Standard and Generalized Synthetic Control Methods [*]

Yi Chen

May 4, 2024

1 Introduction

Researchers are often interested in the causal effect of a certain policy or event taking place at an aggregate level. To estimate that effect, they usually use comparable units as their controls. For example, researchers may want to investigate the effect of a new policy imposed in 2010 on country A and they know that country B does not have such a policy in practice while the two countries are very similar and comparable. Thus, they will use country B to serve as control units to estimate the causal effect of the new policy. This is an ideal scenario.

However, in reality, it is often difficult to find a single unit that is almost the same as the treated unit (the unit where the policy takes place). To address this problem, Alberto Abadie and Hainmueller (2010) propose synthetic control methods[1]. The basic idea behind this method is to find a combination of different units that can work together to serve as a comparable unit to the treated unit. This combination is called a synthetic control, which is a weighted average of the available control units. It makes explicit the relative contribution of each control unit to the counterfactual of interest. Meanwhile, it also makes clear the similarities between the unit affected by the policy of interest and the synthetic control, in terms of pre-intervention outcomes and other predictors of post-intervention outcomes.

Despite the smart idea of the standard synthetic methods, it has some critical limitations: (1) the restrictive constraints are too strong; (2) there is no inference theory of this method. Therefore, Xu (2017) proposes a generalized synthetic control (GSC) method using an interactive fixed model. This model adopts the same function form as in the standard synthetic control method while employing a latent approach to address causal inference problems and provide valid simulation-based uncertainty estimates.

The rest of the report is organized as follows. Section 2 introduces the framework of standard synthetic controls. Section 3 sets up the generalized synthetic control method. Section 4 compares the performance of the two methods via simulation. The last section concludes.

[*] This is the final project report for STA640. The video and the reproducible code for the project can be found here.

[1] To distinguish with the generalized synthetic control methods I introduce later, in this report, I will use "standard synthetic control methods" to refer to this method.

2 Standard synthetic control methods

Suppose that we observe $J+1$ regions. Without loss of generality, suppose only the first region is exposed to the intervention of interest, that is, it serves as the treated unit so we have J remaining regions as the potential controls. Let Y_{it}^N be the outcome for region i at time t in the absence of the intervention, for $i = 1, ..., J+1$ and $t = 1, ..., T$. Let T_0 be the number of pre-intervention units, with $1 \leq T_0 \leq T$. Let Y_{it}^I be the outcome for unit i at time t if uit i is exposed to the intervention in periods $T_0 + 1$ to T. We assume that the intervention has no effect on the outcome before the implementation period, so for $t \in \{1, ..., T_0\}$ and all $i \in \{1, ..., J+1\}$, we have $Y_{it}^I = Y_{it}^N$. Then let $\alpha_{it} = Y_{it}^I - Y_{it}^N$ be the effect of the intervention for unit i at time t. Thus

$$Y_{it}^I = Y_{it}^N + \alpha_{it}$$

Let D_{it} be an indicator that equals to 1 when unit i is exposed to the intervention at time t and 0 otherwise. Then the observed outcome for unit i at time t is

$$Y_{it} = Y_{it}^N + \alpha_{it} D_{it}$$

As in our context, only the first region is exposed to the intervention and only after period T_0, we should have

$$D_{it} = \begin{cases} 1 & \text{if } i = 1 \text{ and } t > T_0, \\ 0 & \text{otherwise.} \end{cases}$$

We aim to estimate $(\alpha_{1,T_0+1}, ..., \alpha_{1T})$, where for $t > T_0$, $\alpha_{1t} = Y_{it}^I - Y_{it}^N = Y_{it} - Y_{it}^N$. Suppose Y_{it}^N is given by the following factor model

$$Y_{it}^N = \delta_t + \beta X_{it} + \lambda_t \mu_i + \epsilon_{it}, \tag{1}$$

where δ_t is a time fixed effect, X_{it} is a $(L \times 1)$ vector of observed covariates, β is a $(1 \times L)$ vector of parameters, λ_t is an unknown common factor with unit-specific factor loadings μ_i, and ϵ_{it} is the error term with zero mean for all i. Note X_{it} may contain pre- and post-intervention values of time-varying variables, as long as they are not affected by the intervention.

