Synthetic Controls Note

Vasco Villas-Boas

December 19, 2025

1 Econometrics

In the simplest case where one can apply synthetic controls, we assume there are $j \in [0 : J]$ units[1] that are observed in periods $t \in [T]$. Unit $j = 0$ is exposed to an intervention at times $t \in \{T_0 + 1, ..., T\}$ and the remaining J units are an untreated reservoir of potential controls (ie., a set of "donors"). We assume that Y_{jt}^A is the outcome that would be observed for unit j at time t in the absence of the intervention and that Y_{jt}^I is the outcome that would be observed for unit j at time t if j is then exposed to the treatment. We also let Y_{jt} be the actual observed outcome for unit j at time t. Given these definitions we have that

$$\begin{aligned} &Y_{jt} = Y_{jt}^A \; \forall j \in [J], t \in [T] \\ &Y_{0t} = Y_{0t}^A \; \forall t \in [T_0] \text{ and } Y_{0t} = Y_{0t}^I \; \forall t \in [T] \setminus [T_0] \end{aligned}$$

The goal is to estimate $\tau_{0t} := Y_{0t}^I - Y_{0t}^A = Y_{0t} - Y_{0t}^A$ for $t \in [T] \setminus [T_0]$. Since Y_{0t} is observed, it suffices to estimate Y_{0t}^A to get an estimate of τ_{0t}. We also assume that we have access to *pre-intervention* characteristics $X_j \in \mathbb{R}^K$ for each $j \in [0 : J]$. As useful auxiliary quantities, define $X := [X_1, ..., X_J] \in \mathbb{R}^{K\times J}$ and $Y_t := [Y_{1t}, ..., Y_{Jt}]' \in \mathbb{R}^J$ which contain the control unit covariates and outcomes at a time t, respectively.

For a currently unspecified symmetric PSD matrix V,[2] the synthetic controls estimator estimates τ_{0t} by first solving the problem[3]

$$w^*(V) = \underset{w\in\Delta^{J-1}}{\arg\min}(X_0 - Xw)'V(X_0 - Xw) \tag{1}$$

and then estimating

$$\hat{\tau}_{0t} := Y_{0t} - (w^*(V))'Y_t \tag{2}$$

for $t \in [T] \setminus [T_0]$. Finally, in one approach to choose V, it's picked to solve the following problem

$$V^* := \underset{V}{\arg\min} \sum_{t\in[T_0]} \left(Y_{0t} - (w^*(V))'Y_t\right)^2$$

where $w^*(V)$ is determined as the solution to the problem in Equation (1).

What is the intuition behind this estimator? We would like to know the effect of the intervention at intervention times $t \in [T] \setminus [T_0]$ on the impacted unit $j = 0$. To come up with this estimated effect, we would like to have a good idea of what the trajectory of unit $j = 0$ would have been during the intervention periods, in the absence of the intervention, to know Y_{0t}^A. In the estimator, we seek to find convex weights w^* such that we can form a synthetic version of the intervened unit $j = 0$ using the control units $[J]$ that tracks the unit in the absence of the treatment. We do that by minimizing the distance parametrized by V between the pre-intervention covariates of unit $j = 0$ and the convex hull of the control unit covariates X. If we see that the synthetic unit tracks the pre-intervention outcomes of unit $j = 0$ well (as measured by

[1] I use the notation that $[N] := \{1, ..., N\}$ and $[0 : N] = \{0, 1, ..., N\}$.

[2] In many cases, V is taken to be a PSD diagonal matrix in which case each diagonal entry can be interpreted as the relative predictive power of that covariate on the outcome.

[3] I use the notation that $\Delta^{J-1} := \{x \in \mathbb{R}^J : x_j \geq 0 \, \forall j \in [J] \text{ and } \sum_{j\in[J]} x_j = 1\}$.

$\frac{1}{T_0}\sum_{t\in[T_0]}(Y_{0t}-(w^*(V))'Y_t)^2$ being small), then we might believe that the synthetic unit would also track unit $j = 0$ in absence of the treatment in the post-intervention periods. Since the synthetic unit is "observed" (or rather all of its constituents are observed) in the post-intervention period, under the assumption that the synthetic unit reasonably tracks the intervened unit in the post-intervention periods in the absence of the treatment, we can form the estimator written in Equation (2) for the treatment effect in period t on unit $j = 0$.

Even if we assume that we have perfect knowledge of unit covariates and outcomes for $j \in [0 : J]$ and $t \in [T]$, there's still uncertainty in the estimated treatment effect since we might not fully believe that our synthetic unit perfectly tracks the counterfactual outcome for the intervened unit (in the intervention periods). In this setting, large sample asymptotics are not the approach that we would like to take to conduct inference. In the setting of synthetic controls, the inference strategy proposed by Abadie et al. (2010) is one of placebo or permutation tests to conduct exact inference. The idea is that we run an exercise where we one-at-a-time consider labeling the other $j \in [J]$ units as the intervened unit and $j = 0$ as a non-intervened unit. We then compute $\hat{\tau}_{jt}$ for each unit $j \in [0 : J]$ when it's a placebo treated unit using the same approach as above and for each $t \in [T]$. Finally, for any unit $j \in [0 : J]$, we form the test statistic:

$$r_j := \frac{R_j(T_0+1,T)}{R_j(1,T_0)} \tag{3}$$

$$\text{where } R_j(t_1,t_2) = \left(\frac{1}{t_2-t_1+1}\sum_{t=t_1}^{t_2}\hat{\tau}_{jt}^2\right)^{1/2}$$

