Synthetic Controls: Methods and Practice

Alberto Abadie
MIT

Mixtape Sessions
April 27-28, 2023

Introduction

- Synthetic control methods were originally proposed in Abadie and Gardeazabal (2003) and Abadie et al. (2010) with the aim to estimate the effects of aggregate interventions.
- Many events or interventions of interest naturally happen at an aggregate level affecting a small number of large units (such as cities, regions, or countries).
- Even in experimental settings micro-interventions may not be feasible (e.g., fairness) or effective (e.g., interference).

In this talk, I will use the terms "event", "intervention", and "treatment" interchangeably.

Applications

- Synthetic controls have been applied to study the effects of right-to-carry laws (Donohue et al., 2017), legalized prostitution (Cunningham and Shah, 2018), immigration policy (Bohn et al., 2014), corporate political connections (Acemoglu et al., 2016) and many other policy issues.
- They have also been adopted as the main tool for data analysis across different sides of the issues in recent prominent debates on the effects of immigration (Borjas, 2017; Peri and Yasenov, 2017) and minimum wages (Allegretto et al., 2017; Jardim et al., 2017; Neumark and Wascher, 2017; Reich et al., 2017).
- Synthetic controls are also applied outside economics in the social sciences, biomedical disciplines, engineering, etc. (see, e.g., Heersink et al., 2017; Pieters et al., 2017).

Applications

- Outside academia, synthetic controls have found considerable coverage in the popular press (see, e.g., Guo, 2015; Douglas, 2018) and have been widely adopted by multilateral organizations, think tanks, business analytics units, governmental agencies, and consulting firms.
- For example, the synthetic control method plays a prominent role in the official evaluation of the effects of the massive Bill & Melinda Gates Foundation's *Intensive Partnerships for Effective Teaching* program (Gutierrez et al., 2016).
- Widely applied in tech industry.

The Washington Post

Wonkblog

Seriously, here's one amazing math trick to learn what can't be known

THE WALL STREET JOURNAL

REAL TIME ECONOMICS | ECONOMICS

How an Analysis of Basque Terrorism Helps Economists Understand Brexit

A method pioneered by an MIT professor has also been used to estimate the economic effect of a tobacco ban, German reunification, legalization of prostitution and gun rights

Plan for the talk

1. A primer on synthetic control estimation
2. Why use synthetic controls?
3. Contextual requirements
4. Data requirements
5. Robustness and diagnostic checks
6. A penalized synthetic control estimator
7. Synthetic controls for experimental design
8. Closing remarks

Literature is large, and there is much I will not cover ...

- **Matrix/tensor completion:** Amjad, Shah, and Shen, (2018), Agarwal, Shah and Shen (2020), Athey, Bayati, Doudchenko, Imbens, and Khosravi (2021), Bai and Ng (2020)
- **Bias correction:** Abadie and L'Hour (2021), Arkhangelsky, Athey, Hirshberg, Imbens, and Wager, (2019), Ben-Michael, Feller, and Rothstein (2021)
- **Inference:** Cattaneo, Feng, Titiunik (2021), Chernozhukov, Wüthrich, and Zhu (2021), Firpo and Possebom (2018)
- **Functional and distributional outcomes:** Chernozhukov, Wüthrich, and Zhu (2019), Gunsilius (2020)
- **Large-T:** Botosaru and Ferman (2019), Ferman (2021), Li (2020)
- **In bandits:** Chen (2023)

... and many more (and many, many, empirical applications).

A primer on synthetic control estimation

- When the units of analysis are a few aggregate entities, a combination of comparison units (a "synthetic control") often does a better job reproducing the characteristics of a treated unit than any single comparison unit alone.
- A synthetic control is selected as the weighted average of all potential comparison units that best resembles the characteristics of the treated unit(s).

A primer on synthetic control estimation

- Suppose that we observe $J + 1$ units in periods $1, 2, \ldots, T$.
- Unit "one" is exposed to the intervention of interest (that is, "treated") during periods $T_0 + 1, \ldots, T$.
- The remaining J units are an untreated reservoir of potential controls (a "donor pool").
- Let Y_{it}^I be the outcome that would be observed for unit i at time t if unit i is exposed to the intervention in periods $T_0 + 1$ to T.
- Let Y_{it}^N be the outcome that would be observed for unit i at time t in the absence of the intervention.
- We aim to estimate the effect of the intervention on the treated unit,

$$\tau_{1t} = Y_{1t}^I - Y_{1t}^N = Y_{1t} - Y_{1t}^N$$

for $t > T_0$, and Y_{1t} is the outcome for unit one at time t.

A primer on synthetic control estimation

- Let $\boldsymbol{W} = (w_2, \ldots, w_{J+1})'$ with $w_j \geq 0$ for $j = 2, \ldots, J+1$ and $w_2 + \cdots + w_{J+1} = 1$. Each value of $\boldsymbol{W}$ represents a potential synthetic control.
- Let $\boldsymbol{X}_1$ be a $(k \times 1)$ vector of pre-intervention characteristics for the treated unit. Similarly, let $\boldsymbol{X}_0$ be a $(k \times J)$ matrix which contains the same variables for the unaffected units.
- The vector $\boldsymbol{W}^* = (w_2^*, \ldots, w_{J+1}^*)'$ is chosen to minimize $\|\boldsymbol{X}_1 - \boldsymbol{X}_0 \boldsymbol{W}\|$, subject to our weight constraints.
- Let Y_{jt} be the value of the outcome for unit j at time t. For a post-intervention period t (with $t \geq T_0$) the synthetic control estimator is:

$$\widehat{\tau}_{1t} = Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt}.$$

