REGRESSION ADJUSTMENT WITH SYNTHETIC CONTROLS IN ONLINE EXPERIMENTS

Congshan Zhang[1], Dominic Coey[1], Matt Goldman[1], and Brian Karrer[1]

[1]Facebook
[1]*cszhang, coey, mattgoldman, briankarrer@fb.com*

October 19, 2021

ABSTRACT

In the setting of online experiments, we propose a two-step procedure to improve efficiency for estimating average treatment effect (ATE) by combining synthetic control methods with the popular regression adjustment framework. In particular, we form a synthetic control for each and every subject in the experiment using a donor pool that consists of k nearest-neighbors (kNN) from outside of the experiment. The predicted counterfactuals are then used in the following regression adjustment step. The asymptotic theory of the method can be shown following [17] and is validated in a realistically calibrated Monte Carlo analysis. For both user-level and cluster experiments at Facebook, we show that the proposed method yields significantly narrower CIs compared with the standard difference-in-mean estimator and a widely used OLS adjusted estimator.

1 Introduction

Regression adjustment methods have been widely used to improve estimation efficiency in randomized experiments of various designs [12, 13, 16, 17, 18]. It is known that asymptotically the estimator after adjustment performs no worse than the standard difference-in-mean estimator [19]; in fact, the stronger the covariates correlate with the outcome, the greater variance reduction the adjusted estimator would achieve [17]. Meanwhile, machine learning (ML) has received growing attention in causal inference literature (e.g., [8, 9, 10, 20]). For variance reduction purpose in the setting of online experiments, in particular, [14] uses ML models to predict the value of outome from many covariates in the hope of generating a proxy that is strongly correlated with the goal metric, in which case the width of confidence interval is shown to be reduced dramatically.

Regression adjustment is known to be valid under mild assumptions that are standard for analyzing online experiments and is very easy to implement. In this article, we propose a new idea of improving estimation efficiency under the framework of regression adjustment that makes use of synthetic control methods [2]. In particular, we construct a synthetic control for each and every subject in the experiment using a donor pool that consists of k nearest-neighbors (kNN) *outside* the experiment. The predicted counterfactuals from synthetic controls are later used as a covariate in the following regression adjustment step (e.g., OLS regression). We emphasize from the outset that individuals outside the experiment could be involved in other experiments and thus receive other interventions, so while they can be useful as control variates, they are not actually part of the control group. In fact, we could consider the donor pool as being literally any time-series of any quantities independent of the intervention – they do not even need to be meaningful subjects such as users.

The synthetic control method is originally proposed in economics to estimate the effects of interventions on aggregate outcomes (e.g., at country level) in observational settings [5]. More recently, the method has also been used for disaggregated data (e.g., [6]). To the best of our knowledge, this is the first study that applies synthetic control methods for the purpose of variance reduction in randomized experiments. Our method distinguishes itself from existing regression adjustment methods in two important ways. First, it exploits both the time-series fluctuation of the outcome variable in the pre-intervention periods and the information of donors outside the experiment, which most existing methods do not fully leverage. Second, it offers great flexibility for customization, and thus should admit broad applications – it can be applied in both individual-level and cluster experiments; the predicted counterfactuals

can be used either alone as the adjustment or as an additional covariate in more complicated nonlinear adjustment models. Our method can be particularly useful in cluster experiments which do not have a large number of clusters for training complex ML models.

2 Methodology

We introduce the synthetic control average treatment effect (SCATE) estimator. For a real vector of matching variables $\boldsymbol{X}$, we define $\|\boldsymbol{X}\| = (\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})^{1/2}$, where $\boldsymbol{V}$ is a diagonal matrix to adjust the scale and relative importance of different components in $\boldsymbol{X}$. As a concrete example, if we only include the time-series values of the outcome variable in pre-intervention periods in $\boldsymbol{X}$, then $\boldsymbol{X}$ is a T_0-vector, where T_0 is the number periods before the experiment starts. If we let $\boldsymbol{V} = I$, $\|\boldsymbol{X}\|$ becomes the Euclidean norm. Let the post-intervention outcome variable be Y. Given an integer k and a diagonal matrix $\boldsymbol{V}$ – the choice of which will be discussed below – we propose the following algorithm.