The motivation for dividing the test statistic by the "pre-intervention" fit is that the "pre-intervention" fit on the outcome may be of different quality for different units and what we care about is if there's a significant change in "pre" and "post" intervention fit. We then look to see where r_0 lives in the distribution $\{r_0, ..., r_J\}$ and use its ranking to determine a p-value.[4]

Synthetic controls estimators are typically motivate the counterfactual outcome with a linear-factor model data-generating process:

$$Y_{it}^A = \delta_t + \mu_t'\theta_i + \epsilon_{it},\ \mathbb{E}_{it}[\epsilon_{it} \mid \delta_t, \mu_t, \theta_i, D_{it}] = 0 \tag{4}$$

where δ_t, μ_t are time latent factors, θ_i are unit latent factors, ϵ_{it} is a disturbance that's mean independent of these latent factors and D_{it}, and D_{it} is a dummy for whether unit i's treatment status changes at t (and remains changed onwards during the period in question). The synthetic controls estimator assumes that so long as a unit's treatment status doesn't change significantly during this period, the outcome is given by this model.

2 When to use

Synthetic controls are useful to estimate the effects of aggregate interventions in comparative case studies. In many settings, micro-interventions are not feasible for instance because there will be spill-over effects between units at the micro-level that make potential outcomes undefined. For instance, we could consider an intervention where we offer money to some individuals in a town and we wish to see how their financial situation changes relative to if we hadn't offered them money. The financial outcomes of those that don't receive money in the town cannot be the counterfactual outcome for the receiving individuals, had they not received money, because the non-receivers might be impacted by the intervention status of the receivers (eg., the receivers might employ them). As a result, we might wish to run this experiment at city levels where we consider offering money to some individuals in one town but not in other towns. The question then becomes what city can serve as a counterfactual for the intervened city in the post-treatment periods? The synthetic controls estimator posits that a convex average of non-treated cities can do a better job forming the counterfactual outcome than any city alone. As another application of synthetic controls, there are settings where we might just have data at the aggregate level and this aggregate study is the best we can. To list some contextual requirements, synthetic controls assume there's no interference between treated units, the volatility in the outcome is small relative to the intervention effect, there exists a comparison group, there's no anticipation of treatment (though this can be solved by backdating), and the non-intervention counterfactual sequence of the treated unit lies in the convex hull of the outcome sequence of the

[4] As a remark, the validity of this permutation test is econometrically clear to me in a linear factor model of Equation (4) if $\epsilon_{it}|\delta_t, \mu_t, \theta_i \overset{iid}{\sim} (0, \sigma_t^2)$ as then intuitively one can permute the treatment assignment and produce a sound null distribution of no effect. Under unit heteroskedasticity, the validity of the permutation seems less clear to me though it's possible that the test statistic r_j, as constructed to divide by the pre-treatment fit, allows for this. If there's any econometrics that argue for the validity of this test under a more general error structure I'd be curious to learn about it.

non-treated units. Finally, on this topic, we could consider synthetic controls estimators over differences-and-differences strategies when we don't necessarily believe parallel trends hold and there are few treated individuals.

3 Watch out

We'd like to briefly discuss one non-headline but important point to be wary of when using synthetic controls estimators. One of the main things one should be wary of is not overfitting to the pre-intervention trends for unit $j = 0$ with w^*. We don't wish to track noise or volatility in the outcome process. To help with not overfitting, we can consider using less untreated units in the "donor" pool. Moreover, to check that we're not overfitting, we can consider picking w^* and V using not all periods before T_0 (assuming we have sufficient data) and checking our fit on the remaining periods: we keep some extra pre-intervention periods to validate our ability to tract the untreated sequence of outcomes for the intervened unit pre-intervention. We can also consider using the penalized synthetic controls estimator of Abadie and L'Hour (2021).

References

Alberto Abadie. Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, 59(2):391–425, June 2021. URL: `https://www.aeaweb.org/articles?id=10.1257/jel.20191450`, `doi:10.1257/jel.20191450`.

Alberto Abadie and Jaume Vives i Bastida. Synthetic controls in action, 2022. URL: `https://arxiv.org/abs/2203.06279`, `arXiv:2203.06279`.

Alberto Abadie and Jérémy L'Hour. A penalized synthetic control estimator for disaggregated data. *Journal of the American Statistical Association*, 116(536):1817–1834, 2021. `arXiv:https://doi.org/10.1080/01621459.2021.1971535`, `doi:10.1080/01621459.2021.1971535`.

Maximilian Kasy. Synthetic controls. Slide deck, 14.385 Nonlinear Econometric Analysis, Department of Economics, MIT, Fall 2022. Accessed: 2025-12-16. URL: `https://maxkasy.github.io/home/files/teaching/Nonlineareconometrics_MIT_2022/synthetic_control_slides.pdf`.