A primer on synthetic control estimation

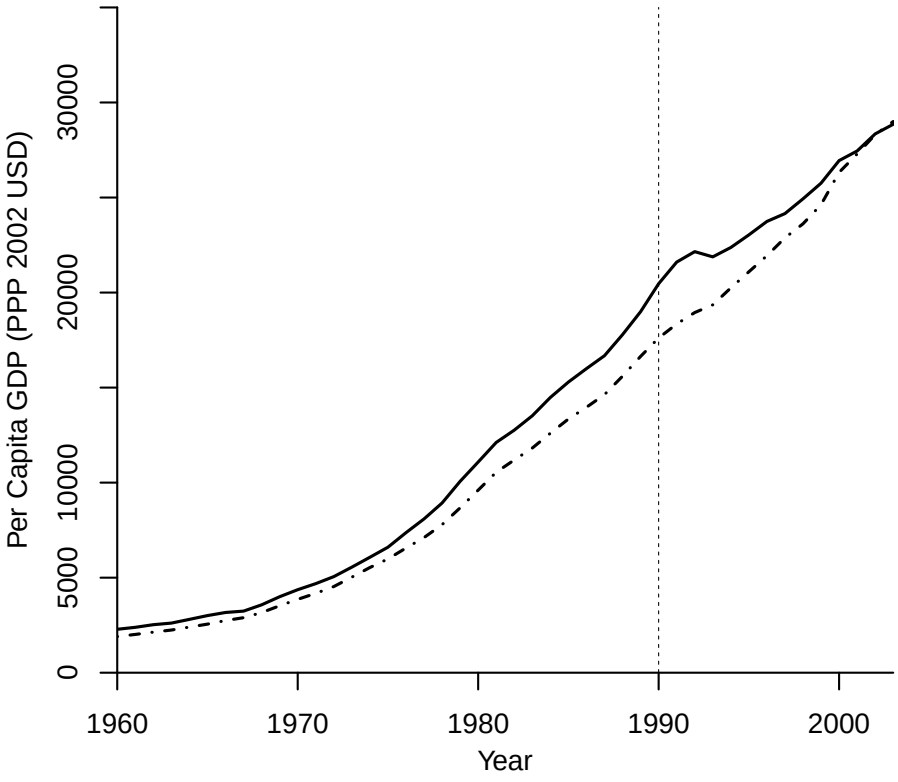
- Typically,

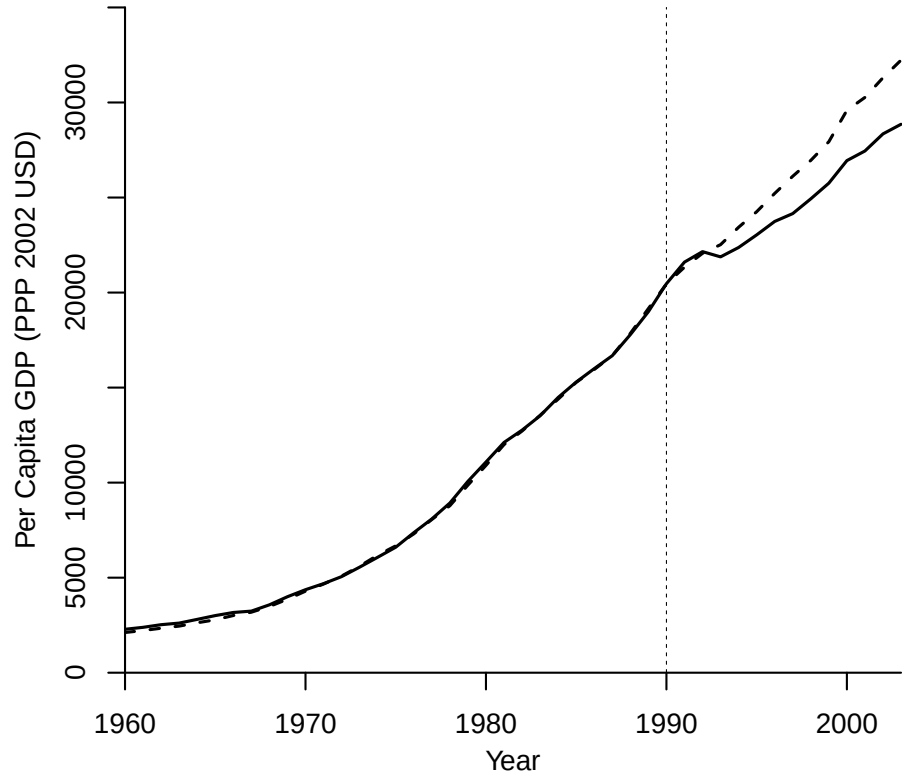
$$\|\boldsymbol{X}_1 - \boldsymbol{X}_0 \boldsymbol{W}\| = \left(\sum_{h=1}^{k} v_h \left(X_{h1} - w_2 X_{h2} - \cdots - w_{J+1} X_{hJ+1} \right)^2 \right)^{1/2}$$

- The positive constants $v_1, \ldots, v_k$ reflect the predictive power of each of the k predictors on Y_{1t}^N.
- $v_1, \ldots, v_k$ can be chosen by the analyst or by data-driven methods.

A primer on synthetic control estimation

Application: German reunification





A primer on synthetic control estimation

Application: German reunification

	West Germany (1)	Synthetic West Germany (2)	OECD Sample (3)
GDP per-capita	15808.9	15802.24	13669.4
Trade openness	56.8	56.9	59.8
Inflation rate	2.6	3.5	7.6
Industry share	34.5	34.5	34.0
Schooling	55.5	55.2	38.7
Investment rate	27.0	27.0	25.9

Note: First column reports $\boldsymbol{X}_1$, second column reports $\boldsymbol{X}_0\boldsymbol{W}^*$, and last column reports a simple average for the 16 OECD countries in the donor pool. GDP per capita, inflation rate, and trade openness are averages for 1981–1990. Industry share (of value added) is the average for 1981–1989. Schooling is the average for 1980 and 1985. Investment rate is averaged over 1980–1984.

A primer on synthetic control estimation

Application: German reunification

country j	W_j^*	country j	W_j^*
Australia	0	Netherlands	0.10
Austria	0.42	New Zealand	0
Belgium	0	Norway	0
Denmark	0	Portugal	0
France	0	Spain	0
Greece	0	Switzerland	0.11
Italy	0	United Kingdom	0
Japan	0.16	United States	0.22

A primer on synthetic control estimation

- Abadie et al. (2010) establish a bias bound under the factor model

$$Y_{it}^N = \boldsymbol{\theta}_t \boldsymbol{Z}_i + \boldsymbol{\lambda}_t \boldsymbol{\mu}_i + \varepsilon_{it},$$

where $\boldsymbol{Z}_i$ are observed features, $\boldsymbol{\mu}_i$ are unobserved features, and ε_{it} is a unit-level transitory shock, modeled as random noise.

- Suppose that we can choose $\boldsymbol{W}^*$ such that:

$$\sum_{j=2}^{J+1} w_j^* \boldsymbol{Z}_j = \boldsymbol{Z}_1, \ \sum_{j=2}^{J+1} w_j^* Y_{j1} = Y_{11}, \ \cdots \ , \ \sum_{j=2}^{J+1} w_j^* Y_{jT_0} = Y_{1T_0}$$

with probability one. In practice, these may hold only approximately.

A primer on synthetic control estimation

Suppose that $E|\varepsilon_{jt}|^p < \infty$ for some $p > 2$. Then,

$$|E[\widehat{\tau}_{1t} - \tau_{1t}]| < C(p)^{1/p} \left(\frac{\bar{\lambda}^2 F}{\underline{\xi}} \right) J^{1/p} \max \left\{ \frac{\bar{m}_p^{1/p}}{T_0^{1-1/p}}, \frac{\bar{\sigma}}{T_0^{1/2}} \right\}$$

where F is the number of unobserved factors,

$$\sigma_{jt}^2 = E|\varepsilon_{jt}|^2, \quad \sigma_j^2 = \frac{1}{T_0} \sum_{t=1}^{T_0} \sigma_{jt}^2, \quad \bar{\sigma}^2 = \max_{j=2,\ldots,J+1} \sigma_j^2,$$

$$m_{pjt} = E|\varepsilon_{jt}|^p, \quad m_{pj} = \frac{1}{T_0} \sum_{t=1}^{T_0} m_{pjt}, \quad \bar{m}_p = \max_{j=2,\ldots,J+1} m_{pj},$$

for p even, $|\lambda_{tf}| \leq \bar{\lambda}$ for all $t = 1, \ldots, T$ and $f = 1, \ldots, F$, and

$$\underline{\xi} \leq \xi(M) = \text{smallest eigenvalue of } \frac{1}{M} \sum_{t=T_0-M+1}^{T_0} \lambda_t' \lambda_t.$$

A primer on synthetic control estimation

- The bias bound is predicated on close fit, and controlled by the ratio between the scale of ε_{it} and T_0.
- In particular, the credibility of a synthetic control depends on the extent to which it is able to fit the trajectory of Y_{1t} for an extended pre-intervention period.

A primer on synthetic control estimation

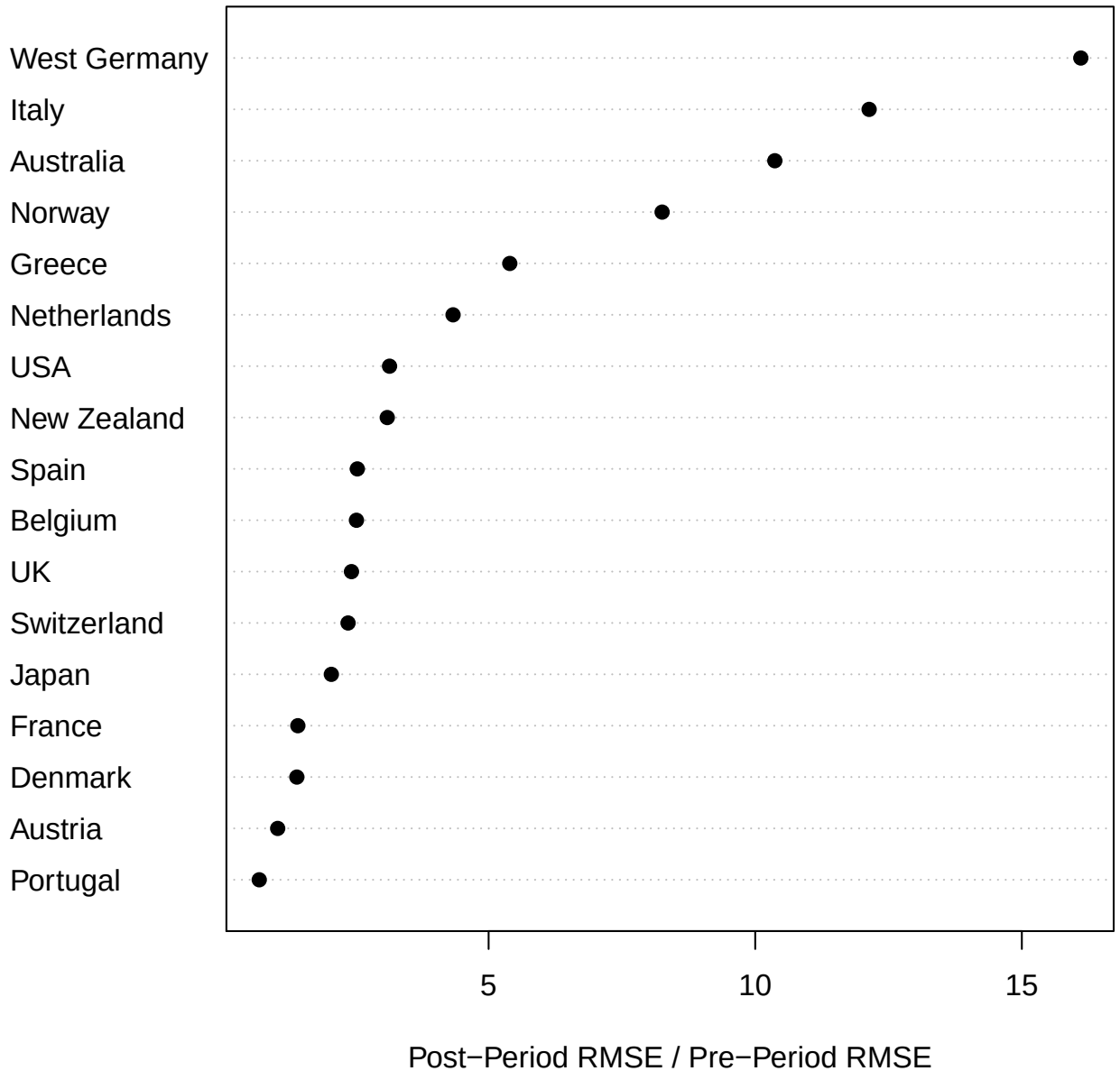
- There are no ex-ante guarantees on the fit. If the fit is poor, Abadie et al. (2010) recommend against the use of synthetic controls.
- Settings with small T_0, large J, and large noise create substantial risk of overfitting.
- To reduce interpolation biases and risk of overfitting, restrict the donor pool to units that are similar to the treated unit.

