Sample Splitting in Bayesian Approaches to Estimate Conditional Average Treatment Effects

STA640 Spring 2022, Final Project

Xige Huang, Youran Wu

2022-04-24

1. Introduction

To obtain valid and accurate estimation in treatment effects, two strategies of reducing model sensitivity are often considered, namely, balancing the covariates in the design stage and specifying a flexible outcome model. A Double-Robust (DR) approach in the frequentist context combines such two strategies and thus usually involves specification of both propensity score model and outcome model; recent literature has provided numerous flexible machine learning models as the outcome model. Additionally, sampling splitting is a key procedure in obtaining good estimates of the causal effects when using machine learning methods. Kennedy (2020) has proven that sample-splitting improves accuracy in estimating the Conditional Average Treatment Effect (CATE) using a DR-estimator.

The question of interest is whether sample splitting helps in the Bayesian paradigm. To answer this question, we design a simulation study with different data generating processes and adopt three methods in the Bayesian paradigm, among which two involve two-step estimation are can be seen as analogue to a DR-approach. Specifically, we will consider Bayesian Additive Regression Trees (BART), BART with propensity score as a covariates (BART-ps), and Bayesian Causal Forest (BCF); the latter two include estimating the propensity score and building the potential outcome model. We consider several different but related data generating processes: homogeneous and heterogeneous treatment effects, low and high dimensional covariates, sparse and non-sparse covariates matrix, randomized and non-randomized treatment assignment.

All implementation details and code can be found at our github repository github-Sample Splitting in Bayesian.

2. Literature Review

There is rich literature in recent years that focus on using flexible machine learning methods to estimate average or individual treatment effects. Machine learning methods such as penalized regression (LASSO, elastic net), trees (CART, random forest), boosting, Bayesian nonparametric models (BART, Gaussian Process, Dirichlet Process) that use regularization to decrease the variance of the estimator for causal effect. However, due to bias and variance tradeoff, the introduction of regularization and overfitting might make the prediction bias larger. Also, good performance in predicting propensity score or outcome model alone does not necessarily translate into good causal estimation.

To address this problem, we can use DML and sample splitting to help limit overfitting. The main idea of DML is to use machine learning methods in double-robust (DR) estimators, which means specifying machine learning models for both propensity score and outcome models (Farrell, 2015). The general process of sample splitting is to split the data into main sample and auxiliary sample, one for training and one for estimating the causal estimand of interest. Since we only use half of the data for estimation in sample splitting, we're not using the data efficiently. One way to solve this problem is cross fitting, which means we can swap the role of main sample and auxiliary sample and re-estimate. We can upgrade cross fitting to subsetting the data into K folds and average the K results. In Chernozhukov et al. (2018), he used a simulation study to show that sample splitting helps remove bias when used with DML for data generated by partial linear regression. We reproduce the simulation here. The machine learning methods we use here for propensity score estimation and outcome model are both random forest.