To construct a synthetic control, consider a $(J \times 1)$ vector of weights $W = (w_2, ..., w_{J+1})'$ such that $w_j \geq 0$ for $j = 2, ..., J+1$ serving as the weight of unit j, and $w_2 + ... + w_{J+1} = 1$. The value of outcome for synthetic control is given by

$$\sum_{j=2}^{J+1} w_j Y_{jt} = \delta_t + \beta \sum_{j=2}^{J+1} w_j X_{jt} + \lambda_t \sum_{j=2}^{J+1} w_j \mu_j + \sum_{j=2}^{J+1} w_j \epsilon_{jt}$$

Suppose that there are $(w_2^*, ..., w_{J+1}^*)$ such that[2]

$$\begin{aligned} \sum_{j=2}^{J+1} w_j^* Y_{jt} = Y_{1t}, \text{ for } t = 1, ..., T_0 \\ \sum_{j=2}^{J+1} w_j^* X_{jt} = X_{1t} \end{aligned} \tag{2}$$

[2] In practice, it is usually impossible to find W that can make the equation (2) hold exactly, we usually try to find $\hat{W} = \arg\min_{W \in \Lambda} \sum_{t=1}^{T_0} (Y_{1t} - \sum_{j=2}^{J+1} w_j Y_{jt})^2$ and $\hat{W} = \arg\min_{W \in \Lambda} \sum_{t=1}^{T_0} (X_{1t} - \sum_{j=2}^{J+1} w_j X_{jt})^2$ instead.

and if $\sum_{t=1}^{T_0} \lambda_t' \lambda_t$ is nonsingular, then

$$Y_{1t}^N - \sum_{j=2}^{J+1} w_j^* Y_{jt} = \sum_{j=2}^{J+1} w_j^* \sum_{s=1}^{T_0} \lambda_t \left(\sum_{n=1}^{T_0} \lambda_n' \lambda_n \right)^{-1} \lambda_s' (\epsilon_{js} - \epsilon_{1s}) - \sum_{j=2}^{J+1} w_j^* (\epsilon_{jt} - \epsilon_{1t})$$

Under standard conditions (see in Appendix B of Alberto Abadie and Hainmueller (2010)), the mean of right-hand side will be close to zero if the number of preintervention periods is large enough. Intuitively, by considering enough preintervention periods, we can account for unobservables. Therefore, we can use the following approximate estimator to estimate α_{1t}:

$$\hat{\alpha}_{1t} = Y_{it} - \sum_{j=2}^{J+1} w_j^* Y_{jt}$$

for $t \in \{T_0 + 1, ..., T\}$.

However, as pointed out by Doudchenko and Imbens (2017), there are three main limitations of this method. First, with no-intercept constraint in the factor model (1), it actually assumes within the synthetic control, the weight of the treated unit is $w_1 = 0$. This rules out the possibility of a systematic difference between the treated unit and the synthetic control unit. Second, the sum-to-one restriction ($\sum_{j=2}^{J+1} = 1$) is implausible if a treated unit is on the extreme end of the distribution units, for example, when the treated unit is the largest or smallest in terms of outcome values. Third, the nonnegative constraint ($w_j \geq 0$) implicitly assumes a positive correlation between outcomes of the treated unit and control units, while the reality may be the opposite.

3 Generalized synthetic control method

Not restricted by this convex combination form of control units, Xu (2017) proposes a generalized synthetic control method to link the synthetic control methods with linear fixed effect models.

The assumption on the function form of Y_{it}^N is the same as in (1) of the standard synthetic method. The following equation just puts α_{it} back into the model for generalization.

$$Y_{it} = \alpha_{it} D_{it} + \beta X_{it} + \lambda_t \mu_i + \epsilon_{it} \tag{3}$$

The factor component of the model, $\lambda_t \mu_i = \lambda_{1t}\mu_{i1} + \lambda_{2t}\mu_{i2} + ... + \lambda_{rt}\mu_{ir}$, takes a linear additive form by assumption. Here r is the dimension of the unobserved factors[3]. Specifically, additive unit and time fixed effects can be viewed as a special case when setting $\lambda_{1t} = 1$ and $\mu_{i2} = 1$.
We can pool all time periods together to rewrite the model of each unit as

$$Y_i = D_i \cdot \alpha_i + X_i \beta + \Lambda \mu_i + \epsilon_i \tag{4}$$

where $Y_i = (Y_{i1}, ..., Y_{iT})'$, $D_i = (D_{i1}, ..., D_{iT})'$ and $\alpha_i = (\alpha_{i1}, ..., \alpha_{iT})'$, $\epsilon_i = (\epsilon_{i1}, ..., \epsilon_{iT})'$ are $(T \times 1)$ vectors; $X_i = (X_{i1}, ..., x_{iT})'$ is a $(T \times L)$ matrix; and $\Lambda = (\lambda_1, ..., \lambda_T)'$ is a $(T \times r)$ matrix. Stacking all control units together, we have

$$Y_{co} = X_{co}\beta + \Lambda M_{co}' + \epsilon_{co}$$

[3] Here we assume r is known. In practice, researchers may have limited knowledge of the exact number of factors r. The chosen of r can be done by using a cross-validation procedure to select models before estimating causal effects. The procedure is described in Xu (2017).

where Y_{co} and ϵ_{co} is a ($T \times N_{co}$[4]) matrix of control units, X_{co} is a three-dimensional ($T \times N_{co} \times L$) matrix, and $M = (\mu_1, ..., \mu_{N_{co}})'$ is a ($N_{co} \times r$) matrix.

