A Design-Based Perspective on Synthetic Control Methods

Lea Bottmer[1] Guido Imbens[1,2,3]
Jann Spiess[2,3] Merrill Warnick[1]

[1]Department of Economics
[2]Graduate School of Business
[3]SIEPR

Stanford University

NBER Summer Institute 2021

July 29, 2021

Our paper

- Synthetic control (SC) estimator typically applied in observational data and justified using factor models
- Instead analyze SC-type estimators under randomization, propose modifications, compare to difference in means (DiM)
- Main results:
    1. Standard SC generically biased under randomization
    2. Bias can be fixed by adding a constraint on weights
    3. DiM variance estimator can be extended to SC-type estimators
- Main take-aways:
    1. For experiment: modified unbiased SC that improves over DiM
    2. For observational data: additional robustness, insights into inference that implicitly assumes randomization

Motivating example

- Setup: Three units $i \in \{\text{AZ}, \text{CA}, \text{NY}\}$ observed over two periods $t \in \{1, 2\}$, with one unit treated $(U_i = 1)$ at $t = 2$ and two pure controls $(U_i = 0)$

- Goal: estimate treatment effect of treated unit in period 2,

$$\tau = \sum_i U_i \big(\overbrace{Y_{i2}(1)}^{= Y_{i2}, \text{ observed}} - \underbrace{Y_{i2}(0)}_{\text{not observed}} \big)$$

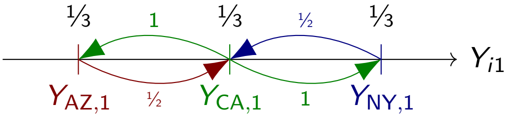
- Synthetic control: learn (positive) weights M_{ij} for $j \neq i$ from previous outcomes Y_{i1} to form

$$\hat{Y}_{i2}(0) = \sum_{j \neq i} M_{ij} Y_{j2}, \qquad \hat{\tau} = \sum_i U_i \Big(Y_{i2} - \hat{Y}_{i2}(0) \Big)$$

by minimizing error $(Y_{i1} - \sum_{j \neq i} M_{ij} Y_{j1})^2$

Motivating example

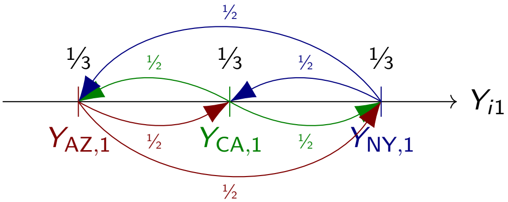
Motivating example



M_{ij}	$j =$ AZ	CA	NY
$i =$ AZ	–	1	0
CA	1/2	–	1/2
NY	0	1	–

$$\mathbb{E}\left[\hat{\tau} - \tau\right] = \frac{1}{3} \sum_i \hat{Y}_{i2}(0) - Y_{i2}(0) = \frac{1}{3} \sum_i \Big(\sum_{j \neq i} M_{ji} - 1 \Big) Y_{i2}(0)$$

Motivating example



M_{ij}	$j =$ AZ	CA	NY
$i =$ AZ	–	1	0
CA	1/2	–	1/2
NY	0	1	–

M_{ij}	$j =$ AZ	CA	NY
$i =$ AZ	–	1/2	1/2
CA	1/2	–	1/2
NY	1/2	1/2	–

$$\mathbb{E}\left[\hat{\tau} - \tau\right] = \frac{1}{3}\sum_i \hat{Y}_{i2}(0) - Y_{i2}(0) = \frac{1}{3}\sum_i \left(\sum_{j \neq i} M_{ji} - 1\right) Y_{i2}(0)$$

Plan for the talk

1. Generalized Synthetic Control (GSC) estimators
2. Properties of GSC estimators under randomization
3. Unbiased and modified unbiased SC estimators
4. Variance estimation
5. Extensions
6. Take-aways for applying GSC estimators

Related literature

- Synthetic Control methodology [Abadie and Gardeazabal, 2003, Abadie et al., 2010, Abadie et al., 2015, Abadie, 2019]
- New estimators in a general SC class [Doudchenko and Imbens, 2016, Abadie and L'Hour, 2017, Ferman and Pinto, 2017, Arkhangelsky et al., 2019, Li, 2020, Ben-Michael et al., 2020]
- Inference for SC estimators [Abadie et al., 2010, Doudchenko and Imbens, 2016, Ferman and Pinto, 2017, Hahn and Shi, 2016, Lei and Candès, 2020, Chernozhukov et al., 2017]
- Randomization inference for causal effects [Neyman, 1990, Imbens and Rubin, 2015, Abadie et al., 2020, Rambachan and Roth, 2020, Sekhon and Shem-Tov, 2020]

