Part D: Panel Data Methods

D5: Synthetic Control Methods and Factor Models

Kirill Borusyak

ARE 213 Applied Econometrics

UC Berkeley, Fall 2025

D5 outline

1. Synthetic control methods
2. Factor models

Recommended reading: Abadie (2021)

Setting

Consider a complete panel with a non-staggered binary treatment again

	$t = 1$	$\dots$	$t = T_0$	$t = T_0 + 1$	$\dots$	$t = T_0 + T_1 \equiv T$
$i = 1$						
$\dots$						
$i = N_0$						
$i = N_0 + 1$				▨	▨	▨
$\dots$				▨	▨	▨
$i = N_0 + N_1 \equiv N$				▨	▨	▨

Setting (2)

For now, assume $N_1 = T_1 = 1$:

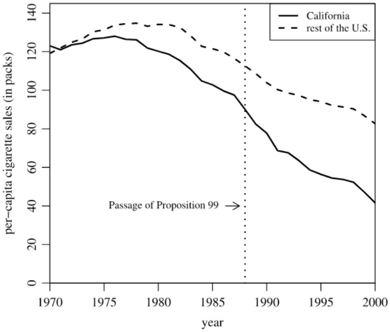
	$t = 1$	$\dots$	$t = T_0$	$t = T$
$i = 1$				
$\dots$				
$i = N_0$				
$i = N$				▨

- We just need to impute $Y_{NT}(0)$ to get $\hat{\tau}_{NT} = Y_{NT} - \widehat{Y_{NT}(0)}$
- We'll come back to inference later

Motivating example: Abadie et al. (2010)

Abadie, Diamond, and Hainmueller (2010) study the effect of California's 1988 tobacco control program (Proposition 99) on cigarette sales per capita

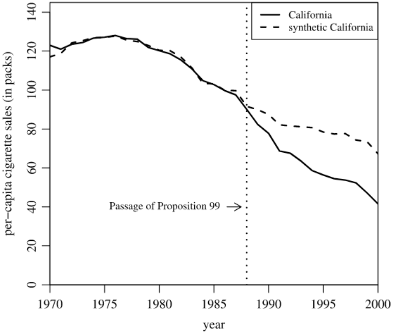
They try **DiD** but it clearly fails:



Motivating example: Abadie et al. (2010)

Abadie, Diamond, and Hainmueller (2010) study the effect of California's 1988 tobacco control program (Proposition 99) on cigarette sales per capita

How about this?



DiD vs. Synthetic control method (SCM)

- DiD uses the simple average of untreated units (“donors”) as the control:

$$\hat{\tau}_{NT} = Y_{NT} - \widehat{Y_{NT}(0)}, \qquad \widehat{Y_{NT}(0)} \equiv \frac{1}{N_0}\sum_{i=1}^{N_0} Y_{iT} + \frac{1}{T_0}\sum_{t=1}^{T_0}\left(Y_{Nt} - \frac{1}{N_0}\sum_{i=1}^{N_0} Y_{it}\right)$$

- What if we found a weighted average of donors that closely traced the pre-treatment path of Y_{Nt}: a “synthetic control” unit?
    - If the same relationship continues into $t = T$, can use $\widehat{Y_{NT}(0)} = \sum_{i=1}^{N_0} \omega_{iT} Y_{iT}$
    - Avoid manually picking comparable states (think Card and Krueger (1994))

How does it work?

For some $\{v_t\}$, we choose weights ω_i on donors to solve

$$\min_{\omega_1,\dots,\omega_{N_0}} \sum_{t=1}^{T_0} v_t \cdot \left(Y_{Nt} - \sum_{i=1}^{N_0} \omega_i Y_{it} \right)^2 \qquad \text{s.t. } \omega_i \geq 0, \quad \sum_{i=1}^{N_0} \omega_i = 1$$