A primer on synthetic control estimation

- Abadie et al. (2010) propose a mode of inference for the synthetic control framework that is based on permutation methods.
- A permutation distribution can be obtained by iteratively reassigning the treatment to the units in the donor pool and estimating "placebo effects" in each iteration.
- The effect of the treatment on the unit affected by the intervention is deemed to be significant when its magnitude is extreme relative to the permutation distribution.

A primer on synthetic control estimation

Application: German reunification

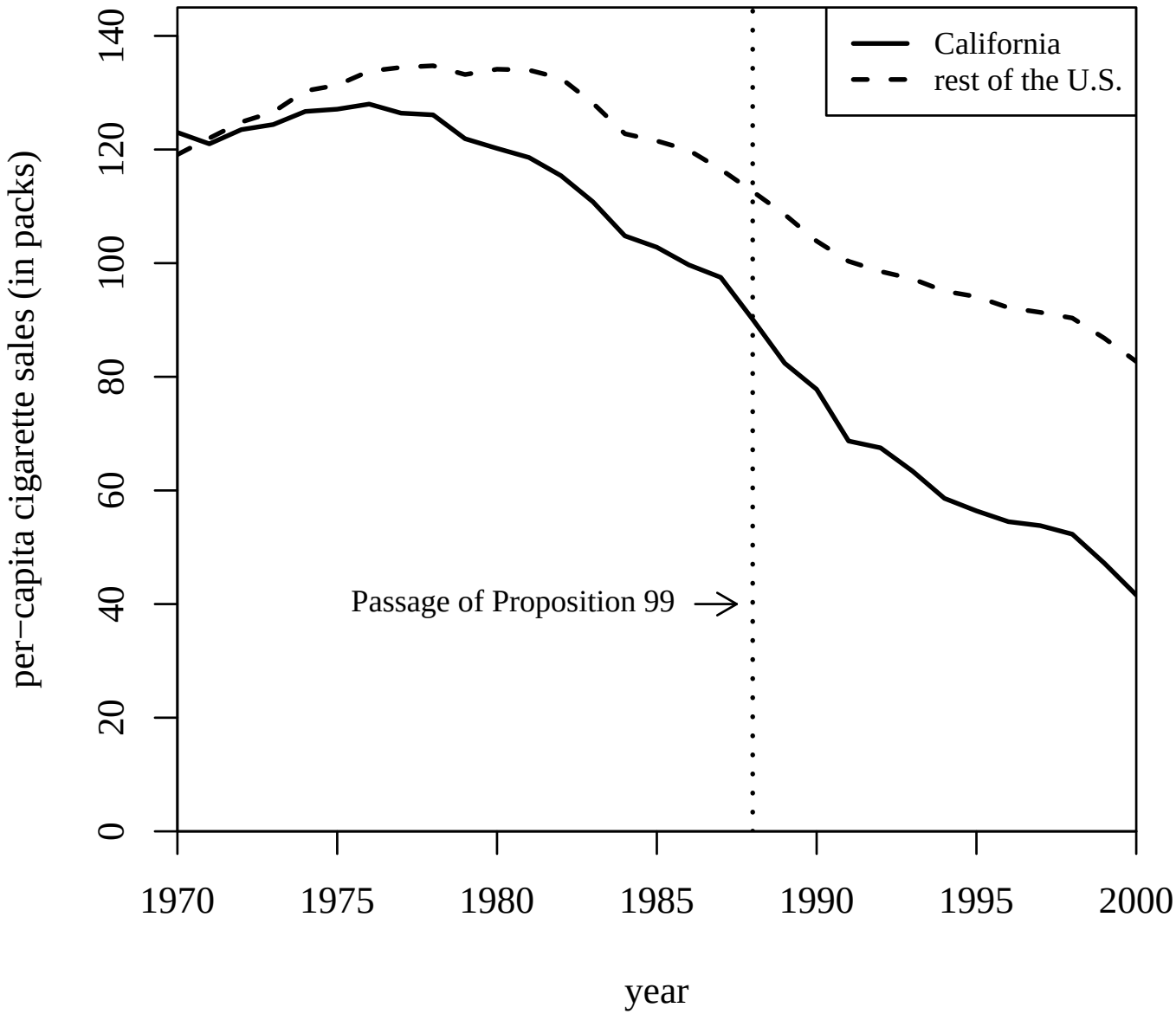


A primer on synthetic control estimation

- The permutation distribution is more informative than mechanically looking at p-values alone.
- Depending on the number of units in the donor pool, conventional significance levels may be unrealistic or impossible.
- Often, one sided inference is most relevant.

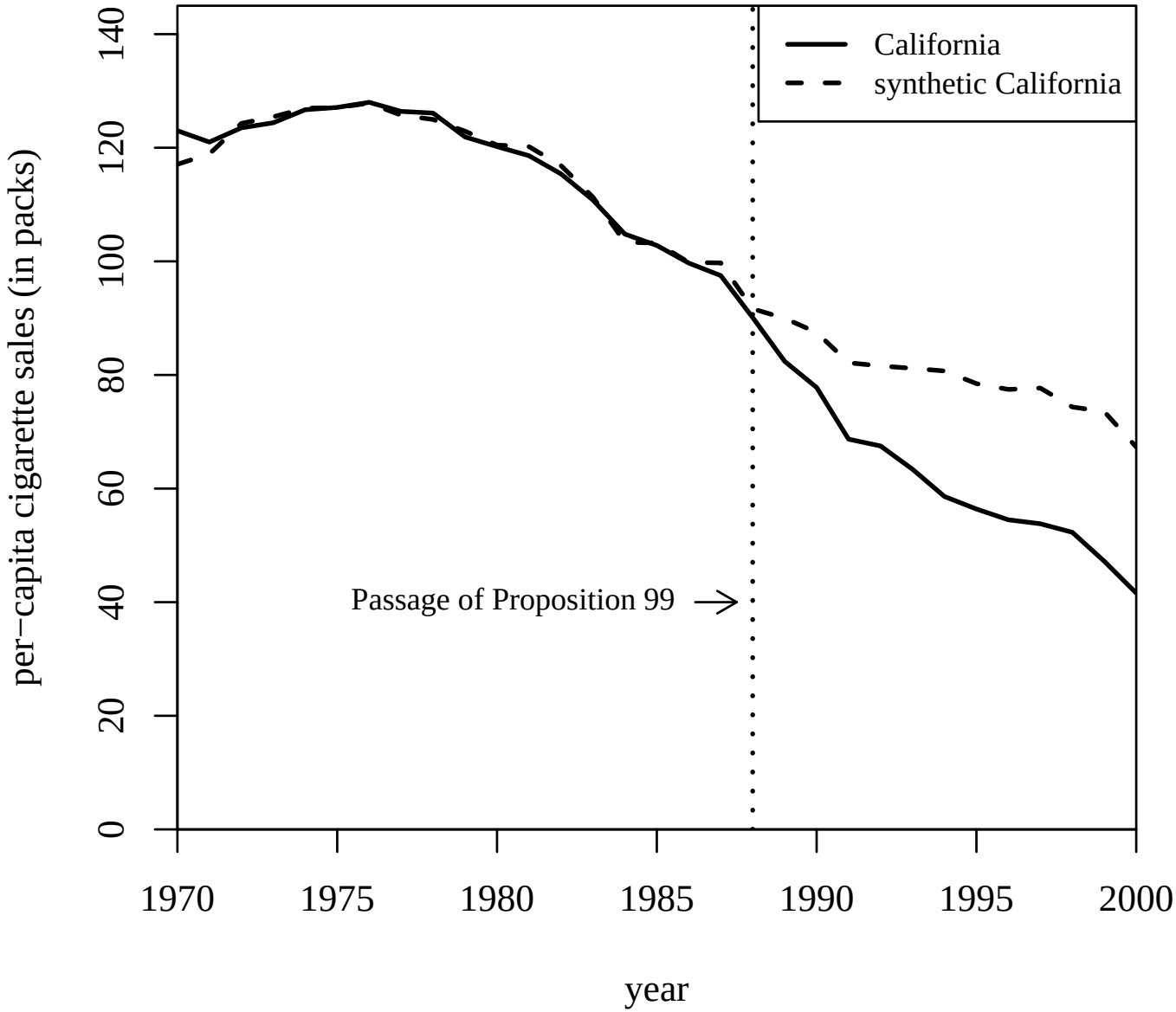
A primer on synthetic control estimation