Algorithm. SCATE

Procedure SCATE(k, $\boldsymbol{V}$)
  for each subject in the experiment ($i = 1, \ldots, n$) **do**
    i) Find k nearest-neighbors indexed by j from $N - n$ subjects outside the experiment based on $\|\boldsymbol{X}_i - \boldsymbol{X}_j\|^2$.
    ii) Compute $\boldsymbol{W}_i^\star \equiv (W_{i,1}, \ldots, W_{i,k})$ that solves

$$\min_{\boldsymbol{W}_i \in \mathbb{R}^k} \left\| \boldsymbol{X}_i - \sum_{j=1}^k W_{i,j} \boldsymbol{X}_j \right\|^2$$

$$\text{s.t.} \quad W_{i,j} \geq 0,$$

$$\sum_{j=1}^k W_{i,j} = 1,$$

    where $W_{i,j}$ is the weight given to donor j to form the synthetic control for subject i.
    iii) Compute predicted counterfactuals as $\hat{Y}_i = \sum_{j=1}^k W_{i,j}^\star Y_j$.
  end for
  Estimation. Get $\hat{\tau}$ from the OLS regression

$$Y_i = \alpha + \tau T_i + \gamma \hat{Y}_i + \phi T_i \left(\hat{Y}_i - \bar{\hat{Y}} \right) + \epsilon_i$$

  as the ATE estimator, where T_i is the treatment indicator.
  CLT. Compute $\hat{\sigma}_\tau^2 = \hat{\sigma}_0^2/(1-\hat{p}) + \hat{\sigma}_1^2/\hat{p} - \hat{\sigma}_{sc}^2[\hat{\gamma}\hat{p} + (\hat{\gamma}+\hat{\phi})(1-\hat{p})]^2/\hat{p}(1-\hat{p})$, where $\hat{\sigma}_1^2$ and $\hat{\sigma}_0^2$ are the sample variances of Y_i in the treatment and control group, respectively; $\hat{\sigma}_{sc}^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{\hat{Y}})^2/n$; and $\hat{p} = \sum_{i=1}^n T_i/n$ (the derivation of $\hat{\sigma}_\tau$ is based on (10) in [21]).
  Use $\sqrt{n}(\hat{\tau} - \tau)/\hat{\sigma}_\tau \to \mathcal{N}(0,1)$ for statistical inference.
end procedure

Comments. Several practical considerations with using synthetic controls are discussed below.

(a) **Lack of unique solutions.** Given our large donor pool outside the experiments, one could possibly find multiple best synthetic controls for some subject if the matching variables of that subject fall into a convex hull of the corresponding values for the donors. Two common solutions to this problem are (i) restricting the donor pool to a smaller number of units that are similar to the target unit until the nonuniqueness issue is no longer a practical concern as in the proposed algorithm and (ii) introducing a penalized term that focuses on pair-wise discrepency between the target unit and each donor as advocated in [6]. The two solutions share the same spirit, but the latter one is more computationally intensive due to the high dimensionality of the problem, especially in the setting of online experiments.

(b) **Choice of k.** Since the synthetic control method finds the projection of each subject onto a convex hull spanned by its k neighbors, adding more neighbors would reduce matching error but at the same time is susceptible to overfitting (i.e., the discussion on the tradeoff between extrapolation bias and interpolation bias in the synthetic control community, e.g., [15]). A practical way of choosing k is to minimize the sum of squared errors $\sum_{i \in \text{control group}} (Y_i - \hat{Y}_i(k))^2$ in the control group.

(c) **Choice of $\boldsymbol{V}$.** A simple choice is to use the inverse variance of each dimension of $\boldsymbol{X}$ to form a diagonal matrix. A more sophisticated method involves using a nested quadratic optimization with out-of-sample validation to internalize the choice of $\boldsymbol{V}$ in the algorithm (e.g., [4]). Based on our empirical analysis with real experiment data, when only values of the outcome variable over pre-intervention periods are included in $\boldsymbol{X}$, setting $\boldsymbol{V} = I$ rarely affects the variance reduction performance, and thus is a reasonable choice for large scale experiments.