- "Simplex constraints" produce a well-defined average and avoid extrapolation
- They are also a form of regularization
  - Otherwise, a **"vertical" regression** of Y_{Nt} on $Y_{1t}, \dots, Y_{N_0 t}$ across $t = 1, \dots, T_0$
    - ⋆ With $N_0 > T_0$ there would be ∞ ways to fit Y_{Nt} in pre-periods perfectly
  - And no reason to get good $\widehat{Y_{NT}(0)}$ — overfitting
  - With the constraints, *typically* get a unique, sparse solution: few non-zero weights
- Sparsity makes the counterfactual transparent

Synthetic California (Abadie et al., 2010)

State	Weight
Utah	0.334
Nevada	0.234
Montana	0.199
Colorado	0.164
Connecticut	0.069
Other 33 states	0

Details

1. Besides pre-period outcomes, can match on any predetermined predictors: $X_i = (Y_{i1}, \ldots, Y_{iT_0}, Z_i)$

2. How to pick weights on predictors, v?

   - Inverse variance of the predictor across all units

   - Or cross-validation:

     - ⋆ Choose $t_0 < T_0$ training periods: $t = 1, \ldots, t_0$

     - ⋆ Search for v to minimize out-of-sample MSE on the validation period $t_0 + 1, \ldots, T_0$

     - ⋆ For estimation, limit the sample to the last t_0 pre-periods & treated period

Details (2)

3. $\{\omega_i\}$ are not unique if the treated unit is in the convex hull of many donors

   - Abadie and L'Hour (2021): try to match with donors that are more similar to the treated unit

     $$\min_{\omega_1,\ldots,\omega_{N_0}} \sum_t \left\{ \left(Y_{Nt} - \sum_{i=1}^{N_0} \omega_i Y_{it} \right)^2 + \lambda \sum_{i=1}^{N_0} \omega_i \left(Y_{Nt} - Y_{it} \right)^2 \right\}$$

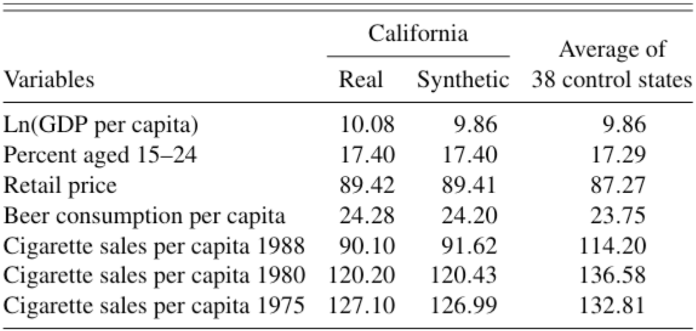
     s.t. $\omega_i \geq 0$, $\sum_{i=1}^{N_0} \omega_i = 1$, with penalty $\lambda > 0$

     - ★ Restores uniqueness and reduces interpolation bias

   - Robbins et al. (2017): pick weights that minimize the distance from equal weights

Abadie et al. (2010) example (cont.)

As predictors, use several observables (averaged in 1980–88) and outcome in three pre-periods



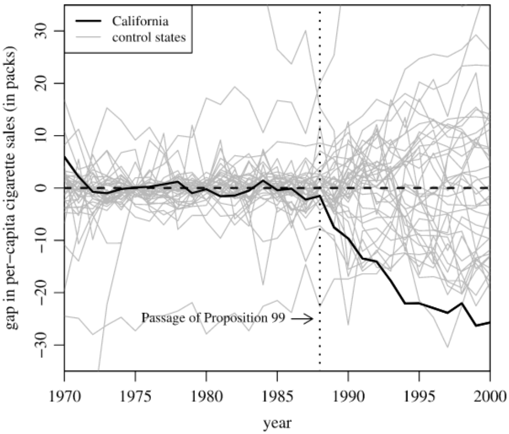
	California		Average of
Variables	Real	Synthetic	38 control states
Ln(GDP per capita)	10.08	9.86	9.86
Percent aged 15–24	17.40	17.40	17.29
Retail price	89.42	89.41	87.27
Beer consumption per capita	24.28	24.20	23.75
Cigarette sales per capita 1988	90.10	91.62	114.20
Cigarette sales per capita 1980	120.20	120.43	136.58
Cigarette sales per capita 1975	127.10	126.99	132.81

- *Note:* Drop states that adopted other tobacco restrictions during the sample period