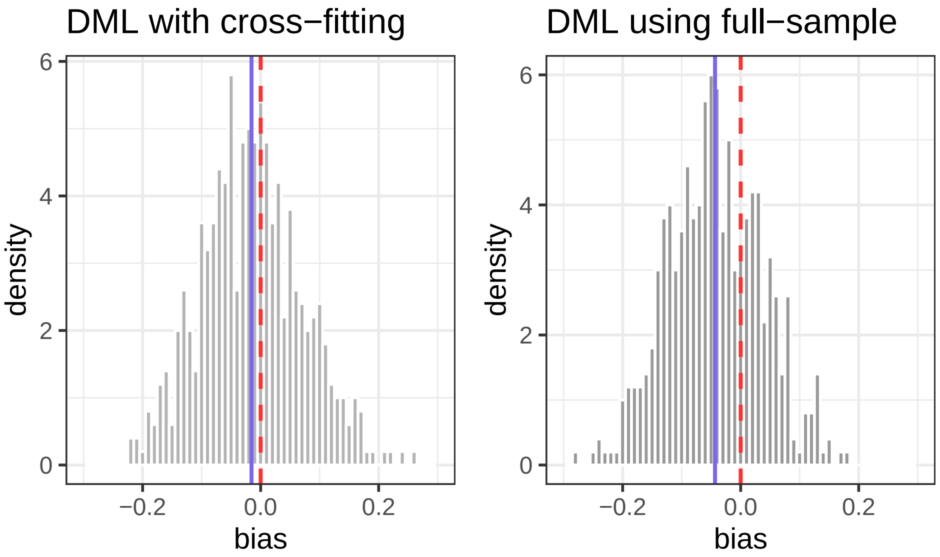


Figure 1: Comparison of full-sample and cross-fitting procesures, DML with Random Forests, blue line: mean of bias distribution

In this figure, we can clearly see the distribution of $\theta - \hat{\theta}$ in DML without cross fitting remarkably shifts to the left of 0 due to overfitting. The distribution of $\theta - \hat{\theta}$ in DML with cross fitting has a mean around 0, whereas distribution of bias using full sample clearly deviates from 0. We can also see that the spread of the two distributions are roughly equal, illustrating the efficiency of using data in cross fitting.

Kennedy (2020) proposes another sample splitting method which divides the sample into three folds. In the first stage, one fold is used to estimate the conditional expectation of the outcome, and another fold is used to estimate propensity score; in the second stage, estimations from the first stage are used to construct an estimate of the pseudo-outcome which is then regressed on X using the remaining fold. Kennedy (2020) has shown that under such double-splitting approaches, faster rates can be achieved in estimating CATE.

3. Method

3.1 Potential outcome framework and assumptions

We employ the Nayman-Rubin potential outcome framework throughout our simulation study. Specifically, $Z_i \in \{0,1\}$ is the treatment assignment indicator, $X_i \in \mathbb{R}^d$ is the covariates of d-dimensional, and $Y_i(Z)$ is the potential outcome of unit i when i is assigned into the treatment group for $Z = 1$ and control group for $Z = 0$. Our desired causal estimand CATE is defined as

$$\tau(x) = \mathbb{E}\left[Y(1) - Y(0) \mid X = x\right]$$

To ensure that our identification and estimation of $\tau(x)$ is valid, we need to make the following assumptions:

1. SUTVA (Stable Unit Treatment Value Assumption): a subject's potential outcome is not affected by other subjects' exposure to treatment and there are no different versions of treatment. This ensures counterfactual consistency.

2. Ignorability:
$$\mathbb{P}(Z_i = 1 \mid X_i, Y_i(0), Y_i(1)) = \mathbb{P}(Z_i = 1 \mid X)$$
This ensures that there is no unmeasured confounder.

3. Overlap:
$$0 < \mathbb{P}(Z_i = 1 \mid X_i, Y_i(0), Y_i(1)) < 1 \quad \forall i$$
This ensures that no subpopulation is entirely located in the treatment or control group.

3.2 BART, BART-ps, and BCF

BART (Hill 2011) is a flexible Bayesian non-parametric method that enables us to directly fit the response surface of potential outcome Y. In the frequentist context, Double-Robust (DR) estimator is a two-step procedure to estimate the treatment effect, usually including both the specification of model propensity score and the model on potential outcome. Although the role of propensity score in the Bayesian paradigm has been controversial, there is still literature showing that incorporating propensity score in Bayesian methods can improve estimation accuracy of the treatment effect. Specifically, estimates of treatment effects using BART are shown to be biased under situations where there exists strong confounding. This is called regularization-induced confounding proven by Hahn et al. (2018). Including propensity score as a covariate in BART (Zigler et al., 2013) will reduce such regularization-induced-bias. Bayesian causal forest proposed by Hahn, P. R., Murray, J. S., and Carvalho, C. M. (2020) is a further extension of incorporating propensity score in BART, where the model is specified as

$$f\left(\mathrm{x}_i, z_i\right) = \mu\left(\mathrm{x}_i, \hat{e}(X)_i\right) + \tau\left(\mathrm{x}_i\right) z_i$$

Bayesian causal forest is proven to outperform BART and BART-ps in accuracy of estimation of CATE especially when sample size is small and true treatment effects are in moderate-to-high homogeneity. Those two approaches can be seen as a Bayesian analogue to the double-robust approach. In our implementation, we will estimate the propensity score from a logistic regression of all covariates X on the treatment indicator Z.

