Title: New Perspectives on the Synthetic Control Method

Authors:

Eli Ben-Michael, UC Berkeley, ebenmichael@berkeley.edu

Avi Feller, UC Berkeley, afeller@berkeley.edu
[presenting author]

Jesse Rothstein, UC Berkeley, rothstein@berkeley.edu

Problem:

Many important interventions and policy changes in education occur at an aggregate level, such as the level of the school, district, or state; prominent examples include school finance policies, curriculum development, and accountability changes. In settings where randomization is infeasible or unethical, education researchers often turn to quasi-experimental research designs based on repeated observations of aggregate data. For example, a researcher might estimate the impact of a new reading program using school-level average test scores at multiple time points surrounding the introduction of the intervention (Jacob, Somers, Zhu, & Bloom, 2016).

The synthetic control method (SCM) is an increasingly popular approach for estimating impacts in this setting (Abadie, Diamond, & Hainmueller, 2010). The idea is to construct a comparison case, known as a synthetic control, which minimizes the imbalance of pre-treatment outcomes between the treated unit and synthetic control. The estimated impact is then the difference in post-treatment outcomes between the observed treated unit and the synthetic control. SCM has been widely applied — the main SCM papers have over 4,000 citations — and has been called "[a]rguably the most important innovation in the policy evaluation literature in the last 15 years" (Athey & Imbens, 2017).

Despite its popularity, however, SCM remains poorly understood, with little guidance for practice. How should researchers implement this approach to yield valid estimates of causal effects and in what settings? How should they make specific implementation choices? When are the results credible? Taken together, this lack of guidance meaningfully hampers education research on important policy questions, since researchers have limited ability to differentiate studies that provide valuable evidence from weaker studies particularly vulnerable to confounding contextual factors.

Prior methodological research:

There is an extensive methodological literature on SCM (Abadie & Cattaneo, 2018). We briefly highlight three relevant threads:

- **SCM estimation.** Kreif, Gruber, Radice, Grieve, & Sekhon (2016), Gobillon & Magnac (2016), and Ferman & Pinto (2018), among others, assess the general performance of SCM methods. Ferman & Pinto (2018), in particular, focus on the important setting in which the SCM fit is imperfect.
- **SCM testing and inference.** Hahn & Shi (2017) and Firpo & Possebom (2018) assess testing and inference in practice, finding that the standard approach for uniform placebo testing
- **SCM extensions.** There have been many important extensions and generalizations to SCM. Doudchenko & Imbens (2016) relax important restrictions for SCM, including the constraint that the weights are non-negative. Xu (2017) estimate the linear factor model directly.

Finally, we build on the recent literature on balancing weights, also known as calibrated propensity scores. Examples of this approach include Hainmueller (2012), Zubizarreta (2015), and \Athey, Imbens, & Wager (2018).

Method:

Our paper uses analogies to balancing estimators to demonstrate SCM's properties. We first show that SCM is an *approximate balancing weight estimator*. Next, we show that it is an inverse propensity score weighting (IPW) estimator. Each type of estimator has received a great deal of attention in the recent literature, and we can use these equivalencies to clarify ambiguity about bias and inference for SCM.

On the basis of the first equivalence, we conclude that SCM is biased in the settings in which it is typically applied. Typical SCM applications involve a small number of treated units, as few as one, and a relatively large number of pre-treatment observations. Exact balance between the treated unit and the weighted synthetic control is only feasible if the treated unit's pre-treatment time series is contained within the convex hull formed by the control units. The curse of dimensionality makes this unlikely when the pre-treatment time series is long. Existing theoretical results for SCM, including the result that it is asymptotically unbiased, assume exact balance. Because this is unlikely, we argue that SCM will generally be biased, analogous to bias for matching estimators with inexact matches. We bound this bias and argue that analysts implementing SCM should pursue dimension reduction strategies to make achieving balance more likely.

We then demonstrate that SCM is an IPW estimator, using pre-treatment outcomes as covariates and (when exact balance is not possible) penalizing the propensity score coefficients via a ridge penalty. We use this equivalence to demonstrate that existing methods for testing and inference, which assume uniform probabilities of selection, yield invalid tests and inferences. In principle, valid tests could be formed by incorporating the propensity score weights into existing inference and testing procedures. In practice, however, this is quite challenging, as the propensity score is poorly estimated and typically only a handful of units have positive estimated propensities. We offer some possible steps forward.

Lastly, we propose an Augmented SCM estimator, analogous to Augmented IPW, that combines SCM's approximate balancing with an outcome model. We give some theoretical guarantees for special cases and outline specific implementation decisions. While attractive, the augmented estimator also comes at a price: the resulting weights on donor units can be negative, which limits interpretability.

Setting:

The setting where SCM is applicable is common in education policy research. We focus on two examples. First, we assess the Reading First (RF) program analyzed in Jacob et al. (2016). In this within-study comparison, the sample for this design includes 69 Reading First schools that were

matched to similar comparison elementary schools in the same US state using propensity score matching. The time series cross-sectional dataset constructed for this purpose includes information about schools' 3rd grade state reading and math test scores, school characteristics (enrollment, student demographic, etc.), and local poverty rates from the Census Bureau's Small Area and Income and Poverty Estimates.

Second, we assess the introduction of universal pre-kindergarten, using the launch of programs in Georgia and Oklahoma in the 1990s on children's NAEP scores, as analyzed in Cascio & Schanzenbach (2013). Going beyond the original difference-in-differences analysis, we can explore the implicit control groups in these estimates using both NAEP and SEDA data, which will also allow us to more thoroughly examine performance by student subgroup (e.g., free lunch recipients).

References

Abadie, A., & Cattaneo, M. D. (2018). Econometric methods for program evaluation. *Annual Review of Economics*.

Abadie, A., Diamond, A., & Hainmueller, J. (2010). 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.

Athey, S., & Imbens, G. W. (2017). The State of Applied Econometrics: Causality and Policy Evaluation. *Journal of Economic Perspectives*, *31*(2), 3–32.

Athey, S., Imbens, G. W., & Wager, S. (2018). Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *74*, 235–28.

Cascio, E. U., & Schanzenbach, D. W. (2013). The Impacts of Expanding Access to High-Quality Preschool Education. *Brookings Papers on Economic Activity*, 127–192.

Doudchenko, N., & Imbens, G. W. (2016). Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis.

Ferman, B., & Pinto, C. (2018). Synthetic Controls with Imperfect Pre-Treatment Fit, 1–45.

Firpo, S., & Possebom, V. (2018). Synthetic control method: Inference, sensitivity analysis and confidence sets. *Journal of Causal Inference*.

Gobillon, L., & Magnac, T. (2016). Regional Policy Evaluation: Interactive Fixed Effects and Synthetic Controls. *Review of Economics and Statistics*, *98*(3), 535–551.

Hahn, J., & Shi, R. (2017). Synthetic Control and Inference. *Econometrics*, *5*(4), 52.

Hainmueller, J. (2012). Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies. *Political Analysis*, *20*(1), 25–46.

Jacob, R., Somers, M.-A., Zhu, P., & Bloom, H. (2016). The Validity of the Comparative Interrupted Time Series Design for Evaluating the Effect of School-Level Interventions. *Evaluation Review*, *40*(3), 167–198.

Kreif, N., Gruber, S., Radice, R., Grieve, R., & Sekhon, J. S. (2016). Evaluating treatment effectiveness under model misspecification: A comparison of targeted maximum likelihood estimation with bias-corrected matching. *Statistical Methods in Medical Research*, *25*(5), 2315–2336.

Xu, Y. (2017). Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis*.

Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, *110*(511), 910–922.