Application: California tobacco control program



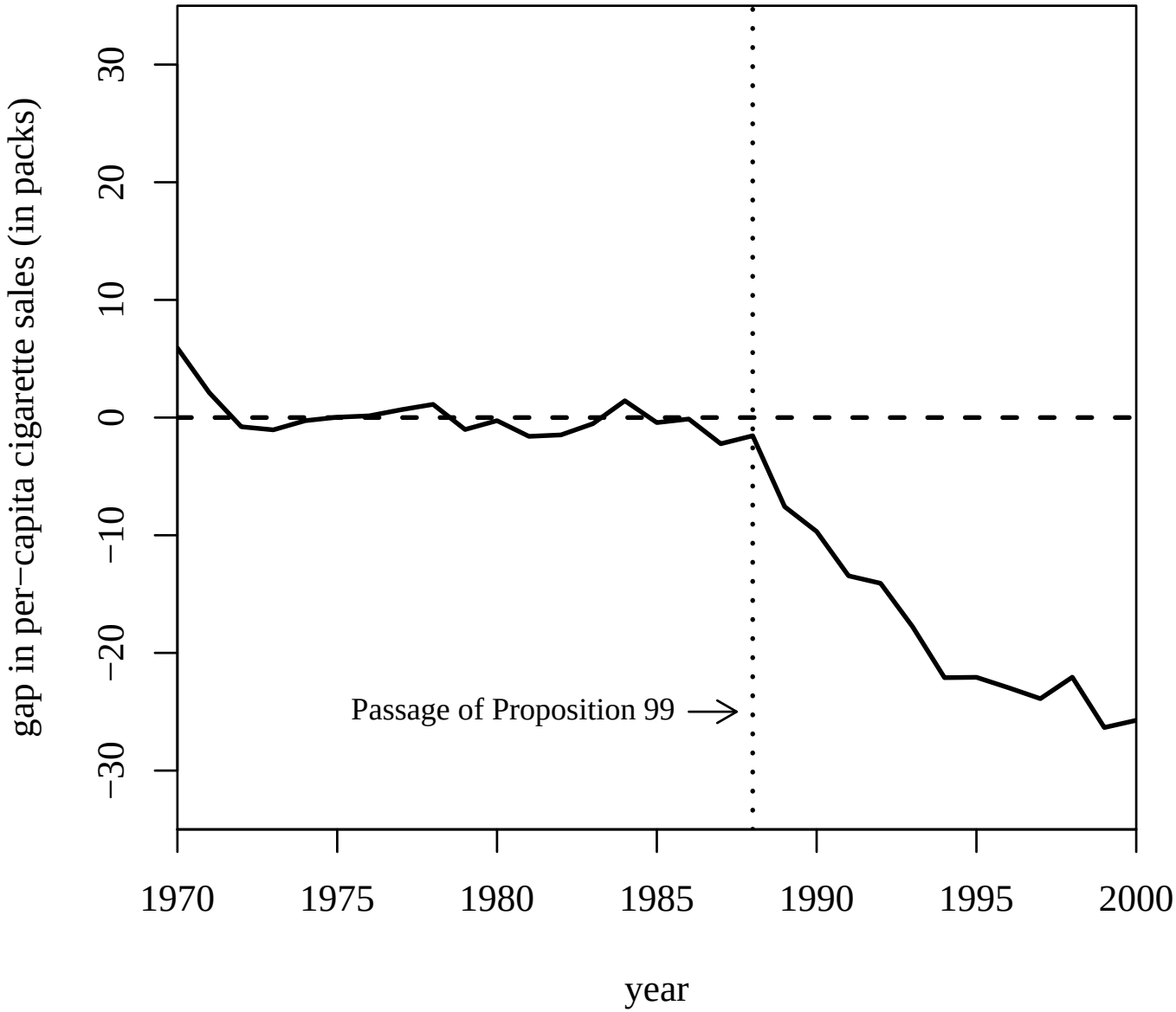
A primer on synthetic control estimation

Application: California tobacco control program



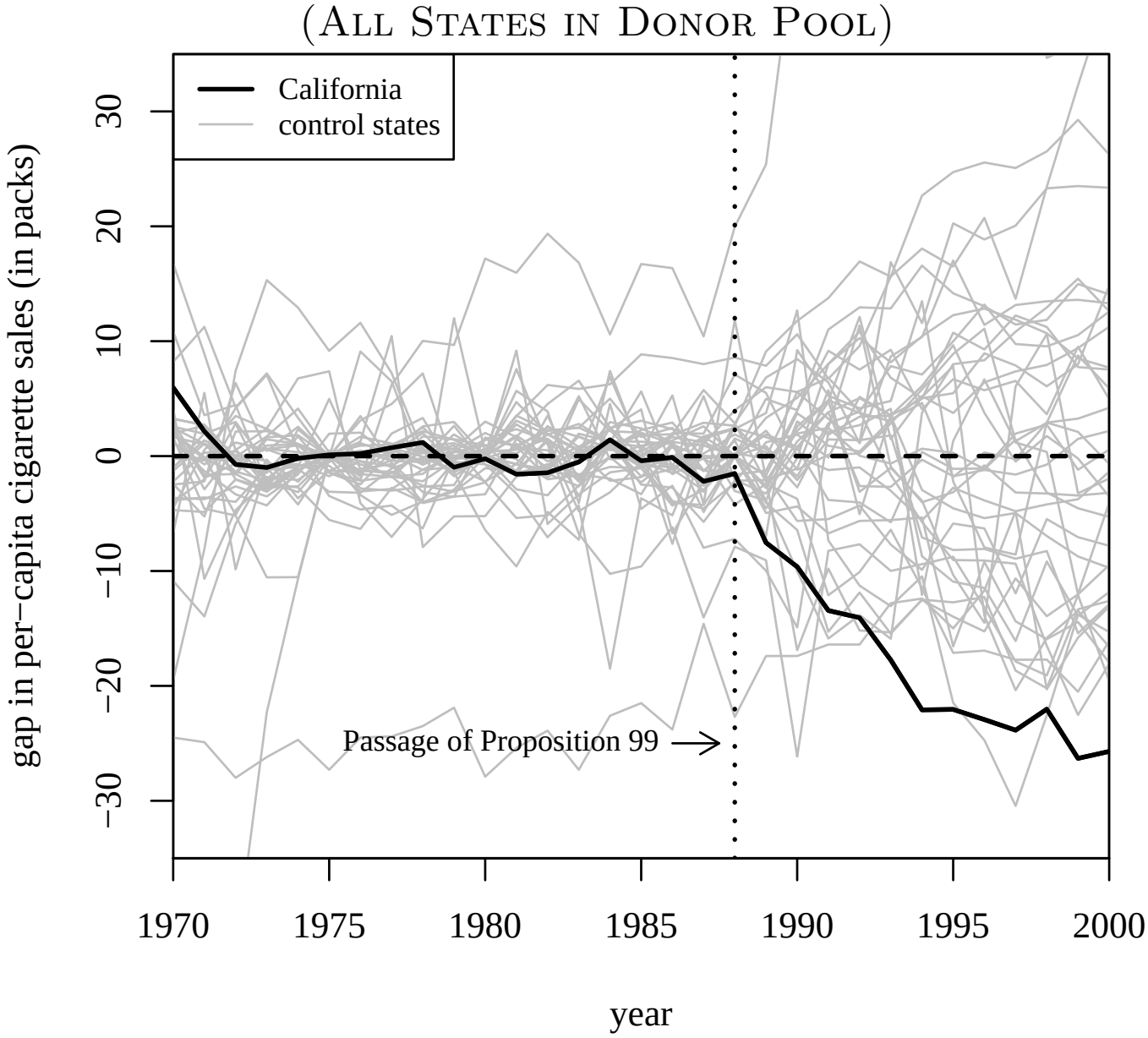
A primer on synthetic control estimation

