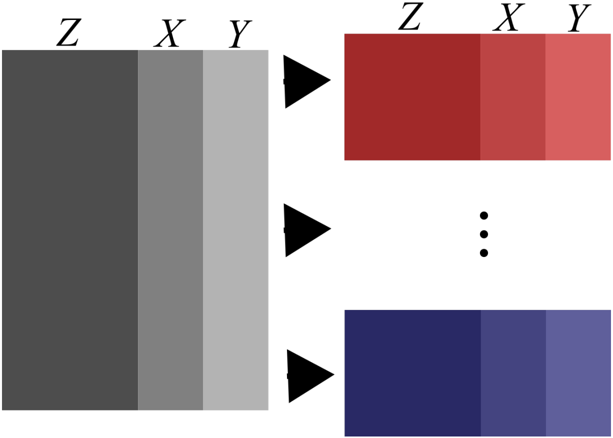
Repeat for p different partitions

1. Partition data into s non-overlapping splits
2. For each s:
   - Estimate $\hat{\pi}^s(Z_i^s)$
   - Estimate $\hat{m}_x^s(Z_i^s)$
3. For each s:
   - Predict $\hat{\pi}^{\bar{s}}(Z_i^s)$
   - Predict $\hat{m}_x^{\bar{s}}(Z_i^s)$
   - Generate $\hat{Y}_i^*$
   - Estimate ψ^s
   - Estimate $Var(\psi^s)$
4. Estimate ψ^p as mean of $\hat{\psi}^s$
5. Estimate $Var(\psi^p)$ as mean of $\widehat{Var}(\hat{\psi}^s)$

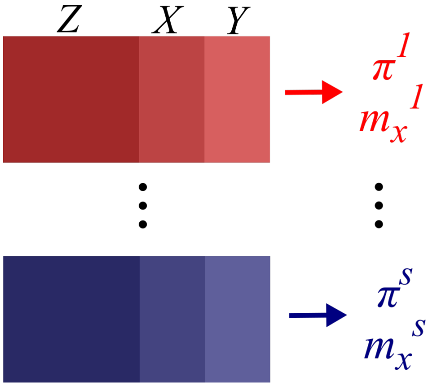
Summarize all partitions with
$\hat{\psi} = \text{median}(\hat{\psi}^p)$; $\widehat{Var}(\hat{\psi})$

UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

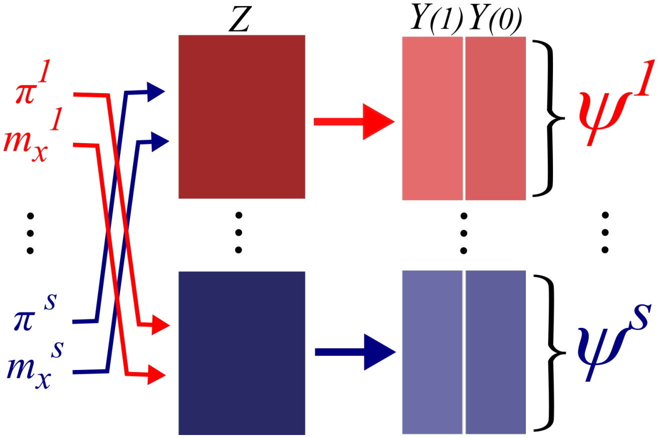
Step 1: Partition Data into Splits



Step 2: Estimate Nuisance Models



Step 3: Estimate ψ^s



Step 3: Estimate ψ^s

Augmented Inverse Probability Weighting

$$\widehat{Y}_i^*(x) = \frac{I(X_i = x)Y_i}{\widehat{\pi}(Z_i)} + \frac{\widehat{m_x}(Z_i)(\widehat{\pi}(Z_i) - I(X_i = x))}{\widehat{\pi}(Z_i)}$$

Then calculate

$$\psi^s = \frac{1}{n_s} \sum_{i \in n_s} \widehat{Y}_i^*(x = 1) - \widehat{Y}_i^*(x = 0)$$

Step 4: Estimate ψ^p

Take the mean of all ψ^s

$$\hat{\psi}^p = \frac{1}{s} \sum_{i=1}^{s} \hat{\psi}^s$$

Step 5: Estimate $Var(\psi^p)$

Take the mean of all $Var(\psi^s)$

$$\widehat{Var}(\hat{\psi}^p) = \frac{1}{s} \sum_{i=1}^{s} \widehat{Var}(\hat{\psi}^s)$$

Repeat for p different partitions

Repeat previous steps for p times

Summarize all different partitions used via

$$\hat{\psi} = \text{median}(\hat{\psi}^p)$$

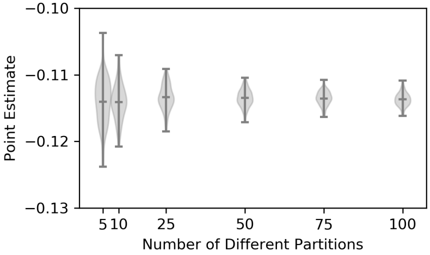
$$\widehat{Var}(\hat{\psi}) = \text{median}\left(\widehat{Var}(\hat{\psi}^p) + (\hat{\psi} - \hat{\psi}^p)^2\right)$$

- Median rather than mean since more stable to outliers

Why p Partitions?