Plan for the talk

1. Generalized Synthetic Control (GSC) estimators
2. Properties of GSC estimators under randomization
3. Unbiased and modified unbiased SC estimators
4. Variance estimation
5. Extensions
6. Take-aways for applying GSC estimators

Panel data setup

- Outcomes Y_{it}:

$$\mathbf{Y} = \begin{pmatrix} Y_{11} & Y_{12} & Y_{13} & \dots & Y_{1T} \\ Y_{21} & Y_{22} & Y_{23} & \dots & Y_{2T} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_{N1} & Y_{N2} & Y_{N3} & \dots & Y_{NT} \end{pmatrix}$$

- Binary treatment-unit and treatment-time vectors **U** and **V**:

$$\mathbf{U} = \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_N \end{pmatrix}, \mathbf{V} = \begin{pmatrix} V_1 \\ V_2 \\ \vdots \\ V_T \end{pmatrix}$$

$U_i = 1$ and $V_t = 1$ correspond to i being treated in t

In this talk: assume treatment in last period, $V_t = 1_{t=T}$

Potential-outcomes setup

- Potential outcomes: $Y_{it}(0)$, $Y_{it}(1)$
- Observed outcome: $Y_{it} = (1 - U_i V_t) Y_{it}(0) + (U_i V_t) Y_{it}(1)$
- Treatment effect on the treated:

$$\begin{aligned} \tau(\mathbf{U}, \mathbf{V}) &= \sum_{i=1}^{N} \sum_{t=1}^{T} U_i V_t \Big(Y_{it}(1) - Y_{it}(0) \Big) \\ &= \sum_{i=1}^{N} \sum_{t=1}^{T} U_i V_t Y_{it} - \sum_{i=1}^{N} \sum_{t=1}^{T} U_i V_t Y_{it}(0) \end{aligned}$$

Potential-outcomes setup, last-period treatment

- Potential outcomes: $Y_{iT}(0), Y_{iT}(1)$
- Observed outcome: $Y_{iT} = (1 - U_i)Y_{iT}(0) + U_i Y_{iT}(1)$
- Treatment effect on the treated:

$$\begin{aligned}\tau(\mathbf{U}) &= \sum_{i=1}^{N} U_i \Big(Y_{iT}(1) - Y_{iT}(0) \Big) \\ &= \sum_{i=1}^{N} U_i Y_{iT} - \sum_{i=1}^{N} U_i Y_{iT}(0)\end{aligned}$$

Generalized Synthetic Control (GSC) estimators

$$\hat{\tau}(\mathbf{U}, \mathbf{Y}|\mathcal{M}) = \sum_{i=1}^{N} U_i \left\{ \mathbf{M}_{i0} + \sum_{j=1}^{N} \mathbf{M}_{ij} Y_{jT} \right\}$$

$$\mathbf{M}(\mathbf{Y}, \mathcal{M}) \equiv \arg\min_{\mathbf{M} \in \mathcal{M}} \sum_{i=1}^{N} \sum_{t=1}^{T-1} \left(\mathbf{M}_{i0} + \sum_{j=1}^{N} \mathbf{M}_{ij} Y_{jt} \right)^2$$

- $\mathbf{M}_{ij}$ gives weight for control unit j given treated unit i
- $\mathbf{M}_{i0}$ is the intercept
- $\mathcal{M}$ is the set of possible weight matrices

Synthetic Control (SC) estimator [Abadie et al., 2010]

$$\mathcal{M}^{SC} = \left\{ \mathbf{M} \middle| \mathbf{M}_{ii} = 1 \forall j \geq 1; \mathbf{M}_{ij} \leq 0 \forall i \geq 1, j \neq i; \right.$$
$$\left. \sum_{j=1}^{N} \mathbf{M}_{ij} = 0 \ \forall i \geq 1; \mathbf{M}_{i0} = 0 \ \forall i \right\}$$

Example:

$$\mathbf{M} = \begin{pmatrix} 0 & 1 & -0.00 & -1.00 & -0.00 \\ 0 & -0.00 & 1 & -0.00 & -1.00 \\ 0 & -0.45 & -0.40 & 1 & -0.15 \\ 0 & -0.03 & -0.74 & -0.24 & 1 \end{pmatrix}$$