Application: California tobacco control program



A primer on synthetic control estimation

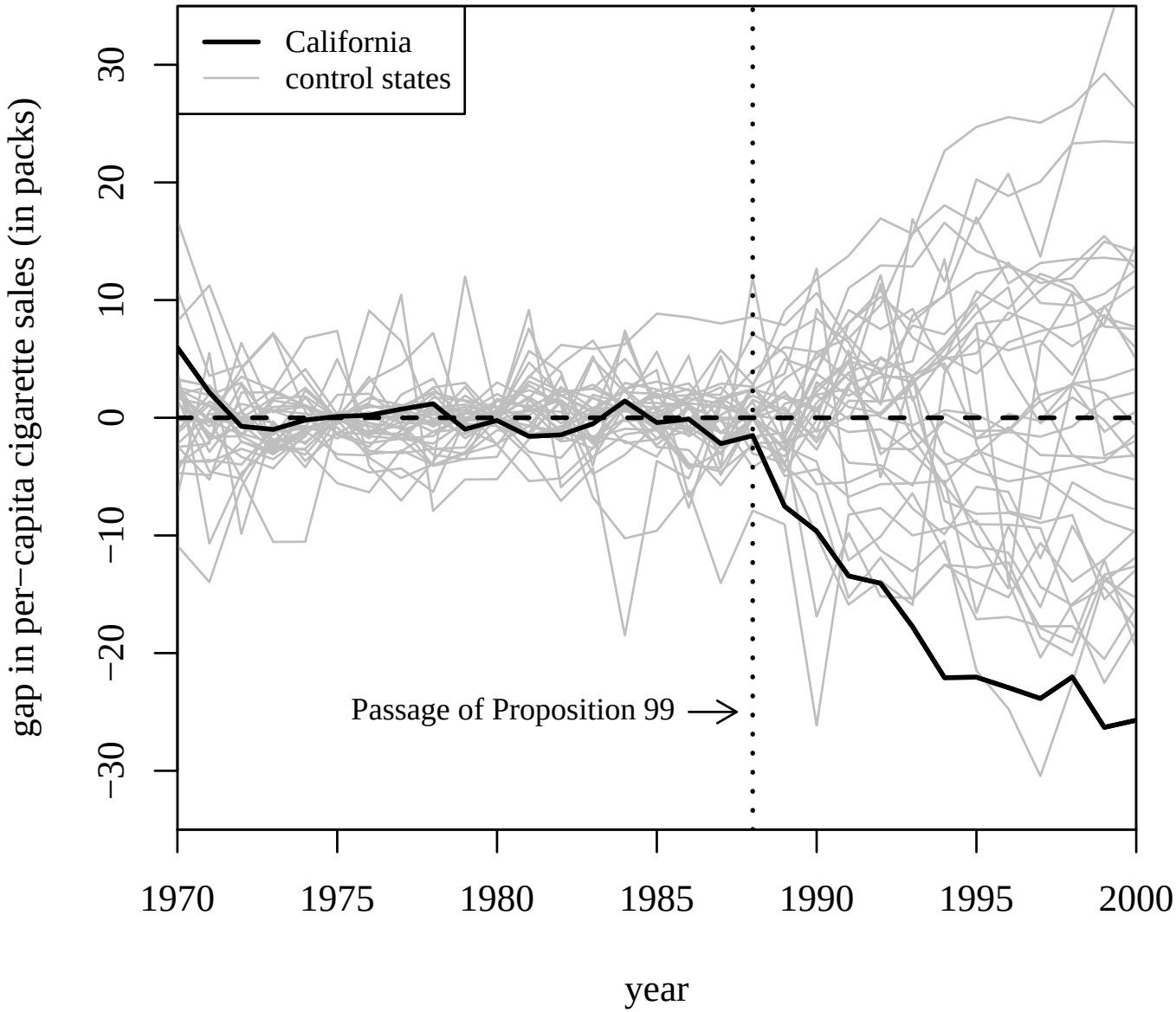
Application: California tobacco control program



A primer on synthetic control estimation

Application: California tobacco control program

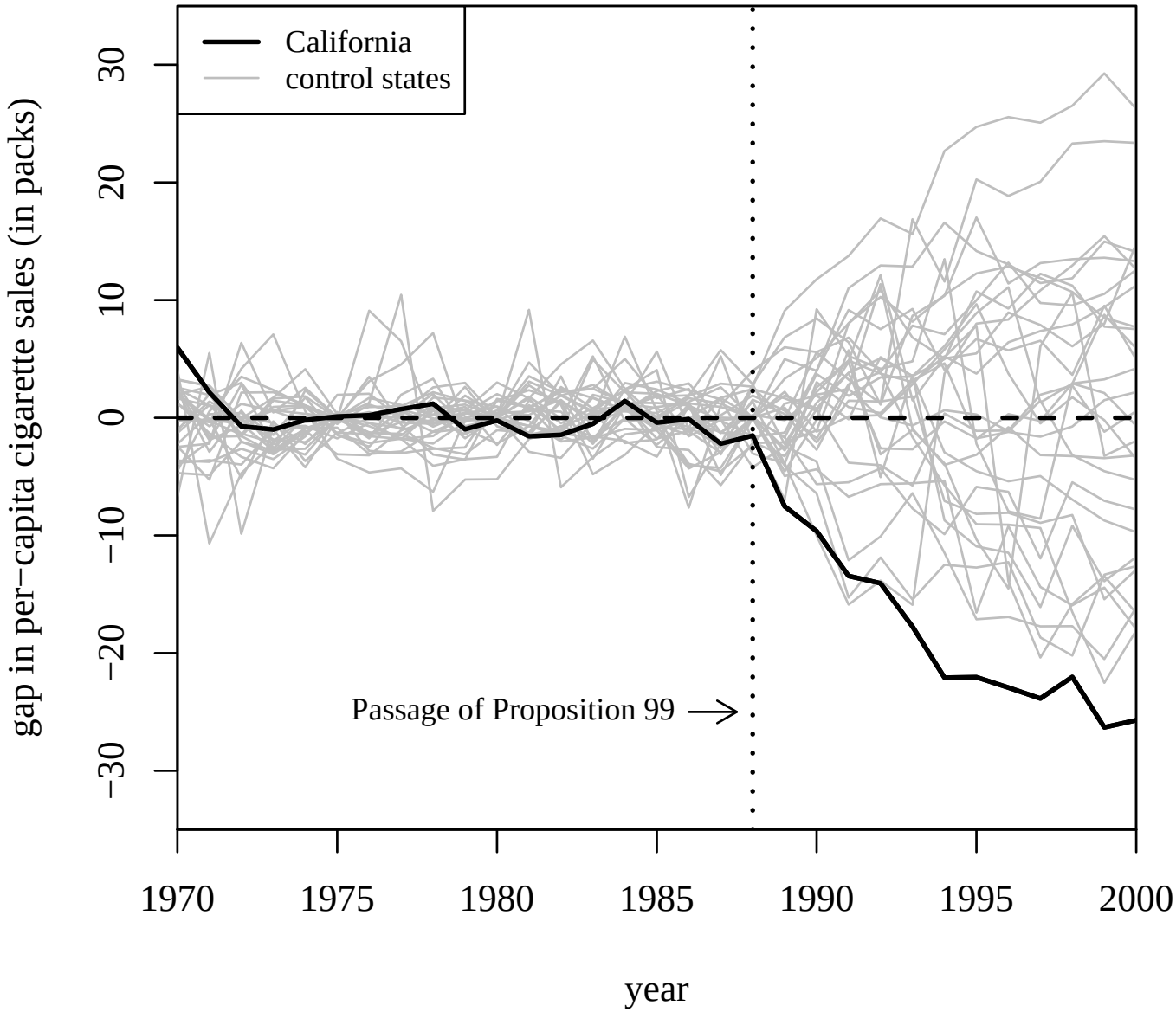
(Pre-Prop. 99 MSPE $\leq$ 20 Times Pre-Prop. 99 MSPE for CA)



A primer on synthetic control estimation

Application: California tobacco control program

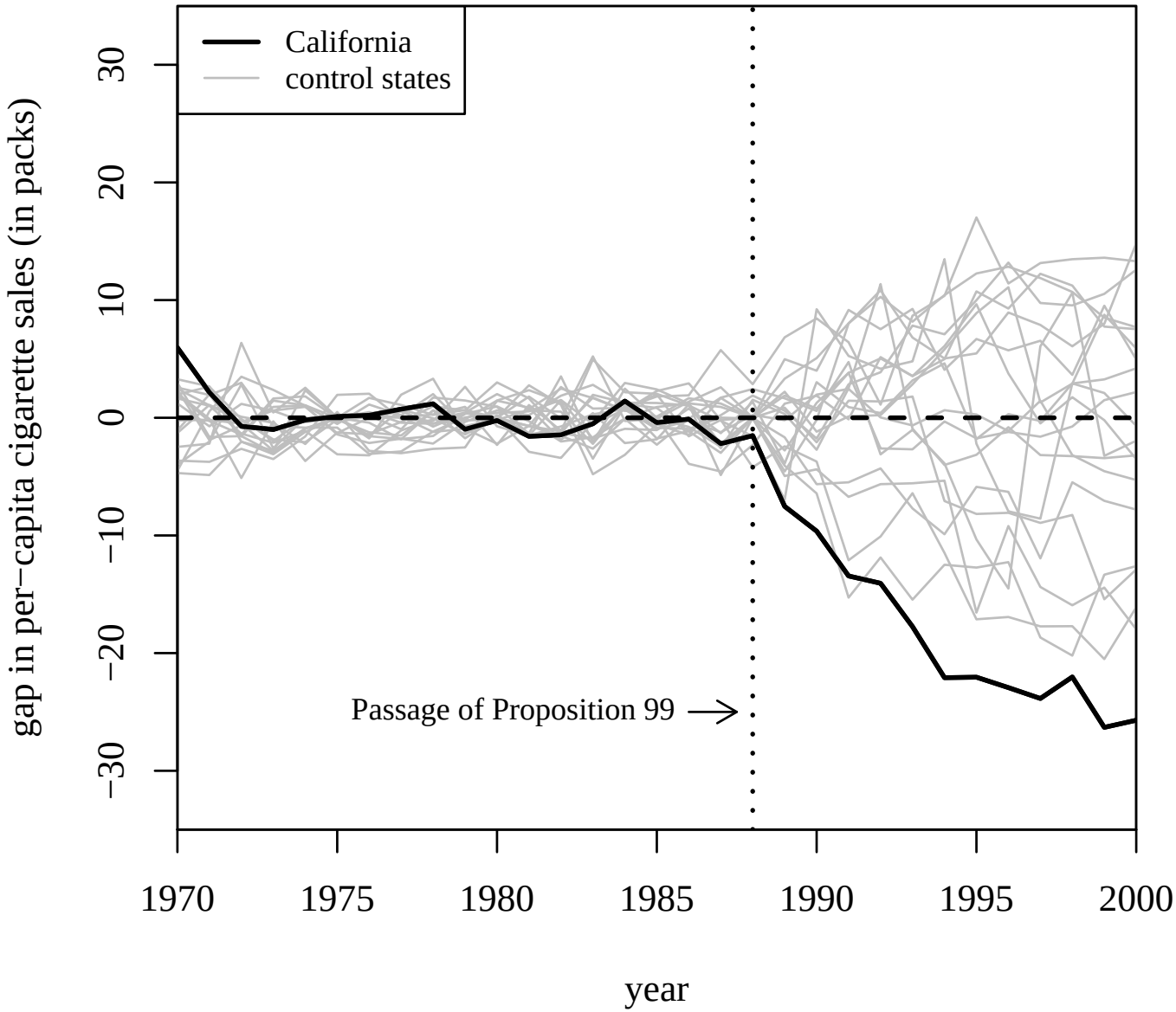
(Pre-Prop. 99 MSPE $\leq$ 5 Times Pre-Prop. 99 MSPE for CA)