We will compare under 10 different data generating processes detailed in the next section. We consider using cross fitting on our synthetic dataset. Theoretically, in order to get a consistent estimator, we would want our nuisance function (the underlying structural form for prediction of probability of treatment, which is equivalent to propensity score in our context) to have low entropy and is differentiable everywhere to fulfill the Donsker condition. (Chernozhukov et al. (2018)). With cross fitting, we use independent sets for estimating nuisance function and outcome model and thus can treat nuisance function as fixed functions, allowing us to bypass the condition on complexity. (Bickel (1982), Schick (1986)).

3.3 Sample splitting

First we consider sample-splitting in BART. Below is the detailed procedure:

1. Draw a sample from the generated data. For data with $n = 500$, we draw $n = 400$ rows of data without replacement; for data with $n = 2000$, we draw $n = 1500$ rows of data with replacement.

2. Split the data into two equal folds A and B. We use fold A to train a BART and fold B to predict outcome.

3. Repeat step 1 and 2 *M=20* times and average the predicted outcomes.

Next, we consider sample-splitting in BART with propensity score as a covariate and BCF, for which both of the methods need a two-step estimation of first estimating propensity score, and then use the estimated propensity score in the outcome regression. Below is the detailed procedure that is similar in Kennedy (2020):

1. Draw a sample from the generated data. For data with $n = 500$, we draw $n = 400$ rows of data without replacement; for data with $n = 2000$, we draw $n = 1500$ rows of data without replacement. (same as step one in BART with sample splitting)

2. Split the data into three equal folds A, B, and C. We use fold A to train a model that predicts propensity score. Then we use the obtained model to predict propensity score in fold B. After that, we use fold B to train a BART with the predicted propensity score as one of the covariates.

3. Use the outcome model trained on fold B to predict propensity scores and then outcomes in fold C.

4. Repeat step 1 and 2 *M=20* times and average the predicted outcomes.

4. Simulation

We will consider the following partially linear regression (PLR) model proposed in Robinson (1988) as our general data generation.

$$\begin{aligned} Y = \tau(X)Z + g(X) + U, &\quad \mathbb{E}[U|X, Z] = 0 \\ Z = e(X) + V, &\quad \mathbb{E}[V|X] = 0 \\ \tau(X) = t(X) + W, &\quad \mathbb{E}[W|X] = 0 \end{aligned}$$

In this model, Y is a continuous outcome variable, Z is the treatment status, $g(X)$ is a smooth function, $\tau(X)$ is the true treatment effect, and $e(X)$ is the propensity score. The covariates $X = (X_1, \ldots, X_p)$ is a p-dimensional vector and follows $X_i \sim \mathcal{N}(0, \Sigma)$ with Σ being a fixed correlation matrix. Error terms $U, V, W \sim \mathcal{N}(0, 1)$.

We consider 5 factors of data generating process, namely, large and small sample size, high and low dimensional, homogeneous and heterogeneous treatment effect, sparse and non-spare covariate matrix, and

Table 1: Data Generating Process Settings

	A/F	B/G	C/H	D/I	E/J
size	2000/500	2000/500	2000/500	2000/500	2000/500
dimension	50	50	50	10	10
treatment.effect	$\tau = 0.5$	$\tau = 0.5$	heterogenous	$\tau = 0.5$	heterogenous
sparsity	non-sparse	non-sparse	sparse	non-sparse	sparse
Z	RCT: e(X) = 0.5	RCT: e(X) = 0.2	dependent on X	RCT: e(X) = 0.2	dependent on X

randomized (including balanced and imbalanced treatment group assignment) and non-randomized treatment assignment. For all data generating processes, we set the outcome Y to be continuous, and let $g(X) = cos(X \cdot b)$ be a smooth function with $b = \frac{1}{l}$ for $l = \{1, 2, \ldots, k\}$. Our representation of the different scenarios of data generating process follows the table that efficiently describes all simulated data from Jacob 2019 (see Section 3, Simulation Study).