Inference

- Inference is difficult with only 1 treated unit
- Abadie et al. (2010) "spaghetti plot": randomization inference
  - Under the null of zero effect, the treated unit is no different than others
  - For each i (including $i = N$), construct synthetic i and compute prediction error: $R_i = \sum_{t > T_0} \left(Y_{it} - \widehat{Y_{it}(0)} \right)^2$
  - Reject the null if R_N is extreme in the set of $R_1, \ldots, R_N$; p-value $= \frac{1}{N} \sum_{i=1}^{N} \mathbf{1} \left[R_N \geq R_i \right]$

Inference in Abadie et al. (2010): "Spaghetti plot"



Inference details

- Complication 1: for some i, synthetic control may not match the pre-treatment trajectory well
  - Can drop spaghettis with high pre-treatment MSE
  - Or use $R_i =$ post-treatment MSE divided by pre-treatment MSE
- Complication 2: how can we get a confidence interval?
  - Test inversion: for each τ_0 test $\tau = \tau_0$; collect all τ_0 that are not rejected
  - To test $\tau = \tau_0$, replace Y_{NT} with $Y_{NT} - \tau_0$ and test $\tau = 0$

Inference details (2)

- Problem: randomization inference is not valid without randomization
  - Test has a 5% significance level in the sense that for 5% of units if they were treated the correct null would be rejected
    - Not very useful since the treated unit was not chosen randomly with equal probabilities
  - Abadie, Diamond, and Hainmueller (2015) study the economic impacts of German reunification on West Germany
    - What does assigning this treatment to another country mean?
- One alternative (Chernozhukov et al. (2021)): “conformal inference” based on permuting *periods* instead of units

Diagnostic testing and robustness

- Placebo test: **“backdating”**

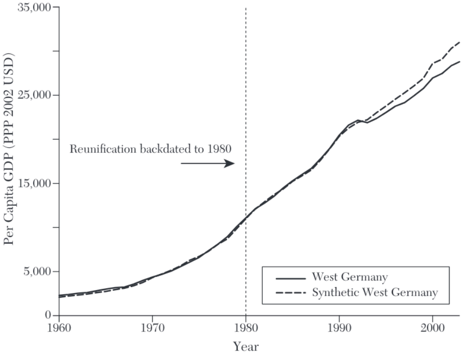


Figure 3. Backdating the 1990 German Reunification Application

(From Abadie (2021))

- Robustness check: dropping each contributing donor at a time

Some extensions

- Everything works with multiple treated periods
- Multiple treated units:
  - Create a synthetic control for each treated unit and estimate the effects, then average (Abadie and L'Hour (2021))
  - Or a single synthetic control for the average of treated units (Robbins et al. (2017))
- Sun, Ben-Michael, and Feller (2025): construct a single synthetic control for multiple outcome variables
- Ben-Michael, Feller, and Rothstein (2021): Augmented SCM estimators, *augsynth*
  - Adjust the estimator for mismatches in the pre-periods for double-robustness

SCM: When and why?

In which contexts should one use synthetic controls?

- Abadie (2021):
    - No anticipation, no spillovers
    - Donor units shouldn't experience large idiosyncratic shocks in the post-periods
    - Big effects and low-volatility outcomes (when few treated units)
    - When a good synthetic control exists
- But which outcome models imply existence of good synthetic controls?
    - And is SCM the best for those models?
    - SCM is usually motivated by factor models (although see Arkhangelsky and Hirshberg (2023))

Outline

1. Synthetic control methods

2. Factor models

Factor model

Factor (a.k.a. latent factor, interactive fixed-effect) model:

$$Y_{it}(0) = \alpha_i + \beta_t + L_{it} + \varepsilon_{it}, \qquad L_{it} = S_i' F_t \equiv \sum_{r=1}^{R} S_{ir} F_{tr}, \qquad \mathbb{E}\left[\varepsilon_{it}\right] = 0$$

with an unknown small number R of unobserved time-varying factors F_t and unobserved unit sensitivities (factor loadings) S_i

