# Scalable Bootstrap Algorithms for Causal Inference with Large Real-World Data

Matthew Kosko [1] Lin Wang [1] Michele Santacatterina [2]

[1]The George Washington University [2]New York University

## Introduction

The bootstrap is an intuitive and powerful technique to quantify the uncertainty in treatment effect estimates. Its application, however, can be prohibitively demanding computationally in settings involving large datasets. In the past years, large datasets have become increasingly prevalent in medicine and epidemiology. For instance, during the COVID-19 pandemic, the effectiveness of mRNA COVID-19 vaccines was evaluated in datasets with more than 1 million subjects [1]. Alternatives to the standard bootstraps have been proposed. For instance, [5], introduced the Bag of Little Bootstraps (BLB), a robust and efficient way of quantifying the uncertainty in an estimate while having computationally superiority on large datasets. In this talk, we discuss the implementation of weighted-BLB, a modified version of the BLB aimed at quantifying the uncertainty of treatment effects estimates in large datasets. We evaluate the performance of the proposed technique in terms of bias, standard errors coverage of the true 95% confidence interval, and computational time in a simulation study. We apply the proposed technique in the evaluation of treatment effects in a large observational study containing more than 100.000 subjects. In the past few decades, various estimators of treatment effects have been proposed in addition to the standard, inverse probability weighting, and regression, such as, for instance, optimal weighting based on kernel methods [3]. The proposed methodology provides a unified way to quantify the uncertainty in treatment effect estimates regardless of the estimator used.

## Full sample bootstrap algorithms

The partition method we propose does not use any explicit model for the outcomes. We can, however, include an outcome model to impute potential outcomes and directly simulate the treatment randomization distribution

### Traditional

The standard approach is, for each bootstrap sample, we re-balance the data by re-computing weights and estimating the ATE [6] .

### Weighted

We can also use weighted bootstrap approaches that do not require re-balancing each bootstrap resample.

### CDF (full)

1 Construct IPW weights $w(X_i)$ in the full sample and construct weighted empirical cdfs for the potential outcomes, $\hat{F}_1(y_1) = \frac{\sum_{i=1}^n W_i w(X_i) \mathbf{1}(Y_i \leq y_1)}{\sum_{i=1} W_i w(X_i)}$ and $\hat{F}_0(y_0) = \frac{\sum_{i=1}^n (1-W_i) w(X_i) \mathbf{1}(Y_i \leq y_0)}{\sum_{i=1} (1-W_i) w(X_i)}$

3 Impute on potential outcomes in this sample using the isotone coupling [2]:

$$\tilde{Y}_i(0) = \begin{cases} Y_i & \text{if } W_i = 0 \\ \hat{F}_0(\hat{F}_1^{-1}(Y_i)) & \text{otherwise} \end{cases} \quad (1)$$

$$\tilde{Y}_i(1) = \begin{cases} Y_i & \text{if } W_i = 1 \\ \hat{F}_1(\hat{F}_0^{-1}(Y_i)) & \text{otherwise} \end{cases} \quad (2)$$

(3)

3 Partition the data and for each subset, for each bootstrap construct a new treatment assignment vector $\mathbf{W}^*$ and calculate the mean difference using the $\tilde{Y}$s.

### Linear model (full)

Rather than empirical cdfs, we can also use an outcome model for the $Y$s to impute. We impute the $Y$s using the correct linear model.

## Partition bootstrap algorithms

**Partition**

for $l \leftarrow 1$ to $s$ do
- Calculate $n_1 = \sum_{i=1} W_i$ and $n_0 = n - n_1$
- Sample a set of indices $I_l = \{i_1, \ldots, i_b\}$ from $\mathcal{I}$ without replacement
- Set $\mathcal{I} = \mathcal{I} / \cap_{j=1}^{l} I_j$
- For the data $\mathbf{Y}_l = (Y_{i_1}, \ldots, Y_{i_b})$, construct a model of the propensity score $\hat{\pi}(X_j)$
- Construct normalized inverse propensity weights for the subset data $(\hat{w}_0(X_i), \hat{w}_1(X_i)) = \left( \frac{1/(1-\hat{\pi}(X_i))}{\sum_{i=1}^b (1-W_i)/(1-\hat{\pi}(X_i))}, \frac{1/\hat{\pi}(X_i)}{\sum_{i=1}^b W_i/\hat{\pi}(X_i)} \right)$
- for $k \leftarrow 1$ to $r$ do
  - Let $b_1^{(k)} = \sum_{j=1}^b W_{i_j}$ and $b_0^{(k)} = b - b_1^{(k)}$
  - Sample $\left( M_1^{1,k}, \ldots, M_{b_1^{(k)}}^{1,k} \right) \sim \text{Multinomial}(n_1, \hat{\mathbf{w}}_1)$
  - Sample $\left( M_1^{0,k}, \ldots, M_{b_0^{(k)}}^{0,k} \right) \sim \text{Multinomial}(n_0, \hat{\mathbf{w}}_0)$
  - Calculate $\hat{\tau}_k^{(j)*} = \frac{1}{n_1} \sum_{i=1}^b W_i Y_i M_i^{1,k} - \frac{1}{n_0} \sum_{i=1}^b (1 - W_i) Y_i M_i^{0,(k)}$
- end for
- $\hat{\tau}^{(l)} \leftarrow \frac{1}{r} \sum_{i=1}^r \hat{\tau}_i^{(l)}$

end for

$\hat{\tau} \leftarrow \frac{1}{s} \sum_{i=1}^s \hat{\tau}^{(i)*}$

This algorithm first partitions the full dataset into $s$ distinct subsets, estimates a propensity score model within each subset, resamples within each subset using the (normalized) IPWs as weights and estimates an ATE for each resample, then summarizes the resample estimates with a chosen statistic and averages those statistics across partitions to obtain an overall bootstrap estimate.