The main quantity of interest is the average treatment effect on the treated (ATT) at time t (when $t > T_0$)

$$ATT_{t,t>T_0} = \frac{1}{N_t} \sum_{i \in \mathcal{T}} [Y_{it}(1) - Y_{it}(0)] = \frac{1}{N_t} \sum_{i \in \mathcal{T}} \alpha_{it}$$

where $\mathcal{T}$ is the set of treated units and N_t is the number of treated units.

To estimate the parameters in this interactive fixed effects (IFE) model, we first use only control group data to obtain $\hat{\beta}, \hat{\Lambda}, \hat{M}_{co}$:

$$(\hat{\beta}, \hat{\Lambda}, \hat{M}) = \arg \min_{\tilde{\beta}, \tilde{\Lambda}, \tilde{M}_{co}} \sum_{i \in \mathcal{C}} (Y_i - X_i \tilde{\beta} - \tilde{\Lambda} \tilde{\mu}_i)'(Y_i - X_i \tilde{\beta} - \tilde{\Lambda} \tilde{\mu}_i)$$

$$\text{s.t. } \tilde{\Lambda}'\tilde{\Lambda}/T = I_r \text{ and } \tilde{M}'_{co}\tilde{M}_{co} = \text{diagonal}$$

The second step estimates factor loadings μ_i for each treated unit i by minimizing the mean squared error of the predicted treated outcome in pretreatment periods:

$$\hat{\mu_i} = \arg \min_{\tilde{\mu}_i} (Y_i^0 - X_i^0 \hat{\beta} - \hat{\Lambda}^0 \tilde{\mu}_i)'(Y_i^0 - X_i^0 \hat{\beta} - \hat{\Lambda}^0 \tilde{\mu}_i)$$

where $\hat{\beta}$ and $\hat{\Lambda}^0$ are from the first-step estimation. Here 0 denotes the pretreatment periods. The last step is to calculate treated counterfactuals based on the estimated parameters and factor loadings

$$\hat{Y}_{it}(0) = \hat{\beta} X_{it} + \hat{\lambda}_t \hat{\mu}_i, \; i \in \mathcal{T}, t > T_0$$

There are two main advantages of this GSC method. First, it generalizes the synthetic control method to cases of multiple treated units and variable treatment periods. Since the IFE model is estimated only once, it does not require obtaining weights for each treated unit one by one as in standard synthetic control methods. Second, it produces frequentist uncertainty estimates, including standard errors and confidence intervals, making the estimates easy to do inference.

4 Simulation

In this section, I compare the performance of the two methods. I start with the following data generating process (DGP)

$$Y_{it} = \alpha_{it} D_{it} + 1 \cdot x_{it,1} + 3 \cdot x_{it,2} + \lambda_t \mu_i + \eta_i + \xi_t + 5 + \epsilon_{it} \tag{5}$$

In this DGP, I include two observed covarites ($x_{it,1}$ and $x_{it,2}$), two unobserved factors (the number of factors in λ_t is 2), and additive two-way fixed effects, η_i and ξ_t. To simplify my analysis, I specify only one treated unit. This is primarily due to the limitation of the standard synthetic control method, as it requires constructing a combination of controls for each treated unit one by one. I set $N_{co} = 40$ to allow for 40 potential control units, $T_0 = 20$ and $T = 30$.

After the generation of simulated data, I use packages `Synth` and `gsynth` in `R` respectively for the standard and generalized synthetic control methods to estimate the treatment effect. The figure (1) visualizes the fit of synthetic controls and estimation results. In the upper panel, the solid

[4] N_{co} is the number of control units.

line is the treated outcome while the dashed line is the estimated $Y(0)$ based on the outcome of constructed synthetic control. We can see that in the pre-intervention periods ($T = 1 \sim 20$), the outcome of synthetic control is not as close to that of the treated unit as we hope, demonstrating a not satisfying performance of the construction of synthetic control. This may be due to the few covariates in the simulated data that make it hard to find suitable weights for each potential control. This is very common in real-world research and it illustrates the limitation of the standard synthetic control methods. Similarly, for the estimation of causal effect in the lower panel, simply from the graph, it does not look to perform very well. Moreover, the biggest issue is that without inference theory to compute confidence intervals for the estimates from this method, we cannot tell how good the estimation is.