- $\alpha_i = \alpha_i \cdot 1$; $\beta_t = 1 \cdot \beta_t$;   unit-specific trend $= \gamma_i \cdot t$
- E.g. $F_t =$ (National minimum wage, Dummy of Democratic administration), $S_i =$ (Share of min.wage occupations in state i, Democratic majority)
- Viewing S_{ir} and F_{tr} as unknown parameters, this is a nonlinear model
- Treatment D_{it} can be correlated with L_{it} (but not with ε_{it})

Factor models vs. SCM

$$Y_{it}(0) = \alpha_i + \beta_t + L_{it} + \varepsilon_{it}, \qquad L_{it} = \sum_{r=1}^{R} S_{ir} F_{tr}, \qquad \mathbb{E}\left[\varepsilon_{it}\right] = 0$$

Why synthetic controls if you have a factor model?

- $Y_{Nt} \approx \sum_{i=1}^{N_0} \omega_i Y_{it}$ for many pre-periods t if $S_{rN} \approx \sum_{i=1}^{N_0} \omega_i S_{ri}$ for each factor r (and for α_i)
- Then can recover $\hat{\alpha}_i + \hat{\beta}_t + \hat{L}_{NT} \equiv \widehat{Y_{NT}(0)}$ with SCM... but there are other methods

Strategy #1: Synthetic DiD (Arkhangelsky et al. (2021))

- Consider a non-staggered case
- DiD puts equal weight on all untreated units and all pre-periods
- But control units similar to the treated ones in the pre-periods are more useful
  - Must have factor loadings similar to $i = N$ (or to average of $N_0 + 1, \ldots, N$)
- Pre-periods similar to the post-period on untreated outcomes are more useful
  - Must have factors similar to $t = T$ (or to average of $T_0 + 1, \ldots, T$)

Strategy #1: Synthetic DiD (Arkhangelsky et al. (2021))

- Obtain weights $v_1, \ldots, v_{T_0}$ and $\omega_1, \ldots \omega_{N_0}$ that add up to one; then

$$\widehat{ATT} = \overline{Y}_{\text{treated,post}} - \sum_{i=1}^{N_0} \hat{\omega}_i \overline{Y}_{i,\text{post}} - \sum_{t=1}^{T_0} \hat{v}_t \left(\overline{Y}_{\text{treated},t} - \sum_{i=1}^{N_0} \hat{\omega}_i Y_{it} \right)$$

  - Estimate via TWFE regression weighted by $\hat{v}_t \cdot \hat{\omega}_i$ (with $\hat{v}_{T_0+1} = \ldots = \hat{v}_T = 1$ and $\hat{\omega}_{N_0+1} = \ldots = \hat{\omega}_N = 1$)
  - How to choose $\hat{v}_t$ and $\hat{\omega}_i$?

Choosing $\hat{v}_t$ and $\hat{\omega}_i$

- Choose $\hat{v}_t$ to make $\overline{Y}_{i,\text{post}}$ close to $\beta_{\text{post}} + \sum_{t=1}^{T_0} v_t Y_{it}$ across untreated units
  - β_{post} captures period FEs
  - Horizontal regression of $\overline{Y}_{i,\text{post}}$ on a constant and $Y_{i1}, \ldots, Y_{iT_0}$ with restrictions $\sum_{t=1}^{T_0} v_t = 1$ and $v_t \geq 0$
- Choose $\hat{\omega}_i$ to make $\overline{Y}_{\text{treated},t}$ close to $\alpha_{\text{treated}} + \sum_{i=1}^{N_0} \omega_i Y_{it}$ across periods
  - α_{treated} captures unit FEs
  - Vertical regression; because $N_0 > T_0$, regularize (using ridge)

$$\min_{\alpha_{\text{treated}},\{\omega_i\}} \frac{1}{T_0} \sum_{t=1}^{T_0} \left(\overline{Y}_{\text{treated},t} - \alpha_{\text{treated}} - \sum_{i=1}^{N_0} \omega_i Y_{it} \right)^2 + \lambda \sum_{i=1}^{N_0} \omega_i^2$$

s.t. $\sum_{i=1}^{N_0} \omega_i = 1, \quad \omega_i \geq 0$ (See paper for choice of penalty λ)