(d) **Computational feasibility.** In the above algorithm, we use the synthetic control method proposed in the seminal paper [3] to get the predicted counterfactuals (later referred to as the ADH method). For this method, both finding kNN and optimizing for weights can be computationally intensive at scale. There exist highly scalable solutions for similarity search (e.g., FAISS [1]), which makes it possible to find kNN among millions of units within seconds. In addition, finding optimal weights can be converted into a matrix completion problem, which can be solved via iterative

thresholded SVD [7]. We remark that other methods for counterfactual prediction could also be used to replace the ADH method. To name two alternatives, (a) Vertical Regression: Given the kNNs, we could simply run time-series regressions for pre-intervention values of the outcome [11], that is, $X_{i,t} = \beta_0 + \beta_1 X_{\text{neighbor}_1,t} + \cdots + \beta_k X_{\text{neighbor}_k,t} + \epsilon_t$ for pre-intervention periods $t = 1, \ldots, T_0$, with $k < T_0$, and the estimated regression model can then be used in the post-treatment periods to generate counterfactual predictions according to $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 Y_{\text{neighbor}_1} + \cdots + \hat{\beta}_k Y_{\text{neighbor}_k}$; (b) Elastic Net: We could apply the elastic net synthetic control method studied in [11], which readily allows for the possibility that $N - n \gg T_0$ by solving $\min_{\boldsymbol{W}_i \in \mathbb{R}^{N-n}} \|\boldsymbol{X}_i - \sum_{j=1}^{N-n} W_{i,j} \boldsymbol{X}_j\|^2 + \lambda(\alpha \|\boldsymbol{W}_i\|_1 + \frac{1-\alpha}{2} \|\boldsymbol{W}_i\|^2)$.

Given our purpose of variance reduction in online experiments, we propose two simple extensions of the above algorithm that might better fit certain scenarios in practice. First, since synthetic controls are derived from matching subjects in and outside the experiment, they contain "orthogonal" information to the covariates that are usually used in non-linear regression adjustment methods. It is therefore natural to use the predicted counterfactuals as additional covariates in those non-linear models. For example, under the framework of [14], we can add the predicted counterfactual as an additional feature in the GBDT model. Hypothetically, one could also throw all time-series data of the focal subjects and donors into a large ML model to generate counterfactual predictions. However, due to the considerable computational needs and the implementation complexity of a large scale ML model, we believe the synthetic control approach has distinct value to extract useful information from the raw data. Second, when cluster experiments are considered, the synthetic control method could be applied to the aggregated data at cluster level. When the number of clusters is not large enough for training highly complex ML models with many covariates, SCATE clearly has its advantage and could be used instead to gain efficiency.

From a theoretical standpoint, synthetic controls and regression adjustment methods have very good synergy – the predicted counterfactuals are independent of the treatment assignment, since they only rely on the values of the outcome variable in the pre-intervention periods, the outcome values from donors outside the experiment, and potentially some time-invariant covariates. Under the assumption of no interference across the subjects, the argument in [17] can be directly applied to justify the asymptotic validity of the algorithm. In cluster experiments where clusters are treated as subjects, the inference theory holds under the assumption of no cross-cluster interference.

3 Numerical study

3.1 Consistency

As discussed above, SCATE always generates consistent estimates of the average treatment effects. To verify this property, we examine two metrics used in Facebook cluster experiments. The metric values are pre-aggregated into the cluster level. We have in total 10,000 clusters, from which we randomly select 100–1,000 clusters into the experiment with the remaining clusters outside to mimic the size of the real experiments. We simulate both A/A and A/B tests with constant treatment effects, and compare the average finite-sample bias of SCATE with that of the standard difference-in-mean estimator across 1,000 Monte Carlo trials. For the A/B settings, we consider two effect sizes: 1% and 5%. As expected, our proposed method almost always yields smaller biases compared with the difference-in-mean estimator in finite sample.

Table 1. Finite-sample bias in percentage term (%)

Model	Diff-in-mean	SCATE
A. Topline Metric I.		
A/A Test	0.0072	0.0025
A/B Test: 1% effect	0.0110	0.0068
A/B Test: 5% effect	(0.0207)	(0.0083)
B. Topline Metric II.		
A/A Test	(0.0001)	(0.0045)
A/B Test: 1% effect	0.0128	0.0055
A/B Test: 5% effect	(0.0039)	(0.0002)

3.2 Variance reduction performance

Finally, we compare the empirical performance of SCATE against two popular estimators – difference-in-mean and OLS adjusted estimator – in simulated A/A tests for a set of real outcome metrics. As in the previous section, we

simulate 1,000 Monte Carlo repetitions for each metric and calculate average improvement of efficiency. [1] Below, we present some highlights of our results for both user-level and cluster experiments.