General restrictions, Modified Synthetic Control (MSC) estimator [Doudchenko and Imbens, 2016, Ferman and Pinto, 2017]

$$\mathcal{M}^0 = \left\{ \mathbf{M} \middle| \mathbf{M}_{ii} = 1 \forall j \geq 1; \mathbf{M}_{ij} \leq 0 \forall i \geq 1, j \neq i; \sum_{j=1}^{N} \mathbf{M}_{ij} = 0 \ \forall i \geq 1 \right\}$$

Example:

$$\mathbf{M} = \begin{pmatrix} -0.17 & 1 & -0.00 & -1.00 & -0.00 \\ 0.10 & -0.03 & 1 & -0.27 & -0.70 \\ -0.00 & -0.44 & -0.42 & 1 & -0.14 \\ -0.06 & -0.00 & -1.00 & -0.00 & 1 \end{pmatrix}$$

Difference-in-means (DiM) estimator

$$\mathcal{M}^{\text{DiM}} = \left\{ \mathbf{M} \middle| \mathbf{M}_{ii} = 1 \forall i \geq 1; \mathbf{M}_{ij} = -\frac{1}{N-1} \forall i \geq 1, j \neq i; \mathbf{M}_{i0} = 0 \ \forall i \right\}$$

Example:

$$\mathbf{M} = \begin{pmatrix} 0 & 1 & -1/3 & -1/3 & -1/3 \\ 0 & -1/3 & 1 & -1/3 & -1/3 \\ 0 & -1/3 & -1/3 & 1 & -1/3 \\ 0 & -1/3 & -1/3 & -1/3 & 1 \end{pmatrix}$$

Plan for the talk

1. Generalized Synthetic Control (GSC) estimators
2. Properties of GSC estimators under randomization
3. Unbiased and modified unbiased SC estimators
4. Variance estimation
5. Extensions
6. Take-aways for applying GSC estimators

Treatment assignment

Last-period assignment

$$V_t = 1_{t=T}$$

Random treatment assignment

$$\text{pr}(\mathbf{U} = \mathbf{u}) = \begin{cases} \frac{1}{N} & \text{if} \quad u_i \in \{0,1\} \, \forall i, \quad \sum_{i=1}^{N} u_i = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Bias of GSC estimators

$$\text{Bias} = \mathbb{E}\left[\hat{\tau} - \tau\right] = \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{j=1}^{N}\mathbf{M}_{ji}\right)Y_{iT}(0)$$

- SC (and MSC) generally biased under unit randomization

$$\mathbf{M} = \left(\begin{array}{ccccc|c} 0 & 1 & -0.00 & -1.00 & -0.00 & 0 \\ 0 & -0.00 & 1 & -0.00 & -1.00 & 0 \\ 0 & -0.45 & -0.40 & 1 & -0.15 & 0 \\ 0 & -0.03 & -0.74 & -0.24 & 1 & 0 \\ \hline & 0.53 & -0.14 & -0.24 & -0.15 & \end{array}\right)$$

Plan for the talk

1 Generalized Synthetic Control (GSC) estimators

2 Properties of GSC estimators under randomization

3 Unbiased and modified unbiased SC estimators

4 Variance estimation

5 Extensions

6 Take-aways for applying GSC estimators

Unbiased Synthetic Control (USC) estimator

$$\mathcal{M}^{USC} = \left\{ \mathbf{M} \in \mathcal{M}^0 \middle| M_{i0} = 0 \ \forall i, \sum_i \mathbf{M}_{ij} = 0 \ \forall j \geq 1 \right\}$$

Example:

$$\mathbf{M} = \begin{pmatrix} 0 & 1 & -0.00 & -1.00 & -0.00 \\ 0 & -0.08 & 1 & -0.00 & -0.92 \\ 0 & -0.62 & -0.30 & 1 & -0.08 \\ 0 & -0.30 & -0.70 & -0.00 & 1 \end{pmatrix} \begin{array}{|c} 0 \\ 0 \\ 0 \\ 0 \end{array}$$

