Synthetic Control As Online Linear Regression

Jiafeng (Kevin) Chen

Harvard Business School

July 28, 2022 | NBER Summer Institute | Labor Studies

Synthetic control

Abadie and Gardeazabal, 2003; Abadie, Diamond, and Hainmueller, 2010; Abadie, Diamond, and Hainmueller, 2015

FIGURE 1 Trends in per Capita GDP: West Germany versus Rest of the OECD Sample

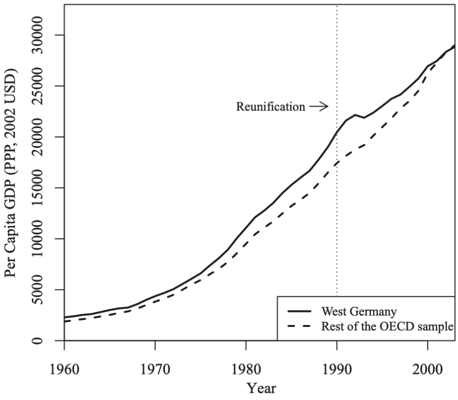
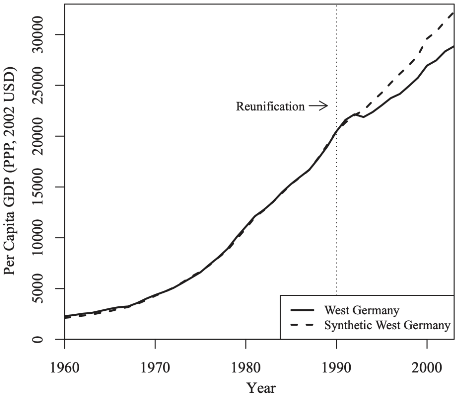


FIGURE 2 Trends in per Capita GDP: West Germany versus Synthetic West Germany



Outcome models and their discontents

- Most statistical guarantees on synthetic control are derived under outcome models
  - → Assume that the untreated potential outcomes $\mathbf{Y} = (Y_{it}(0))$ follow a linear factor model; theoretical results under $T \to \infty$ asymptotic approximation (Abadie, Diamond, and Hainmueller, 2010; Ferman and Pinto, 2021; Ben-Michael, Feller, and Rothstein, 2019; Ben-Michael, Feller, and Rothstein, 2021; Ferman, 2021; Amjad, Shah, and Shen, 2018; Hirshberg, 2021)
- In comparative case study settings, often not obvious how to model the outcomes realistically
  - → What's a realistic sampling thought experiment for US state-year crime rates? (Manski and Pepper, 2018)
- Yet, perhaps the popularity of synthetic control suggests that its appeal goes beyond outcome models
- Can we say anything about synthetic control without relying on outcome modeling?

This paper

- I offer novel guarantees for synthetic control, which do not rely on outcome models
- Over any bounded potential outcomes, on average over time, with large T,
  - → **Result 1**: Synthetic control predictions are never much worse than the predictions made by the best possible weighted matching estimator
  - → **Result 2**: Synthetic control on differenced data never performs much worse than the best possible weighted difference-in-differences estimator
- "On average over time" is averaging with respect to hypothetical treatment timings
  - → Admits a design-based interpretation under random treatment timing (Bottmer, Imbens, Spiess, and Warnick, 2021)
- Key: Thinking of the the panel prediction problem as an online learning problem

Notation

- $\mathbf{Y}$ is the $(N+1) \times T$ matrix of untreated potential outcomes, assumed bounded
- Unit 0 is the treated unit
- $\tau \in [T] \equiv \{1, \dots, T\}$ is a treatment time
- The data analyst wants to predict $y_{0\tau}$ (one-step ahead)
- The control units have outcomes $\mathbf{y}_t$
- Synthetic control algorithm:
  - → Picks weights

$$\theta_\tau = \arg\min_{\theta \in \Theta} \sum_{t \le \tau - 1} (y_{0t} - \theta' \mathbf{y}_t)^2 \qquad \Theta = \left\{ (\theta_1, \dots, \theta_N) \mid \sum_{i=1}^{N} \theta_i = 1, \forall i : \theta_i \ge 0 \right\}$$