For user-level experiments, SCATE yields significantly narrower confidence intervals compared with both the standard difference-in-mean estimator and a commonly used estimator adjusted by the pre-intervention outcome values with OLS. In particular, we show that for four top company critical metrics, SCATE yields about 58% narrower CIs compared with the standard difference-in-mean estimator and 16% narrower CIs compared with the OLS adjusted estimator.

For cluster experiments, which are also widely used at Facebook, on average for tests with 100–1,000 clusters, the CIs from SCATE are approximately 34% narrower relative to those generated by the standard difference-in-mean estimator. When compared against the OLS adjusted estimator, SCATE produces 15% narrower CIs. Details can be found in Table 2.

Table 2. Ratio of the width of CI

Metric	SCATE / Diff-in-Mean	SCATE / OLS-Ajusted
A. User-level experiments		
Metric 1	0.5656	0.8879
Metric 2	0.4233	0.8551
Metric 3	0.5840	0.8809
Metric 4	0.2456	0.7201
B. Cluster experiments		
Metric 1	0.9070	0.9641
Metric 2	0.3491	0.6959
Metric 3	0.4531	0.8614
Metric 4	0.4919	0.7533
Metric 5	0.9326	0.9762
Metric 6	0.9182	0.9606
Metric 7	0.9918	0.9966
Metric 8	1.0074	0.9991
Metric 9	0.5932	0.7849
Metric 10	0.9968	0.9950
Metric 11	0.7874	0.8706
Metric 12	0.6839	0.8739
Metric 13	0.6189	0.9931
Metric 14	0.4653	0.6811
Metric 15	0.5371	0.7605
Metric 16	0.4000	0.6340
Metric 17	0.8836	0.9762

References

[1] Facebook AI similarity search. `https://ai.facebook.com/tools/faiss/`.

[2] Alberto Abadie. Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, 59(2):391–425, 2021.

[3] Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American statistical Association*, 105(490):493–505, 2010.

[4] Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Comparative politics and the synthetic control method. *American Journal of Political Science*, 59(2):495–510, 2015.

[5] Alberto Abadie and Javier Gardeazabal. The economic costs of conflict: A case study of the basque country. *American economic review*, 93(1):113–132, 2003.

[6] Alberto Abadie and Jérémy L'Hour. A penalized synthetic control estimator for disaggregated data. *Working paper*, 2021.

[7] Susan Athey, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, pages 1–15, 2021.

[8] Susan Athey, Guido Imbens, and Stefan Wager. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society Series B*, 80(4):597–623, 2018.

[1] We compared the variance reduction performance between the SCATE method and our extension idea of replacing the optimization step with a time-series regression. Given their comparable empirical performance, we only present results from the regression for speedy implementation.

[9] Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *Annals of Statistics*, 47(2):1148–1178, 2019.

[10] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. 2018.

[11] Nikolay Doudchenko and Guido W Imbens. Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. Technical report, National Bureau of Economic Research, 2016.

[12] Colin B Fogarty. Regression-assisted inference for the average treatment effect in paired experiments. *Biometrika*, 105(4):994–1000, 2018.

[13] Kevin Guo and Guillaume Basse. The generalized oaxaca-blinder estimator. *Journal of the American Statistical Association*, (just-accepted):1–35, 2021.

[14] Yongyi Guo, Dominic Coey, Mikael Konutgan, Wenting Li, Chris Schoener, and Matt Goldman. Machine learning for variance reduction in online experiments. *arXiv preprint arXiv:2106.07263*, 2021.

[15] Maxwell Kellogg, Magne Mogstad, Guillaume Pouliot, and Alexander Torgovitsky. Combining matching and synthetic control to trade off biases from extrapolation and interpolation. Technical report, National Bureau of Economic Research, 2020.

[16] Xinran Li and Peng Ding. Rerandomization and regression adjustment. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):241–268, 2020.

[17] Winston Lin. Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. *Annals of Applied Statistics*, 7(1):295–318, 2013.

[18] Hanzhong Liu and Yuehan Yang. Regression-adjusted average treatment effect estimates in stratified randomized experiments. *Biometrika*, 107(4):935–948, 2020.

[19] Jerzy Splawa-Neyman, Dorota M Dabrowska, and TP Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pages 465–472, 1990.

[20] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

[21] Li Yang and Anastasios A Tsiatis. Efficiency study of estimators for a treatment effect in a pretest–posttest trial. *The American Statistician*, 55(4):314–321, 2001.