Regarding heterogeneous/homogeneous treatment in table 1, homogenous treatment means $\tau(X) = 0.5$, which is a fixed number across a synthetic dataset. Heterogeneous treatment means the treatment assignment is dependent on covariates and varies across datasets. In our setting, $\tau(X) = 1 + 2X_2X_5 + X_{10}$. To create sparse datasets, we randomly replace around 60% of each X_i by 0.

Concerning different treatment assignment mechanisms, we use the standard normal distribution's cumulative density function (CDF) to create probabilities and plug them in a binomial function to get binary treatment variables. We define $a(X)$ to be the dependence of covariates within the CDF. We consider three variations here: balanced group, imbalanced group, and linear dependency of treatment assignment on covariates.

Specifically, in case of random assignment:

$$e(X) = c \quad \text{with} \quad c = \begin{cases} 0.2, & \text{imbalanced assignment} \\ 0.5, & \text{balanced assignment} \end{cases}$$

In case of linear dependence:

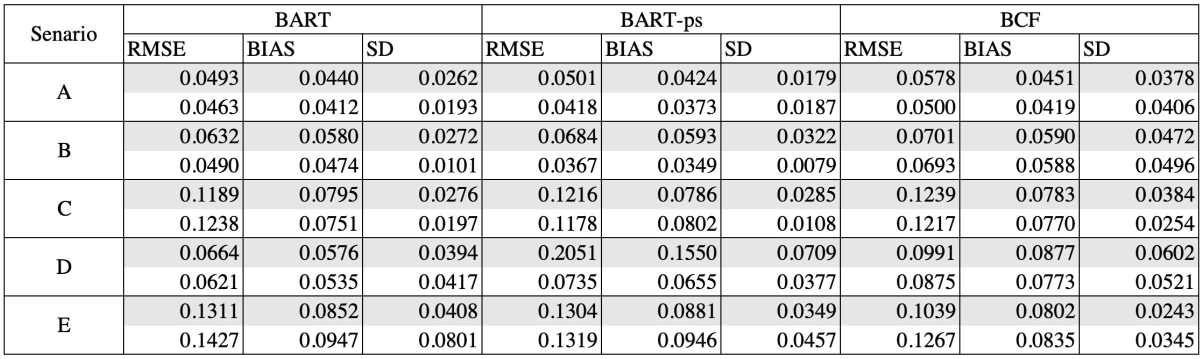
$$a(X) = \begin{cases} X_2 + X_5 - X_8, & \text{when} \quad p = 10 \\ X_5 + X_{10} - X_{22} + X_{25} + X_{31} - X_{40}, & \text{when} \quad p = 50 \end{cases}$$

For each of the 10 data generating processes, i.e., scenario A to J, we simulate 50 train sets, train the three Bayesian approaches on the 50 train sets. We evaluate the performance of the three approaches under cross-fitting and non-cross-fitting by averaging RMSE, absolute bias, and standard deviation of estimates over the 50 simulated train sets on the test set.