A primer on synthetic control estimation

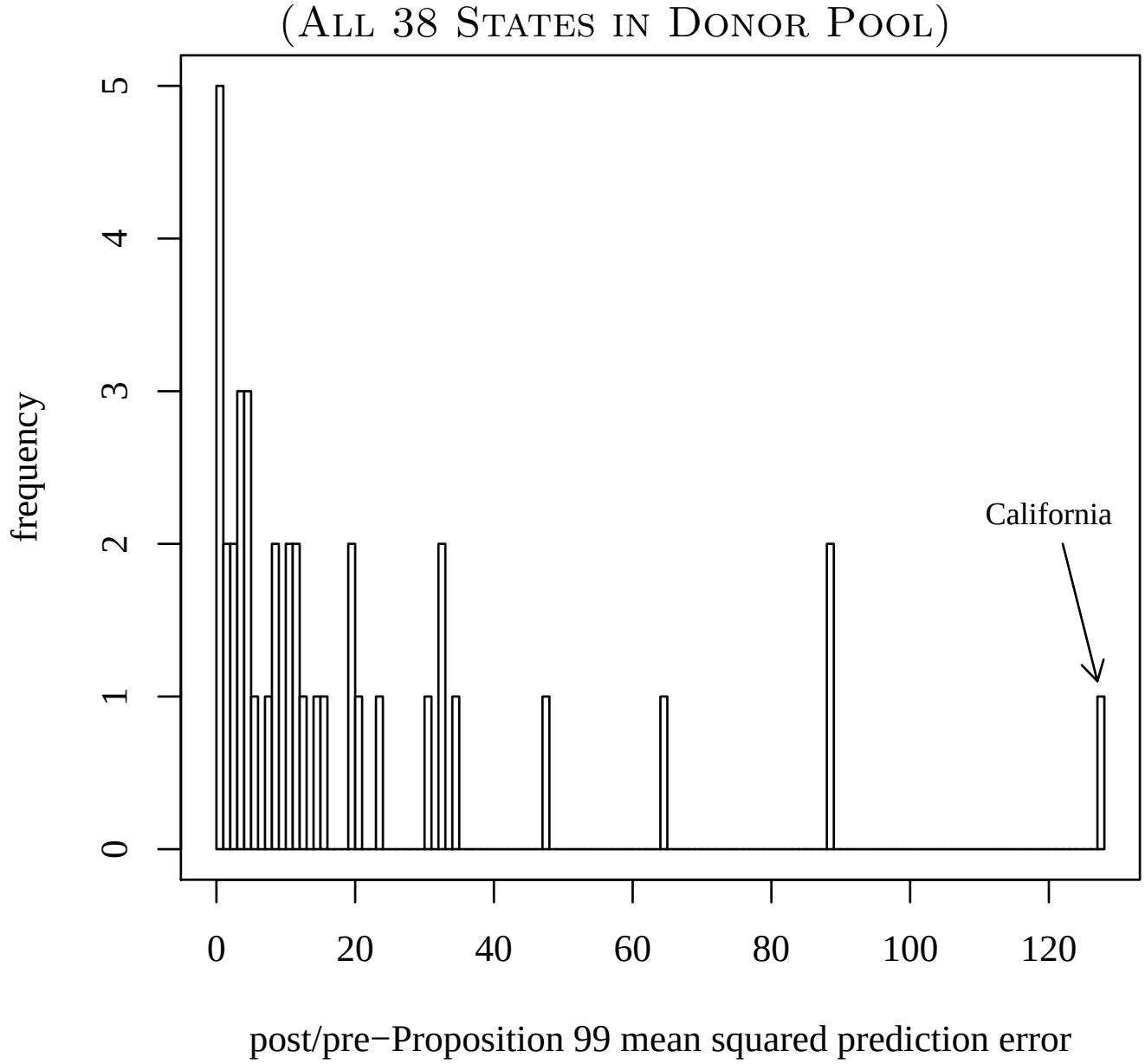
Application: California tobacco control program

(Pre-Prop. 99 MSPE $\leq$ 2 Times Pre-Prop. 99 MSPE for CA)



A primer on synthetic control estimation

Application: California tobacco control program



A primer on synthetic control estimation

- The availability of a well-defined procedure to select the comparison unit makes the estimation of the effects of placebo interventions feasible.
- The permutation method we just described does not attempt to approximate the sampling distributions of test statistics.
- Sampling-based inference is often complicated in a synthetic control setting, sometimes because of the absence of a well-defined sampling mechanism and sometimes because the sample is the same as the population.

A primer on synthetic control estimation

- This mode of inference reduces to classical randomization inference (Fisher, 1935) when the intervention is randomly assigned, a rather improbable setting.
- More generally, this mode of inference evaluates significance relative to a benchmark distribution for the assignment process, one that is implemented directly in the data.

The uniform benchmark is often employed in practice, but departures from uniformity are possible (see, Firpo and Possebom, 2018).

Why use synthetic controls?

- Compare to linear regression. Let:
  - $\boldsymbol{Y}_0$ be the $(T - T_0) \times J$ matrix of post-intervention outcomes for the units in the donor pool.
  - $\overline{\boldsymbol{X}}_1$ and $\overline{\boldsymbol{X}}_0$ be the result of augmenting $\boldsymbol{X}_1$ and $\boldsymbol{X}_0$ with a row of ones.
  - $\widehat{\boldsymbol{B}} = (\overline{\boldsymbol{X}}_0 \overline{\boldsymbol{X}}_0')^{-1} \overline{\boldsymbol{X}}_0 \, \boldsymbol{Y}_0'$ collects the coefficients of the regression of $\boldsymbol{Y}_0$ on $\overline{\boldsymbol{X}}_0$.
- $\widehat{\boldsymbol{B}}' \overline{\boldsymbol{X}}_1$ is a regression-based estimator of the counterfactual outcome for the treated unit without the treatment.
- Notice that $\widehat{\boldsymbol{B}}' \overline{\boldsymbol{X}}_1 = \boldsymbol{Y}_0 \boldsymbol{W}^{reg}$, with

$$\boldsymbol{W}^{reg} = \overline{\boldsymbol{X}}_0' (\overline{\boldsymbol{X}}_0 \overline{\boldsymbol{X}}_0')^{-1} \overline{\boldsymbol{X}}_1.$$

- The components of $\boldsymbol{W}^{reg}$ sum to one, but may be outside $[0, 1]$, allowing extrapolation, and will not be sparse.

Why use synthetic controls?

Application: German reunification

country j	W_j^{reg}	country j	W_j^{reg}
Australia	0.12	Netherlands	0.14
Austria	0.26	New Zealand	0.12
Belgium	0.00	Norway	0.04
Denmark	0.08	Portugal	-0.08
France	0.04	Spain	-0.01
Greece	-0.09	Switzerland	0.05
Italy	-0.05	United Kingdom	0.06
Japan	0.19	United States	0.13

Why use synthetic controls?