  - → Predicts $\hat{y}_{0\tau} = \theta'_\tau \mathbf{y}_\tau$,
  - → Suffers loss $(y_{0\tau} - \hat{y}_{0\tau})^2 = (y_{0\tau} - \theta'_\tau \mathbf{y}_\tau)^2$

Average loss

- Consider the average loss over hypothetical treatment timings, holding fixed the untreated potential outcomes $\mathbf{Y}$:

$$L_T(\mathbf{Y}) = \frac{1}{T}\sum_{\tau=1}^{T}(y_{0,\tau} - \theta_\tau' \mathbf{y}_\tau)^2$$

- Design-based interpretation from random treatment timing

$$L_T(\mathbf{Y}) = \mathbb{E}_{\tau\sim\mathrm{Unif}[T]}[(y_{0,\tau} - \theta_\tau' \mathbf{y}_\tau)^2]$$

Interpretation as online learning

$$L_T(\mathbf{Y}) = \frac{1}{T}\sum_{t=1}^{T}(y_{0,t} - \theta_t'\mathbf{y}_t)^2$$

- Imagine a sequential prediction game
- At each time t, the analyst is prompted for a decision θ_t
- The analyst may pick θ_t based on past data $\mathbf{Y}_s, s < t$
- After time t, $\mathbf{Y}_t$ is revealed to the analyst, and the analyst suffers $(y_{0,t} - \theta_t'\mathbf{y}_t)^2$
- At the end of the game, $L_T(\mathbf{Y})$ is the analyst's average loss
- Synthetic control only looks at pre-treatment data $\implies$

Averaging over treatment timings $\approx$ Online decision making

Online learning

- This latter interpretation as online learning is the key to my results
- In an online learning setup, **Follow-The-Leader** (FTL) is a general class of algorithms that picks decisions (θ_t) by greedily minimizing past loss:

$$\theta_t = \arg\min_{\theta \in \Theta} \sum_{s<t} \ell_s(\theta)$$

- In our case, $\ell_s(\theta) = (y_{0,s} - \theta' \mathbf{y}_s)^2$
- This is exactly what synthetic control does
- Hence, results about FTL $\implies$ results about synthetic control!

Regret control

- No assumptions on $\mathbf{Y} \implies$ No guarantees on $L_T(\mathbf{Y})$
- Consider the difference of L_T against the best fixed alternative

$$\overline{\mathrm{Regret}}_T(\mathbf{Y};\Theta) = L_T(\mathbf{Y}) - \underbrace{\min_{\theta\in\Theta} \frac{1}{T}\sum_{\tau=1}^{T}(y_{0\tau} - \theta'\mathbf{y}_\tau)^2}_{\text{Can only use one } \theta\text{, but knows } \mathbf{Y}}$$

- How well does synthetic control stack up against an oracle weighted matching estimator?

Main result

- **Theorem**. If $\|\mathbf{Y}\|_\infty \leq 1$ and $T > N > 2$, then synthetic control has logarithmic regret:
$$\overline{\text{Regret}}_T(\mathbf{Y}; \Theta) \leq CN\frac{\log T}{T}$$
where C is a universal constant
  - → Immediate from Hazan, Agarwal, and Kale, 2007
- On average over hypothetical treatment timing, synthetic control is never much worse than the best matching estimator, regardless of the potential outcomes
- Caveat: despite a finite-sample result, its usefulness lies in the $\frac{\log T}{T}$ rate in T
  - → Worst-case over all outcomes is (deliberately) conservative
  - → Still useful that a good convergence rate is achievable

Three interpretations of regret control

1. Uniformly over realizations of (bounded) untreated potential outcomes $\mathbf{Y}$,

$$L_T(\mathbf{Y}) \le \underbrace{\min_{\theta\in\Theta} \frac{1}{T}\sum_{\tau=1}^{T}(y_{0\tau} - \theta'\mathbf{y}_\tau)^2}_{\text{best fixed weighted match}} + O\left(\frac{\log T}{T}\right)$$