$$\begin{array}{cccc} \hline 0 & 0 & 0 & 0 \end{array}$$

Modified Unbiased Synthetic Control (MUSC) estimator

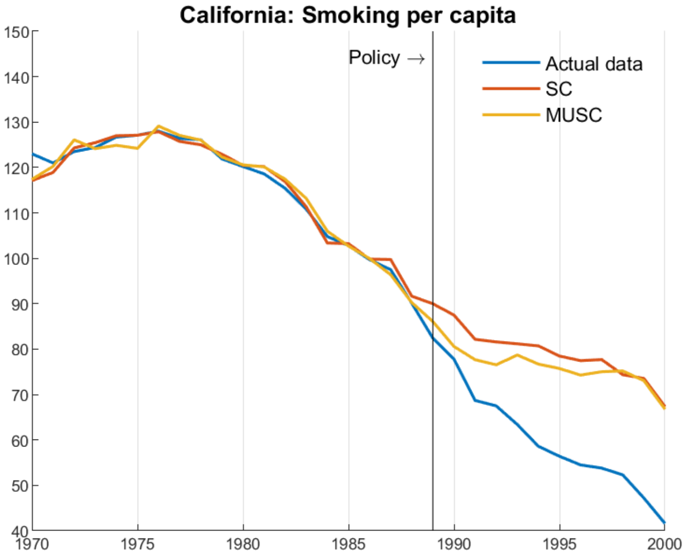
$$\mathcal{M}^{MUSC} = \left\{ \mathbf{M} \in \mathcal{M}^0 \middle| \sum_i \mathbf{M}_{ij} = 0 \ \forall j \geq 1 \right\}$$

Example:

$$\mathbf{M} = \begin{pmatrix} -0.17 & 1 & -0.00 & -1.00 & -0.00 \\ 0.14 & -0.30 & 1 & -0.00 & -0.70 \\ 0.07 & -0.63 & -0.07 & 1 & -0.30 \\ -0.04 & -0.07 & -0.93 & -0.00 & 1 \end{pmatrix} \begin{array}{|c} 0 \\ 0 \\ 0 \\ 0 \end{array}$$

$$\begin{array}{cccc} \hline 0 & 0 & 0 & 0 \end{array}$$

California smoking study [Abadie et al., 2010]



Simulation

- Data set: Current Population Survey (CPS)
- Outcome variables: log wages ($sd = 0.44$), unemployment rate ($sd = 0.02$), hours ($sd = 1.34$)
- $N = 50$ states
- $T = 21 \cdots 40$ years
- Since there is no actual treatment, $Y_{iT}(0) = Y_{iT}(1)$, so $\tau = 0$ and we can calculate bias and RMSE

Simulation results: bias (at $T = 40$)

Outcome	DiM	SC	MSC	USC	MUSC
Log Wages	0	-0.0067	-0.0025	0	0
Hours	0	0.1128	0.0188	0	0
Unemp. Rate	0	-0.0010	-0.0006	0	0

- DiM, USC and MUSC are indeed unbiased
- SC and MSC are biased

Simulation results: RMSE (averaged over $T = 21 \cdots 40$)

Outcome	DiM	SC	MSC	USC	MUSC
Log Wages	0.1047	0.0510	0.0533	0.0516	0.0530
Hours	1.1974	0.9180	0.8658	0.9136	0.9031
Unemp. Rate	0.0150	0.0130	0.0129	0.0131	0.0129

- Large RMSE improvement over DiM for all other estimators for log wages and hours
- Smaller RMSE improvement over DiM for unemploment rate
- SC, MSC, USC and MUSC all have similar RMSE

Plan for the talk

1. Generalized Synthetic Control (GSC) estimators
2. Properties of GSC estimators under randomization
3. Unbiased and modified unbiased SC estimators
4. **Variance estimation**
5. Extensions
6. Take-aways for applying GSC estimators

Variance of unbiased GSC estimators

$$\mathbb{V}(\mathbf{M}) = \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{M}_{i0} + \sum_{j=1}^{N}\mathbf{M}_{ij}Y_{jT}(0)\right)^{2}$$

- Depends on realized weights and untreated potential outcomes
- Also the mean-squared error for biased GSC estimators

Unbiased variance estimator

$$\hat{\mathbb{V}}(\mathbf{U},\mathbf{Y},\mathbf{M}) = \sum_i U_i \Bigg\{ \frac{1}{N-3} \sum_{\substack{k=1\\k\neq i}}^{N} \left(\sum_{\substack{j=1\\j\neq i}}^{N} \mathbf{M}_{kj}(Y_{kT}-Y_{jT}) \right)^2 - \frac{1}{(N-2)(N-3)} \sum_{\substack{k=1\\k\neq i}}^{N} \sum_{\substack{j=1\\j\neq i}}^{N} \mathbf{M}_{kj}^2 (Y_{kT}-Y_{jT})^2 + \frac{2}{N-2} \sum_{\substack{k=1\\k\neq i}}^{N} \mathbf{M}_{k0} \left(\sum_{\substack{j=1\\j\neq i}}^{N} \mathbf{M}_{kj}(Y_{jT}-Y_{kT}) \right) + \frac{1}{N} \sum_{k=1}^{N} \mathbf{M}_{k0}^2 \Bigg\}$$

- Finite-sample unbiased for $\mathbb{V}(\mathbf{V},\mathbf{M})$
- Variance estimator can be negative
- Reduces to standard DiM var estimator with equal weights