Next, I apply GSC on the simulated data. Figure (2) shows the estimation results of generalized synthetic methods. Similar to Figure (1), the upper panel visualizes the outcomes of treated unit and synthetic control. Observed from the pre-intervention period, it is clear that this method performs much better in estimating $Y(0)$ for the treated than the standard synthetic method. The lower panel shows the estimated causal effect and the true causal effect. The 95% interval is based on bootstraps of 1000 times. It shows that the estimated causal effect for the treated outcome fits the data well and is very close to the true effects. The results together demonstrate the advantages of the generalized synthetic methods over the standard one.

5 Conclusion and discussion

This report introduces and compares standard and generalized synthetic control methods. The GSC method adopts the foundational idea of the standard synthetic control method. It uses pre-treatment data to construct a reweighting scheme for control units, enhancing predictions of treated counterfactuals. For the three limitations of standard synthetic control methods mentioned in Section 2, the generalized method solves the last two, that is, the sum-of-one constraint and the positive weight restriction. Since the estimation of parameters are solely based on control data, the first limitation about symmetric difference between treated units and control units are not solved. This can be further improved by using both control and treatment group for parameter estimation, as Pinkney does (Pinkney, 2021). Besides handling the last two limitations, the method makes two more improvements. First, it supports multiple treated units with variable treatment timings. Second, it provides clear uncertainty estimates, including standard errors and confidence intervals.

Research and development within this domain continue to thrive. Building on the Generalized Synthetic Control (GSC) model, scholars have investigated a Bayesian alternative synthetic control method (Pinkney, 2021; Pang et al., 2022). Specifically, they use shrinkage priors to choose the number of unobserved factors and decide whether and how to include a covariate. Additionally, other researchers, such as Nazaret et al. (2023), have focused on addressing the issues related to the misspecification of linear assumptions in synthetic control methods. This ongoing work highlights the dynamic and evolving nature of research in synthetic control methodologies.

References

Alberto Abadie, A. D., & 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. https://doi.org/10.1198/jasa.2009.ap08746

Doudchenko, N., & Imbens, G. W. (2017). Balancing, regression, difference-in-differences and synthetic control methods: A synthesis.

Nazaret, A., Shi, C., & Blei, D. M. (2023). On the misspecification of linear assumptions in synthetic control.

Pang, X., Liu, L., & Xu, Y. (2022). A bayesian alternative to synthetic control for comparative case studies. *Political Analysis*, *30*(2), 269–288. https://doi.org/10.1017/pan.2021.22

Pinkney, S. (2021). An improved and extended bayesian synthetic control.

Xu, Y. (2017). Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis*, *25*(1), 57–76. https://doi.org/10.1017/pan.2016.2