Multiple splits

- Does not matter asymptotically
- For moderate n, $\hat{\psi}^p$ depends on the chosen split
  - Increases *statistical* efficiency
  - Decreases *computational* efficiency



Recommendations for Practical Application

1. Flexible library of learners
    - *k*-fold super-learner
    - Variety of models
        - Flexible regression, tree-based, gradient, etc.
    - Explore multiple tuning parameters
2. Doubly-robust estimator
3. Cross-fitting
4. Include transformations in data

Software

Python

- *zEpid*
  - https://github.com/pzivich/zEpid
  - v0.9.0+ Supports single and double cross-fit
  - Support for both AIPW and TMLE
- https://github.com/pzivich/publications-code

R

- https://github.com/yqzhong7/AIPW
- https://github.com/pzivich/publications-code

Conclusions

Conclusions

Machine learning is a useful tool for nuisance model specification

- Allow for flexible models
- Less concern over misspecification relative to parametric models
  - Captures wider set of densities
- Best benefit in my opinion
  - Allows us to devote time to other biases

> Epidemiology. 2021 Feb 2. doi: 10.1097/EDE.0000000000001332. Online ahead of print.

Machine learning for causal inference: on the use of cross-fit estimators

Paul N Zivich [1], Alexander Breskin

Affiliations + expand

PMID: 33591058 DOI: 10.1097/EDE.0000000000001332

Further Reading

Doubly-robust estimators generally

- Daniel, Double Robustness, In: *Statistics Reference Online* Wiley 2014
- Robins & Ritov, Towards a Curse of Dimensionality Appropriate (CODA) Asymptotic Theory for Semi-Parametric Models, *Statistics in Medicine* 1997
- Kennedy, Semiparametric theory and empirical processes in causal inference. In: *Statistical Causal Inferences and their Applications in Public Health Research* Spring 2016 p141-167

Augmented inverse probability weighting

- Funk et al., Doubly Robust Estimation of Causal Effects. *American Journal of Epidemiology* 2011, 173:7 p761-767
- Keil et al., Resolving an Apparent Paradox in Doubly Robust Estimators. *American Journal of Epidemiology* 2018, 187:4 p891-892
- Bang & Robins, Doubly Robust Estimation in Missing Data and Causal Inference Models. *Biometrics* 2005, 61:4 p962-973

Targeted maximum likelihood estimation

- Schuler & Rose, Targeted Maximum Likelihood Estimation for Causal Inference in Observational Studies. *American Journal of Epidemiology* 2017, 185:1 p65-67

UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

Further Reading

Machine Learning and Super learner

- Bi et al., What is Machine Learning? A Primer for the Epidemiologist. *American Journal of Epidemiology* Oct 2019, 188:12 p2222-2239
- Rose, Mortality Risk Score Prediction in an Elderly Population Using Machine Learning. *American Journal of Epidemiology* 2013, 177:5 p443-452
- Naimi & Balzer, Stacked Generalization: an Introduction to Super Learning. *European Journal of Epidemiology* 2018, 33:5 p459-464
- Keil & Edwards, You Are Smarter Than You Think: (Super) Machine Learning in Context. *European Journal of Epidemiology* 2018, 33:5 p437-440

Sample-splitting for machine learning

- Díaz, Machine learning in the estimation of causal effects: targeted minimum loss-based estimation and double/debiased machine learning, *Biostatistics* April 2020, 21:2 p353–358
- Chernozhukov et al., Double/debiased machine learning for treatment and structural parameters, *The Econometrics Journal* 2018 21, C1–C68
- Zheng & van der Laan, Cross-validated targeted minimum-loss-based estimation. In: *Targeted Learning* Springer 2011 p459–474
- Newey & Robins, Cross-Fitting and Fast Remainder Rates for Semiparametric Estimation, *arXiv*:1801.09138

Machine learning for causal inference

- Naimi et al., Challenges in Obtaining Valid Causal Effect Estimates with Machine Learning Algorithms. *arXiv*:1711.07137

Acknowledgments

pzivich@live.unc.edu @PausalZ pzivich

PNZ is supported by NICHD T32-HD091058