5. Result

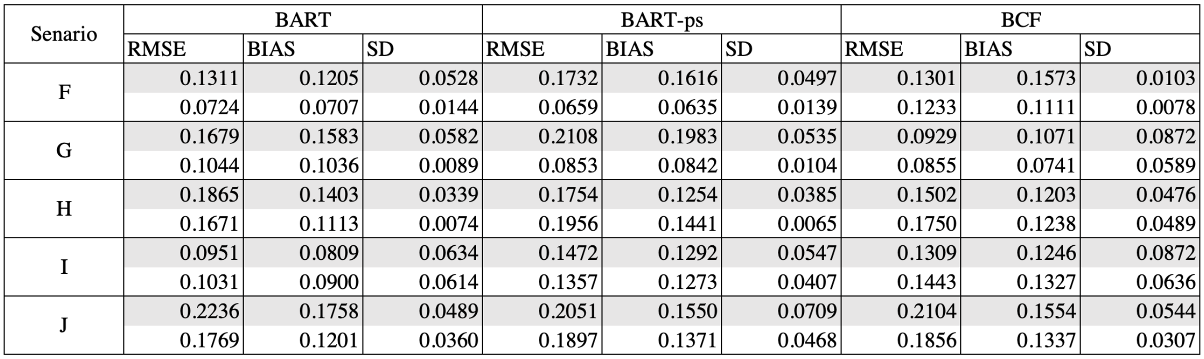
Our results for all 10 data generating processes are presented in figure 2 and figure 3, where detailed settings for A-J are presented in table 1.

All methods evaluated on data generating process E are improved by sample splitting in terms of RMSE, Bias and SD. Scenario E has the following features: a large sample size, low dimensions, heterogeneous treatment effect, sparsity, and the treatment assignment is linearly dependent on X. For other scenarios,



Senario	BART			BART-ps			BCF		
	RMSE	BIAS	SD	RMSE	BIAS	SD	RMSE	BIAS	SD
A	0.0493	0.0440	0.0262	0.0501	0.0424	0.0179	0.0578	0.0451	0.0378
	0.0463	0.0412	0.0193	0.0418	0.0373	0.0187	0.0500	0.0419	0.0406
B	0.0632	0.0580	0.0272	0.0684	0.0593	0.0322	0.0701	0.0590	0.0472
	0.0490	0.0474	0.0101	0.0367	0.0349	0.0079	0.0693	0.0588	0.0496
C	0.1189	0.0795	0.0276	0.1216	0.0786	0.0285	0.1239	0.0783	0.0384
	0.1238	0.0751	0.0197	0.1178	0.0802	0.0108	0.1217	0.0770	0.0254
D	0.0664	0.0576	0.0394	0.2051	0.1550	0.0709	0.0991	0.0877	0.0602
	0.0621	0.0535	0.0417	0.0735	0.0655	0.0377	0.0875	0.0773	0.0521
E	0.1311	0.0852	0.0408	0.1304	0.0881	0.0349	0.1039	0.0802	0.0243
	0.1427	0.0947	0.0801	0.1319	0.0946	0.0457	0.1267	0.0835	0.0345

Figure 2: Comparison between sample-splitting (cell in light grey) and full sample (cell in white), scenarios A-E



Senario	BART			BART-ps			BCF		
	RMSE	BIAS	SD	RMSE	BIAS	SD	RMSE	BIAS	SD
F	0.1311	0.1205	0.0528	0.1732	0.1616	0.0497	0.1301	0.1573	0.0103
	0.0724	0.0707	0.0144	0.0659	0.0635	0.0139	0.1233	0.1111	0.0078
G	0.1679	0.1583	0.0582	0.2108	0.1983	0.0535	0.0929	0.1071	0.0872
	0.1044	0.1036	0.0089	0.0853	0.0842	0.0104	0.0855	0.0741	0.0589
H	0.1865	0.1403	0.0339	0.1754	0.1254	0.0385	0.1502	0.1203	0.0476
	0.1671	0.1113	0.0074	0.1956	0.1441	0.0065	0.1750	0.1238	0.0489
I	0.0951	0.0809	0.0634	0.1472	0.1292	0.0547	0.1309	0.1246	0.0872
	0.1031	0.0900	0.0614	0.1357	0.1273	0.0407	0.1443	0.1327	0.0636
J	0.2236	0.1758	0.0489	0.2051	0.1550	0.0709	0.2104	0.1554	0.0544
	0.1769	0.1201	0.0360	0.1897	0.1371	0.0468	0.1856	0.1337	0.0307

Figure 3: Comparison between sample-splitting (cell in light grey) and full sample (cell in white), scenarios F-J

sample splitting slightly improved the performance in H for BART-ps and BCF, as well as in I for BART and BCF. In the remaining scenarios, the three metrics become larger when using sample splitting.

