Does sample splitting help when estimating causal effects using double machine learning methods?

TJ Tang, Qin He

2022-04-24

## 1. Introduction

Sample splitting is one of the key issues when adapting machine learning methods to causal inference problems. However, there are some anecdotes saying that sample splitting will not help in finite samples. Therefore, this paper study will first go through Chernozhukov et al. (2018) especially about an example that sample splitting reduces the bias, and then will conduct a few simulation studies to figure out the effect of sample splitting when estimating causal effects.

## 2. Double Machine Learning and Sample Splitting

Chernozhukov et al. (2018) establish that in very high dimension cases, machine learning methods will employ regularization to reduce variance but cause the regularization bias and overfitting on the parameters of interest or causal treatment effect parameter, for example average treatment effect. They show that the bias and overfit can be removed by double machine learning methods with sample splitting and cross fitting. Applying double machine learning methods can overcome the regularization bias because it employs an orthogonalized formulation. Meanwhile, sample splitting and cross fitting can reduce the bias due to the overfitting because we can avoid using the same data twice, to train the estimator and predict the missing potential outcomes. The details for orthogonalization and why double machine learning methods reduce the regularization bias is not the focus of this project. Instead, this project will focus more on the sample splitting and the effect to reduce overfitting bias.

Wager and Athey (2018) illustrated double-sample trees as a motivation to reduce bias. They divided training subsamples into two halves and used one half for split selection and the other for within-leaf estimation. Though sampling splitting is sometimes criticized as inefficient as they do not use the full training data at each estimation steps; however, they argued that by aggregating over multiple trees, each sample will be used for split selection or leaf estimation. By doing so, therefore, we can efficiently achieve honesty criterion without wasting half of the samples.

Wager and Athey’s ideas are similar to sample splitting and cross fitting in double machine learning. To perform the sample splitting to control for bias, we need to first split data into main sample and auxiliary sample, and then use the main sample to train the auxiliary sample to estimate. Next, we swap the role of main sample and auxiliary sample to estimate. Finally, we combine the results by averaging over the two estimators. By doing so, we can achieve a good estimator without losing efficiency. The idea of sample splitting is an analogy to -fold cross validation and can be also generalized to -fold splitting and cross-fitting. For -fold cases, we take a -fold random partition of data rather than splitting into half, and at the end we aggregate over machine learning estimators so that we use the full data. In terms of the choice of , Chernozhukov et al. (2018) recommends using moderate values such as 4 or 5.

## 3. Set up of Double Machine Learning with Partial Linear Model

Consider the partial linear model (using the notations in Chernozhukov et al. (2018):

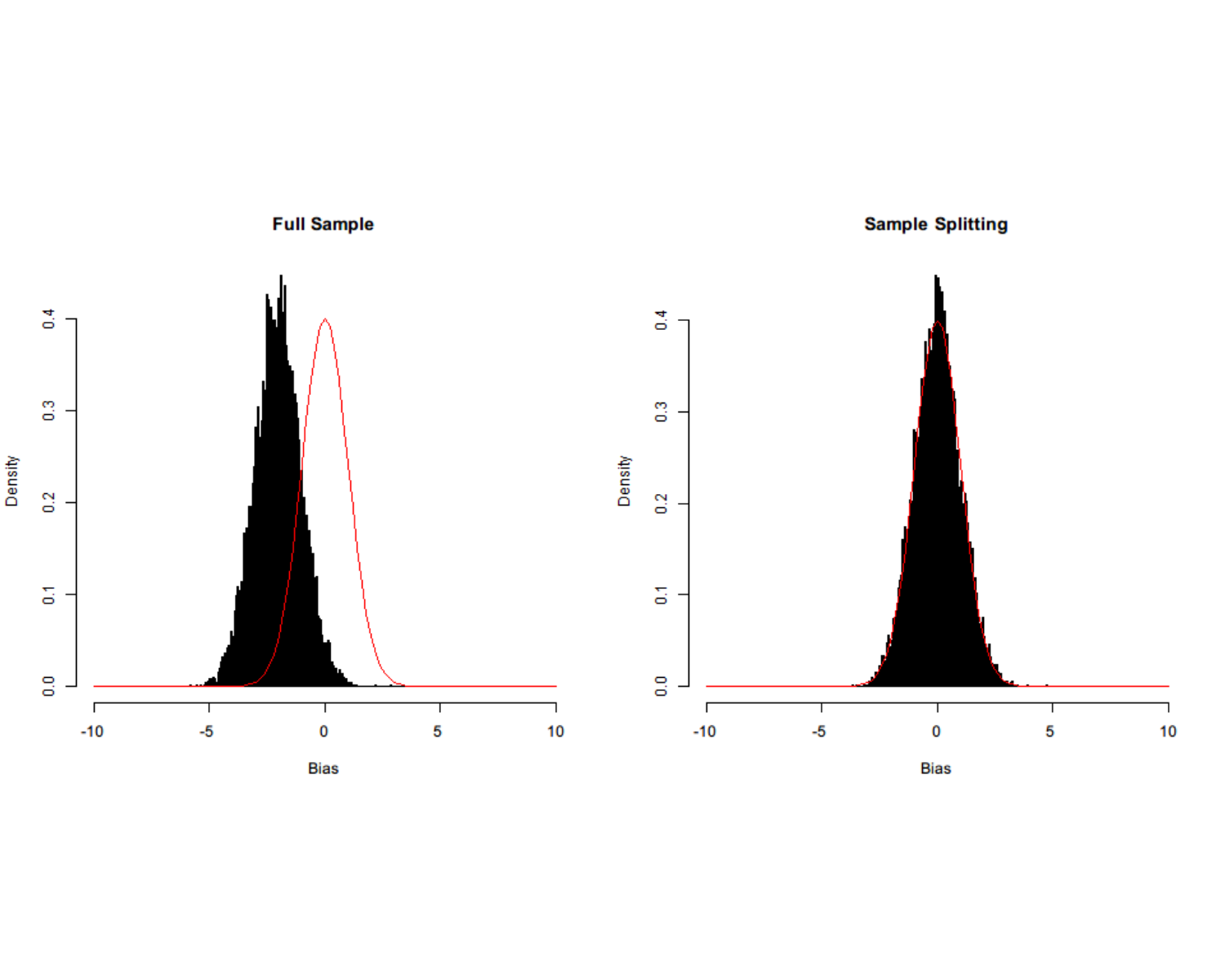
where is the outcome variable, is the treatment indicator, is a vector of confounders, and are disturbances, and is the causal estimand we would like to infer.