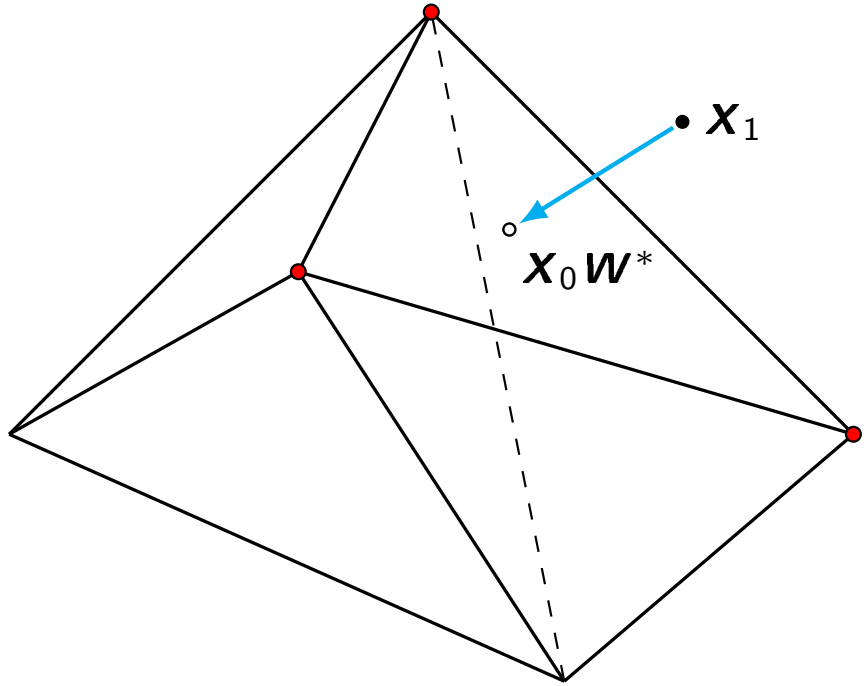
- **No extrapolation.** Synthetic control estimators preclude extrapolation outside the support of the data.
- **Transparency of the fit.** Linear regression uses extrapolation to obtain $\boldsymbol{X}_0 \boldsymbol{W}^{reg} = \boldsymbol{X}_1$, even when the untreated units are completely dissimilar in their characteristics to the treated unit. In contrast, synthetic controls make transparent the actual discrepancy between the treated unit and the convex hull of the units in the donor pool, $\boldsymbol{X}_1 - \boldsymbol{X}_0 \boldsymbol{W}^*$.
- **Safeguard against specification searches.** Synthetic controls do not require access to post-treatment outcomes in the design phase of the study, when synthetic control weights are calculated. Therefore, all design decisions can be made without knowing how they affect the conclusions of the study.

Why use synthetic controls?

- **Safeguard against specification searches (cont.)**
  Synthetic control weights can be calculated and pre-registered before the post-treatment outcomes are realized, or before the actual intervention takes place, providing a safeguard against specification searches and *p*-hacking.
- **Transparency of the counterfactual.** Synthetic controls make explicit the contribution of each comparison unit to the counterfactual of interest.
- **Sparsity.** Because the synthetic control coefficients are proper weights and are sparse, they allow a precise interpretation of the nature of the estimate of the counterfactual of interest (and of potential biases).

Why use synthetic controls?

Sparsity: Geometric interpretation



- If X_1 does not belong to the convex hull of the columns of X_0, the synthetic control $X_0 W^*$ is unique and sparse.
- If X_1 belongs to the convex hull of the columns of X_0, the synthetic control $X_0 W^*$ may not be unique and candidate W^*'s may not be sparse, although sparse solutions always exist (by Carathéodory's theorem).

Contextual requirements

- **Size of the effect and volatility of the outcome.** Small effects will be indistinguishable from other shocks to the outcome of the affected unit, especially if the outcome variable of interest is highly volatile.
- **Availability of a comparison group.** Untreated units that
  - Do not adopt interventions similar to the one under investigation during the period of the study.
  - Do not suffer large idiosyncratic shocks to the outcome of interest during the study period.
  - Have characteristics similar to the characteristics of the affected unit.
- **No anticipation.** Can be addressed by backdating.

Contextual requirements

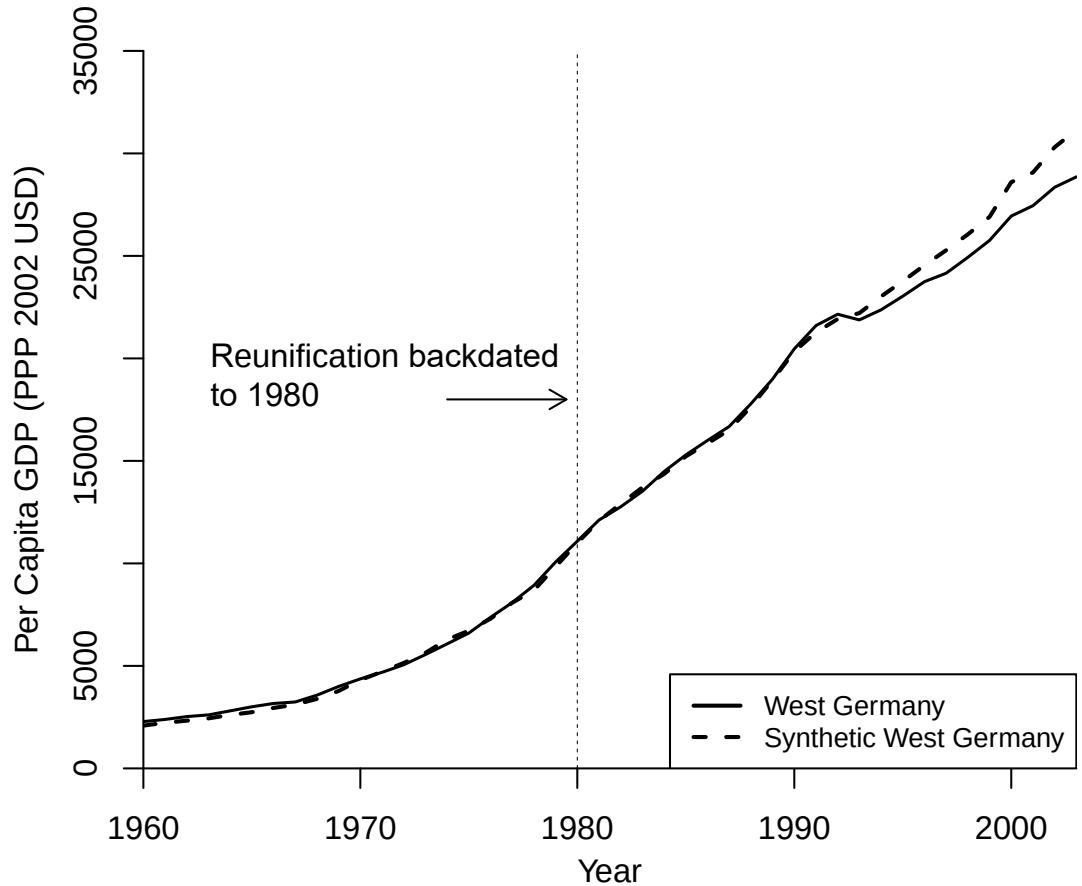
- **No interference.** Sparsity makes it possible to address interference issues.
- **Convex hull condition.** Synthetic control estimates are predicated on the idea that a combination of unaffected units can approximate the pre-intervention characteristics of the affected unit.
- **Time horizon.** The effect of some interventions may take time to emerge or to be of enough magnitude to be quantitatively detected in the data.

Data requirements

- **Aggregate data on predictors and outcomes.** Sometimes, when aggregate data do not exist aggregates of micro-data are employed in comparative case studies.
- **Sufficient pre-intervention information.** The credibility of a synthetic control estimator depends in great part on its ability to steadily track the trajectory of the outcome variable for the affected unit before the intervention. (Recall bias bound.)
- **Sufficient post-intervention information.** This may be problematic if the effect of an intervention is expected to arise gradually over time and if no forward looking measures of the outcome are available.