Performance of variance estimators for $N = 10$ subset and log wages

	DiM	SC	MUSC
$\sqrt{\mathrm{V}}$	0.1031	0.0494	0.0475
$\sqrt{\mathbb{E}\hat{\mathrm{V}}}_{GSC}$	0.1031	0.0494	0.0475
$\sqrt{\mathbb{E}\hat{\mathrm{V}}}_{\text{placebo}}$	0.1038	0.0500	0.0481

- Placebo variance upward biased for these estimators but may be downward biased

Plan for the talk

1 Generalized Synthetic Control (GSC) estimators

2 Properties of GSC estimators under randomization

3 Unbiased and modified unbiased SC estimators

4 Variance estimation

5 Extensions

6 Take-aways for applying GSC estimators

Extensions

1. Time randomization
   - Provides a justification for GSC optimization problem
2. Multiple treated units
   - Extends MUSC to many treated units for estimation of average treatment effect on the treated
3. Propensity scores
   - Allows application if propensity scores vary
   - Allows for optimization of propensity scores when GSC used
   - In observational data provides alternative to binary inclusion

Plan for the talk

1. Generalized Synthetic Control (GSC) estimators
2. Properties of GSC estimators under randomization
3. Unbiased and modified unbiased SC estimators
4. Variance estimation
5. Extensions
6. Take-aways for applying GSC estimators

Take-aways

- Analyze SC-type estimators under randomization
- Main results:
  1. Standard SC generically biased under randomization
  2. Bias can be fixed by adding a constraint on weights
  3. DiM variance estimator can be extended to SC-type estimators
- Main take-aways:
  1. For experiment:
     - MUSC as practical alternative to DiM
     - Variance estimation and multiple units extend
     - Reduces RMSE without sacrificing bias
  2. For observational data:
     - Additional robustness at small or no cost in RMSE
     - Alternative (view on) variance estimation
     - Propensity scores as alternatives to binary inclusion

References I

Abadie, A. (2019).
Using synthetic controls: Feasibility, data requirements, and methodological aspects.
Journal of Economic Literature.

Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. M. (2020).
Sampling-based versus design-based uncertainty in regression analysis.
Econometrica, 88(1):265–296.

Abadie, A., Diamond, A., and 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.

Abadie, A., Diamond, A., and Hainmueller, J. (2015).
Comparative politics and the synthetic control method.
American Journal of Political Science, pages 495–510.

Abadie, A. and Gardeazabal, J. (2003).
The economic costs of conflict: A case study of the basque country.
American Economic Review, 93(-):113–132.

Abadie, A. and L'Hour, J. (2017).
A penalized synthetic control estimator for disaggregated data.
Work. Pap., Mass. Inst. Technol., Cambridge, MA.

Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W., and Wager, S. (2019).
Synthetic difference in differences.
Technical report, National Bureau of Economic Research.

References II

Ben-Michael, E., Feller, A., and Rothstein, J. (2020).
The augmented synthetic control method.
arXiv preprint arXiv:1811.04170.

Chernozhukov, V., Wuthrich, K., and Zhu, Y. (2017).
An exact and robust conformal inference method for counterfactual and synthetic controls.
arXiv preprint arXiv:1712.09089.

Doudchenko, N. and Imbens, G. W. (2016).
Balancing, regression, difference-in-differences and synthetic control methods: A synthesis.
Technical report, National Bureau of Economic Research.

Ferman, B. and Pinto, C. (2017).
Placebo tests for synthetic controls.

Hahn, J. and Shi, R. (2016).
Synthetic control and inference.
Available at UCLA.

Imbens, G. W. and Rubin, D. B. (2015).
Causal Inference in Statistics, Social, and Biomedical Sciences.
Cambridge University Press.

Lei, L. and Candès, E. J. (2020).
Conformal inference of counterfactuals and individual treatment effects.
arXiv preprint arXiv:2006.06138.

References III

Li, K. T. (2020).
Statistical inference for average treatment effects estimated by synthetic control methods.
Journal of the American Statistical Association, 115(532):2068–2083.

Neyman, J. (1923/1990).
On the application of probability theory to agricultural experiments. essay on principles. section 9.
Statistical Science, 5(4):465–472.

Rambachan, A. and Roth, J. (2020).
Design-Based Uncertainty for Quasi-Experiments.
arXiv preprint arXiv:2008.00602.

Sekhon, J. S. and Shem-Tov, Y. (2020).
Inference on a new class of sample average treatment effects.
Journal of the American Statistical Association, pages 1–7.