Strategy #2: Imputation with factor models

- Recall $Y_{it}(0) = \alpha_i + \beta_t + L_{it} + \varepsilon_{it}, \quad L_{it} = \sum_{r=1}^{R} S_{ir} F_{tr}$
- SCM and SDiD balance out latent factors without estimating them
- Alternatively, can estimate L_{it} from untreated observations and use them for $\widehat{Y_{it}(0)}$
- Can estimate factors, loadings, and R using principle component analysis (Bai and Ng (2002))
- Another idea: estimate the L matrix directly
    - Key property: $\text{rank}(L) = R$ is small

Athey et al. (2021) matrix completion approach

Athey, Bayati, Doudchenko, Imbens, and Khosravi (2021): **matrix completion**

L_{it}	$t=1$	$\dots$	$t=T_0$	$t=T_0+1$	$\dots$	$t=T_0+T_1\equiv T$
$i=1$						
$\dots$						
$i=N_0$						
$i=N_0+1$				?	?	?
$\dots$				?	?	?
$i=N_0+N_1\equiv N$				?	?	?

- Recover L from:
  - observing Y_{it} for untreated observations = noisy version of $\alpha_i+\beta_t+L_{it}$
  - knowing that L has low rank

Athey et al. (2021) matrix completion approach

Athey et al. (2021) solve:

$$\min_{\{\alpha_i\},\{\beta_t\},L} \sum_{it:\ D_{it}=0} \left(Y_{it} - \alpha_i - \beta_t - L_{it}\right)^2 + \lambda \left\| L \right\|_*$$

- $\left\| L \right\|_*$ is the "nuclear" norm of matrix L: small when L is close to some low-rank matrix
- Computationally efficient (unlike searching among low-rank matrices)
- Then take average of $Y_{it} - \hat{\alpha}_i - \hat{\beta}_t - \hat{L}_{it}$ among treated observations
- No results on inference

References I

Abadie, A. (2021): "Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects," *Journal of Economic Literature*, 59, 391–425.

Abadie, A., A. Diamond, and J. 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, 493–505, arXiv: 1011.1669v3 ISBN: 0162-1459.

——— (2015): "Comparative Politics and the Synthetic Control Method," *American Journal of Political Science*, 59, 495–510, iSBN: 1540-5907.

Abadie, A. and J. L'Hour (2021): "A Penalized Synthetic Control Estimator for Disaggregated Data," *Journal of the American Statistical Association*, 116, 1817–1834.

Arkhangelsky, D., S. Athey, D. A. Hirshberg, G. W. Imbens, and S. Wager (2021): "Synthetic Difference-in-Differences," *American Economic Review*, 111, 4088–4118.

Arkhangelsky, D. and D. Hirshberg (2023): "Large-Sample Properties of the Synthetic Control Method under Selection on Unobservables," ArXiv:2311.13575 [econ].

References II

Athey, S., M. Bayati, N. Doudchenko, G. W. Imbens, and K. Khosravi (2021): "Matrix Completion Methods for Causal Panel Data Models," *Journal of the American Statistical Association*, 116, 1716–1730, arXiv: 1710.10251.

Bai, J. and S. Ng (2002): "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70, 191–221.

Ben-Michael, E., A. Feller, and J. Rothstein (2021): "The Augmented Synthetic Control Method," *Journal of the American Statistical Association*, 116, 1789–1803.

Card, D. and A. Krueger (1994): "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersry and Pennsylvania," *The American Economic Review*, 84, 772–793.

Chernozhukov, V., K. Wüthrich, and Y. Zhu (2021): "An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls," *Journal of the American Statistical Association*, 116, 1849–1864.

Robbins, M. W., J. Saunders, and B. Kilmer (2017): "A Framework for Synthetic Control Methods With High-Dimensional, Micro-Level Data: Evaluating a Neighborhood-Specific Crime Intervention," *Journal of the American Statistical Association*, 112, 109–126.

References III

Sun, L., E. Ben-Michael, and A. Feller (2025): "Using Multiple Outcomes to Improve the Synthetic Control Method," *Review of Economics and Statistics*.