6 Appendix

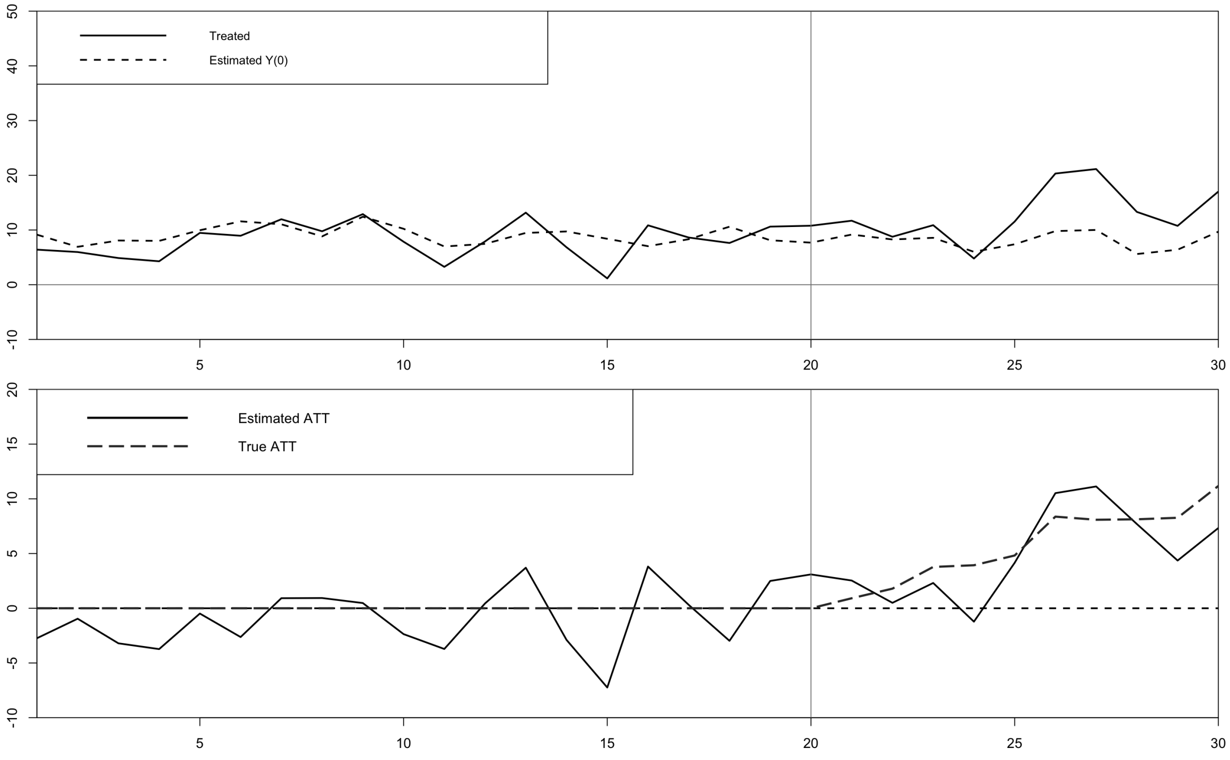

Figure 1: Estimation results of standard synthetic control methods

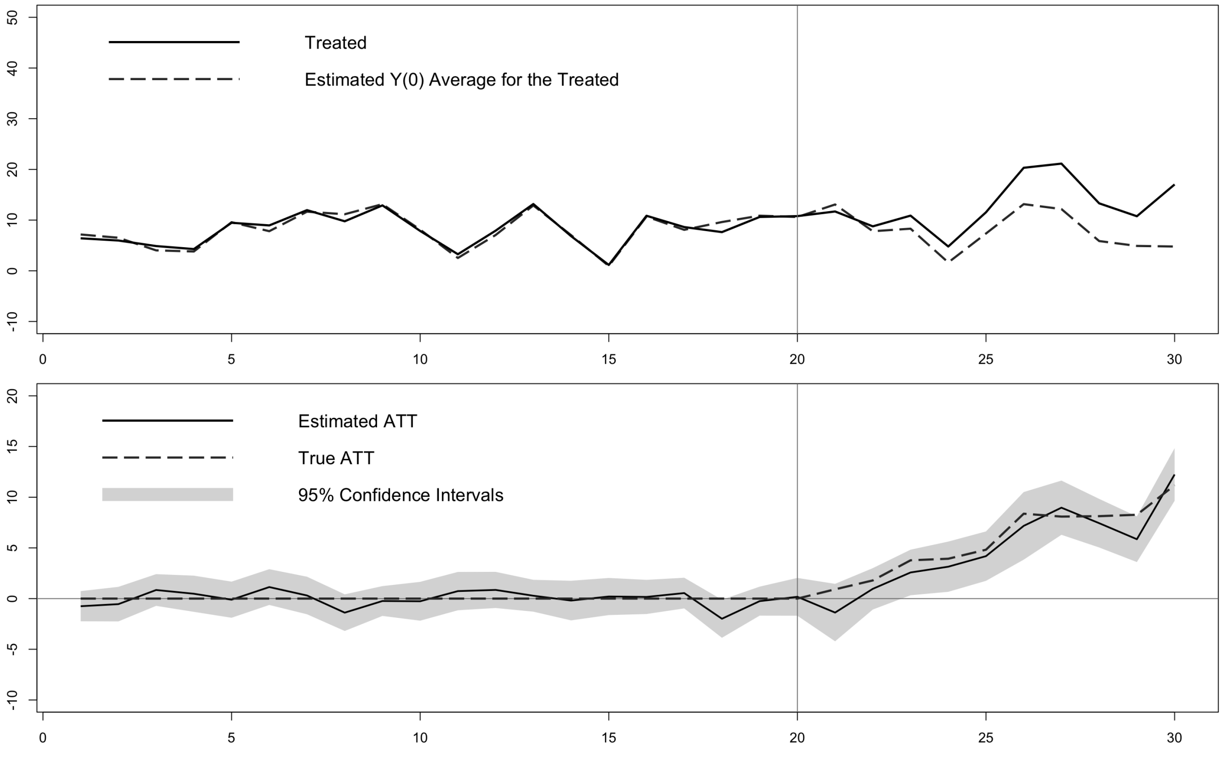

Figure 2: Estimation results of generalized synthetic control methods