2. If we accept the design-based interpretation under $\tau \sim \mathrm{Unif}[T]$, then

$$\mathbb{E}_\tau(y_{0\tau} - \theta'_\tau\mathbf{y}_\tau)^2 \le \min_{\theta\in\Theta}\mathbb{E}_\tau(y_{0\tau} - \theta'\mathbf{y}_\tau)^2 + O\left(\frac{\log T}{T}\right)$$

3. Over any distribution of the (bounded) untreated potential outcomes $\mathbf{Y} \sim P$,

$$\mathrm{Risk} = \mathbb{E}_P\mathbb{E}_\tau(y_{0\tau} - \theta'_\tau\mathbf{y}_\tau)^2 \le \mathbb{E}_P\left[\min_{\theta\in\Theta}\mathbb{E}_\tau(y_{0\tau} - \theta'\mathbf{y}_\tau)^2\right] + O\left(\frac{\log T}{T}\right)$$

→ Synthetic control achieves small risk if there exists a weighted match that tracks y_{0t} well

→ Points 2 and 3 require random treatment timing, somewhat relaxed in the paper

Difference-in-differences

- Our main result shows that synthetic control on y_{it} achieves low regret against the best weighted matching estimator
- A similar argument shows that synthetic control on differenced data,
$$\tilde{y}_{it} = y_{it} - \frac{1}{t-1}\sum_{s<t} y_{is},$$
has low regret against the best weighted difference-in-differences estimator
- A weighted difference-in-differences estimator results from a weighted TWFE regression, and turns out to be weighted matching on the differences $\tilde{y}_{it}$

Recap

- I offer novel, uniform-in-outcome guarantees for synthetic control methods by making a connection to online learning
- Synthetic control is an instance of Follow-the-Leader, a class of online learning algorithms with good regret guarantees
- Regardless of outcomes, synthetic control is as good as the best weighted matching estimator, in terms of average performance over hypothetical treatment timings
- Ditto for synthetic control on differenced data and weighted difference-in-differences

References I

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller (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, pp. 493–505.

— (2015). “Comparative politics and the synthetic control method”. *American Journal of Political Science* 59.2, pp. 495–510.

Abadie, Alberto and Javier Gardeazabal (2003). “The economic costs of conflict: A case study of the Basque Country”. *American economic review* 93.1, pp. 113–132.

Amjad, Muhammad, Devavrat Shah, and Dennis Shen (2018). “Robust synthetic control”. *The Journal of Machine Learning Research* 19.1, pp. 802–852.

Ben-Michael, Eli, Avi Feller, and Jesse Rothstein (2019). “Synthetic controls and weighted event studies with staggered adoption”. *arXiv preprint arXiv:1912.03290*.

References II

Ben-Michael, Eli, Avi Feller, and Jesse Rothstein (2021). “The augmented synthetic control method”. *Journal of the American Statistical Association* 116.536, pp. 1789–1803.

Bottmer, Lea et al. (2021). “A Design-Based Perspective on Synthetic Control Methods”. *arXiv preprint arXiv:2101.09398*.

Ferman, Bruno (2021). “On the properties of the synthetic control estimator with many periods and many controls”. *Journal of the American Statistical Association* 116.536, pp. 1764–1772.

Ferman, Bruno and Cristine Pinto (2021). “Synthetic controls with imperfect pretreatment fit”. *Quantitative Economics* 12.4, pp. 1197–1221.

Hazan, Elad, Amit Agarwal, and Satyen Kale (2007). “Logarithmic regret algorithms for online convex optimization”. *Machine Learning* 69.2-3, pp. 169–192.

Hirshberg, David A (2021). “Least squares with error in variables”. *arXiv preprint arXiv:2104.08931*.

References III

Manski, Charles F and John V Pepper (2018). “How do right-to-carry laws affect crime rates? Coping with ambiguity using bounded-variation assumptions”. *Review of Economics and Statistics* 100.2, pp. 232–244.