In summary, sample-splitting tends to worsen the results more for BART-ps and BCF, especially for the scenarios with small sample size. One reason to consider is that by splitting into three folds, each time we draw a subsample, we can only use less than one third of the data to train the propensity score model and the outcome model. The resulting two models are very likely to be wrongly specified and don't generalize well to the rest two thirds of the data. As a result, we might use the incorrectly estimated propensity scores to train the outcome model for predicting CATE, which means we're enlarging the effect of the bad propensity score model.

6. Conclusion and Discussion

We find that sample splitting almost does not improve accuracy in estimating CATE or reduce bias and variance of estimation using Bayesian approaches in all data generating processes we have considered. Especially in complicated data sets, for example, high dimensional with low sample size, and confounded treatment assignment, sample splitting can even worsen the results. Moreover, drawing subsets and cross-fitting is computationally expensive when we use Bayesian methods. Therefore, we recommend using the full sample to train the propensity score model as well as the outcome model to acheive higher accuracy and efficiency.

Two Bayesian estimators we used to estimate CATE are analogous of DR-estimators in the frequentist context, where the two-step procedure involves first estimating the propensity score and then using the estimated propensity score in the outcome model. The future direction of this study can be extended to evaluate the effect of cross-fitting on X-learners (Künzel 2019). First the response functions of the treated and control are estimated by any supervised learning algorithm (equivalent to a T-learner), then the treatment effects for units in the treated group are imputed based on control outcome estimators and vice versa; CATE is estimated by a weighted average of the two estimates from step 2. X-learners utilize information from the control group to derive better estimators for the treatment group and vice versa, and thus will especially outperform other algorithms when there exists substantial imbalance between the number of units in the treatment and control group. Another potential area of interest is to study whether sample splitting will help improve the accuracy when the outcome model (BART with propensity scores or BCF) is unsuitable for the data or the propensity score model is wrongly specified.

7. Reference

Bickel, P. J. (1982). On Adaptive Estimation. The Annals of Statistics, 10(3). https://doi.org/10.1214/aos/1176345863

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal, 21(1), C1–C68. https://doi.org/10.1111/ectj.12097

Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. Journal of Econometrics, 189(1), 1–23. https://doi.org/10.1016/j.jeconom.2015.06.017

Hahn, P. R., Carvalho, C. M., Puelz, D., & He, J. (2018). Regularization and Confounding in Linear Regression for Treatment Effect Estimation. Bayesian Analysis, 13(1). https://doi.org/10.1214/16-ba1044

Hahn, P. R., Murray, J. S., & Carvalho, C. M. (2020). Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects (with Discussion). Bayesian Analysis, 15(3). https://doi.org/10.1214/19-ba1195

Hill, J. L. (2011). Bayesian Nonparametric Modeling for Causal Inference. Journal of Computational and Graphical Statistics, 20(1), 217–240. https://doi.org/10.1198/jcgs.2010.08162

Edward H Kennedy. (2019). Optimal doubly robust estimation of heterogeneous causal effects. ArXiv preprint. https://arxiv.org/abs/2004.14497

Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. Proceedings of the National Academy of Sciences, 116(10), 4156–4165. https://doi.org/10.1073/pnas.1804597116

Schick, A. (1986). On Asymptotically Efficient Estimation in Semiparametric Models. The Annals of Statistics, 14(3). https://doi.org/10.1214/aos/1176350055

Zigler, C. M., Watts, K., Yeh, R. W., Wang, Y., Coull, B. A., & Dominici, F. (2013). Model Feedback in Bayesian Propensity Score Estimation. Biometrics, 69(1), 263–273. https://doi.org/10.1111/j.1541-0420.2012.01830.x