Next, consider the double machine learning procedure with sample splitting, following Chernozhukov et al. (2018):  
- Take a -fold random partition of observation indices such that the size of each fold is And define \ for each . - Using machine learning methods to predict and using X by and respectively.  
- Get residuals and   
- Regress on to get the estimate

Following the procedures, the estimator is a con

## 4. Replicate Chernozhukov et al. (2018)

We found the codes for generating the Figure 2 in Chernozhukov et al. (2018) on GitHub. We replicated and rewrote the codes in and generated a similar plot.



To illustrate that sample splitting can combat overfitting bias, Chernozhukov et al. (2018) uses an artificial example. Let and fix true be 1. And the data is generated by and , where and are error terms and follow . The overfit is specified as . More specifically, Then two estimators are calculated, one using the full sample, and the other adopting two-fold sample splitting and cross fitting. The double machine learning estimator is formulated as

Note that, when we use two-fold sample splitting and cross-fitting and if we partition our data into and , then for is calculated using the estimated with observations in , and vice versa. And we simply average over for and for to get our causal estimand using the full data. For the above plot, we fix . From the plots, we may observe that using the full sample and without sample splitting (left) may induce large bias; however sample splitting with cross fitting (right) can remove the bias very well.

## 5. Simulation Study

We explored 128 different scenarios in the data generation process with following configurations.  
- Number of covariates: 5 or 100  
- Sample size : 50 or 1000  
- Number of folds : 2, 5  
- Variance of the error term and : 0.1 or 1 - Distribution of coefficients for covariates : or - Outcome model: linear or non-linear with sine functions

We fixed the true causal effect be 1. And the machine learning methods we considered are LASSO and Extreme Gradient Boosting (XGBoost). We fit LASSO using glmnet package and the value of for LASSO is determined by 5-fold cross validation, where we compared from 0.001 to 100. We fit XGBoost using xgboost package. We did not tune XGBoost because we found the loss is small using the parameters provided in the documentation. For each scenario, we generated 100 different data sets and conduct double machine learning using the partial linear models with sample splitting and without sample splitting. The codes for simulations and the simulation results are provided.

## 6. Real Datasets

We applied the methods described in last section to a real dataset. It is the Right Heart Catheterization (RHC) dataset we used in Homework 3. The data is from Murphy and Cluff (1990), then re-analyzed by Connors et al. (1996) in an influential study. The data contains 5735 samples and 51 covariates. The outcome variable is dth30, which is the vital status at 30 days after admission. And the treatment indicator is TRUE if RHC was applied within 24 hours of admission, and FALSE if RHC was not applied within 24 hours of admission. We are interested in estimating the average treatment effect.

## 7. Conclusions

## References

* Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, James Robins, Double/debiased machine learning for treatment and structural parameters, The Econometrics Journal, Volume 21, Issue 1, 1 February 2018, Pages C1–C68, <https://doi.org/10.1111/ectj.12097>
* Stefan Wager & Susan Athey (2018) Estimation and Inference of Heterogeneous Treatment Effects using Random Forests, Journal of the American Statistical Association, 113:523, 1228-1242, DOI: 10.1080/01621459.2017.1319839
* Victor Chernozhukov DML Slides, “Double/Debiased Machine Learning for Causal and Treatment Effects,” May 31, 2018. (Accessed via Sakai)
* Seo M, White IR, Furukawa TA, et al. Comparing methods for estimating patient-specific treatment effects in individual patient data meta-analysis. Statistics in Medicine. 2021;40:1553–1573. <https://doi.org/10.1002/sim.8859>
* Codes for figures in Section 4: <https://github.com/VC2015/DMLonGitHub/blob/master/Figure2.m>
* Lecture Notes provided by Professor Fan Li (Assessed via Sakai)
* Datasets provided by Professor Fan Li (Assessed via Sakai)