### Additional partition

We also apply the partition to the linear model and weighted empirical cdfs methods described previously. The difference is that imputation is done within each subset.

### Consistency of estimator $\hat{\tau}_k^{(j)*} \xrightarrow{p} \hat{\tau}_{\mathbf{ATE}}$

For the $q$th bootstrap draw, we construct the difference in mean estimator:

$$\hat{\tau}_q^{(j)*} = \frac{1}{n_1} \sum_{i=1}^b W_i Y_i M_i^{1,q} - \frac{1}{n_0} \sum_{i=1}^b (1 - W_i) Y_i M_i^{0,q}$$

Consider a single term of the above summation. By the weak law of large numbers, over $r$ weighted bootstrap resamples (corresponding to $r$ treatment and control multinomial draws), for each $i$

$$\frac{1}{r} \sum_{j=1}^r \left( \frac{W_i Y_i M_i^{1,j}}{n_1} - \frac{(1 - W_i) Y_i M_i^{0,j}}{n_0} \right) \xrightarrow{p} W_i Y_i w_1(X_i) - (1 - W_i) Y_i w_0(X_i)$$

Because the subsets are drawn at random from the full dataset, this implies that $\frac{1}{r} \sum_{i=1}^r \hat{\tau}_i^{(j)*} \xrightarrow{p} \hat{\tau}^{(j)*}$ and $\hat{\tau}^* \xrightarrow{p} \hat{\tau}_{ATE}$

## Simulation data

The simulation uses a data-generating process (DGP) derived from [4]. We generate 100,000 responses from the following DGP.

$$\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{i4}) \sim \text{N}(\mathbf{0}, \mathbf{I}_4)$$
$$\epsilon_i \sim \text{N}(0, 1), \quad i = 1, \ldots, n$$
$$\Pr(W_i = 1 | \mathbf{Z}_i) \equiv e(\mathbf{Z}_i) = \frac{1}{1 + \exp(Z_1 - 0.5Z_2 + 0.25Z_3 + 0.1Z_4)}$$
$$Y_i(0) = 210 + 27.4Z_1 + 13.7Z_2, +13.7Z_3 + 13.7 * Z_4 + \epsilon_i$$
$$Y_i(1) = Y_i(0) + 2\mathbf{1}(\text{Tr} = 1)$$

## Simulation results

To find the true standard error, we generate 1,000 data replicates of the using the correct propensity score specification and calculate $\hat{\tau}_{\text{ATE}} = \sum_{i=1}^n \left( W_i Y_i \frac{\frac{1}{\hat{\pi}(\mathbf{Z}_i)}}{\sum \frac{W_i}{\hat{\pi}(\mathbf{Z}_i)}} - (1 - W_i) Y_i \frac{\frac{1}{1-\hat{\pi}(\mathbf{Z}_i)}}{\sum \frac{1-W_i}{1-\hat{\pi}(\mathbf{Z}_i)}} \right)$

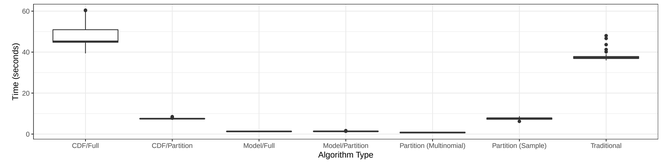


Figure 1. Time elapsed for each algorithm type

| Esimate | Std. Err. | Algorithm |
|---|---|---|
| 1.98 | 0.14 | Traditional |
| 2.01 | 0.22 | CDF/Full |
| 1.98 | 0.19 | Model/Full |
| 1.98 | 0.21 | Partition (Multinomial) |
| 1.95 | 0.19 | Partition (Sample) |
| 1.90 | 0.57 | CDF/Partition |
| 2.08 | 0.31 | Model/Partition |

Table 1. Estimates and standard error for each algorithm type

## References

[1] Noa Dagan, Noam Barda, Eldad Kepten, Oren Miron, Shay Perchik, Mark A Katz, Miguel A Hernán, Marc Lipsitch, Ben Reis, and Ran D Balicer. Bnt162b2 mrna covid-19 vaccine in a nationwide mass vaccination setting. *New England Journal of Medicine*, 2021.

[2] Guido Imbens and Konrad Menzel. A causal bootstrap. *The annals of statistics*, 49(3):1460–1488, 2021.

[3] Nathan Kallus, Brenton Pennicooke, and Michele Santacatterina. More robust estimation of average treatment effects using kernel optimal matching in an observational study of spine surgical interventions. *Statistics in medicine*, 40(10):2305–2320, 2021.

[4] Joseph DY Kang and Joseph L Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 22(4):523–539, 2007.

[5] Ariel Kleiner, Ameet Talwalkar, Purnamrita Sarkar, and Michael I Jordan. A scalable bootstrap for massive data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):795–816, 2014.

[6] Yumin Zhang and Arman Sabbaghi. The designed bootstrap for causal inference in big observational data. *Journal of Statistical Theory and Practice*, 15(4):1–26, 2021.