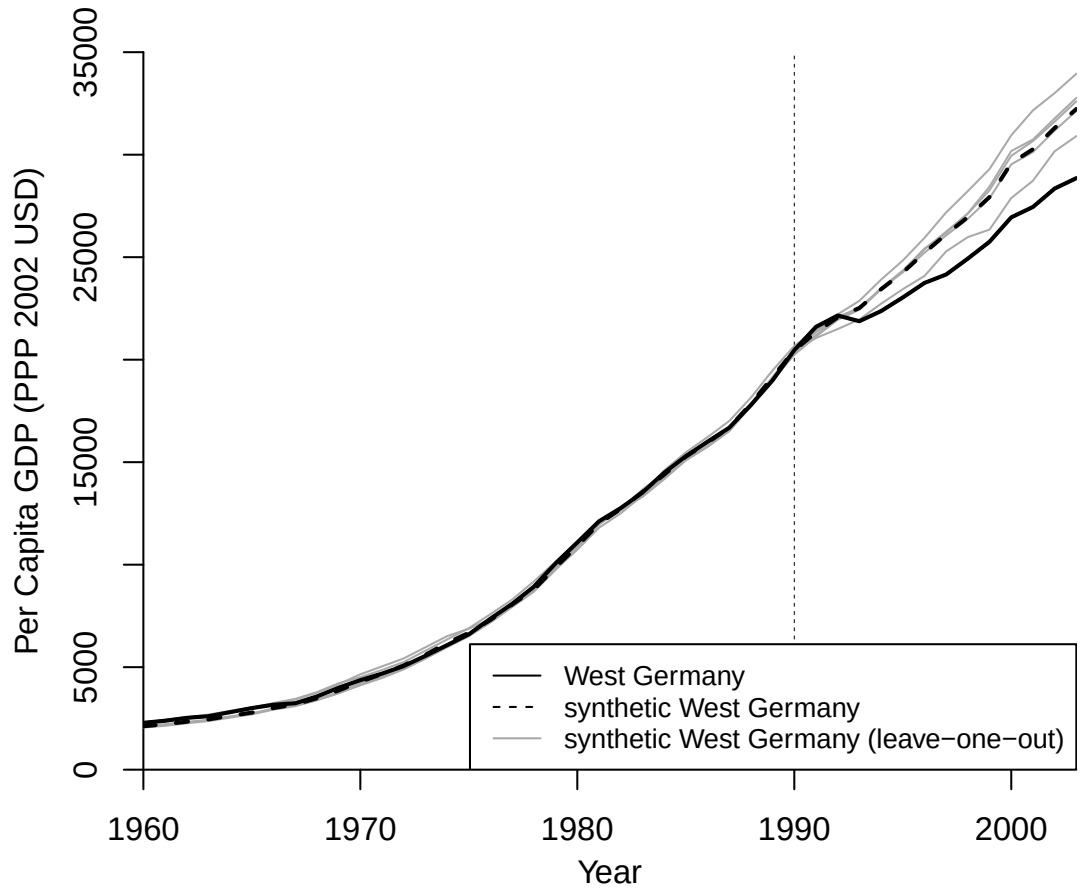
Robustness and diagnosis checks

- **Backdating.** Backdating was discussed before as a way to address anticipation effects on the outcome variable before an intervention occurs. In the absence of anticipation effects, the same idea can be applied to assess the credibility of a synthetic control in concrete empirical applications.



Robustness and diagnosis checks

- **Robustness tests.** With respect to changes in the study design. In the context of synthetic controls:
  - Units in the donor pool
  - Predictors of the outcome variable.



A penalized synthetic control estimator

Penalized synthetic control (Abadie and L'Hour, 2021): $\boldsymbol{W}^*(\lambda)$ solves

$$\min_{\boldsymbol{W}} \left\| \boldsymbol{X}_1 - \sum_{j=2}^{J+1} W_j \boldsymbol{X}_j \right\|^2 + \lambda \sum_{j=2}^{J+1} W_j \left\| \boldsymbol{X}_1 - \boldsymbol{X}_j \right\|^2$$

$$\text{s.t.} \quad W_j \geq 0, \quad \sum_{j=2}^{J+1} W_j = 1.$$

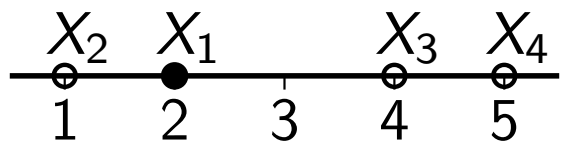
- $\lambda > 0$ controls the trade-off between fitting well the treated and minimizing the sum of pairwise distances to selected control units.
- $\lambda \to 0$: **pure synthetic control** *i.e.* synthetic control that minimizes the pairwise matching discrepancies among all solutions for the unpenalized estimator.
- $\lambda \to \infty$: **nearest neighbor** matching.

A penalized synthetic control estimator

Advantages of the penalized estimator:

1. For any $\lambda > 0$, **solution is unique and sparse** provided that untreated observations are in general position.
2. The presence of the penalization term reduces the **interpolation bias** that occurs when averaging units that are far away from each other.
3. Same **computational complexity** as the unpenalized estimator.

A penalized synthetic control estimator



- $X_1 = 2$ and $X_0 = [1\ 4\ 5]$.
- The (unpenalized) synthetic control has two sparse solutions: $W_1^* = (2/3,\ 1/3,\ 0)$ and $W_1^{**} = (3/4,\ 0,\ 1/4)$.
- W_1^* dominates W_1^{**} in terms of matching discrepancy. Infinite number of non-sparse solutions from convex combinations of these two.
- However, when $\lambda > 0$, the penalized synthetic control has a unique solution:

$$W_1^*(\lambda) = \begin{cases} (2 + \lambda/2,\ 1 - \lambda/2,\ 0)/3 & \text{if } 0 < \lambda \leq 2, \\ (1,\ 0,\ 0), & \text{if } \lambda > 2. \end{cases}$$

- As $\lambda \to 0$, $W_1^*(\lambda) \to W_1^*$, the pure synthetic control. The penalized synthetic control never uses the "bad" match X_4.

Synthetic controls for experimental design

- Suppose a ridesharing company in the US wants to assess the impact of a new pay program with higher incentives for drivers.
- A randomized control trial — or A/B test, where **drivers in a city are randomized into the new program** (treatment arm) or the status-quo (control arm) — is problematic:
  - Fairness: Such an experiment would raise equity concerns, as drivers in different treatment arms obtain different compensations for the same jobs.
  - Interference: If drivers in the active treatment arm respond to higher incentives by working longer hours, they will effectively steal business from drivers in the control arm of the experiment, which will result in biased experimental estimates.

Synthetic controls for experimental design

- The usual solution to interference is to **randomize the treatment across cities**, perhaps assigning half of the cities to treatment and half to control.
- This experiment has clear drawbacks:
  - Could be prohibitively expensive.
  - Could still raise substantial equity concerns.
  - Could create great disappointment among the drives in the treated cities (half of the US drivers!) if the program is rolled back after experimentation.
  - In some cases, the number of cities where the ridesharing company operates could be too small for effective randomization.

Synthetic controls for experimental design

- Suppose instead that the new incentive pay **treatment is applied and advertised as a pilot program** in one city or in a few cities.
- Which city or cities should be treated?
- Which city or cities should be used as a comparison/control?
- This is a setting where **randomization of treatment may create defective designs** where:
  - The treated city/cities are non-representative of the entire set of cities of interest.
  - Treated and control cities are very different in their characteristics.

Synthetic controls for experimental design

- To address these challenges, we use the **synthetic control method as an experimental design** to select treated units in non-randomized experiments, as well as the untreated units to be used as a comparison group.
- We use the name **synthetic control designs** to refer to the resulting experimental designs.
- The choice of the treated unit(s) aims to accomplish two goals:
  - The features of the treated unit(s) should be representative of the features of an aggregate of interest, like the entire country.
  - The treated unit(s) should not be idiosyncratic in the sense that their features cannot be closely approximated by the units in the control arm.

My work on synthetic control for experimental design is joint with Jinglong Zhao (BU Questrom).

Synthetic controls for experimental design

- In contrast to the observational case, in the experimental settings we have **two synthetic units**: one treated and one untreated.
- Related to work by Doudchenko (Google) and co-authors, and Jones and Barrows (Uber).

Synthetic controls for experimental design

- T time periods and J units (cities in the ridesharing company example). T_0 pre-experimental periods, T_1 post-experimental periods.
- Using information available at T_0, the analyst aims to select the set of units that will be administered treatment during the experimental periods, $T_0 + 1, T_0 + 2, \ldots, T$.
- Potential outcomes:

  Y_{jt}^I: potential outcome for unit j at time t under treatment.

  Y_{jt}^N: potential outcome for unit j at time t under no treatment.
- Treatment effects:

$$Y_{jt}^I - Y_{jt}^N,$$

  for $j = 1, \ldots, J$ and $t = T_0 + 1, \ldots, T$.
- Y_{jt}, the observed outcome, is Y_{jt}^I for treated units and Y_{jt}^N for untreated units.

Synthetic controls for experimental design

- Suppose we aim to estimate the **average treatment effect,**

$$\tau_t = \sum_{j=1}^{J} f_j (Y_{jt}^I - Y_{jt}^N),$$

for $t = T_0 + 1, \ldots, T$, where $f_1, \ldots, f_J$ are known weights (e.g., population share).

- An experimenter chooses $\boldsymbol{w} = (w_1, \ldots, w_J)$ and $\boldsymbol{v} = (v_1, \ldots, v_J)$, such that

$$\begin{aligned} & w_j \geq 0, \;\; v_j \geq 0, \\ & w_j v_j = 0, \\ & \sum_{j=1}^{J} w_j = 1, \quad \sum_{j=1}^{J} v_j = 1. \end{aligned}$$

Synthetic controls for experimental design

- Units with $w_j > 0$ are units that will be **assigned to the intervention of interest** from $T_0 + 1$ to T.
- Units with $w_j = 0$ constitute an untreated reservoir of potential control units (a "donor pool"). Among units with $w_j = 0$, those with $v_j > 0$ are **used to estimate average outcomes under no intervention**.
- A **synthetic control estimator** is

$$\tau_t(\boldsymbol{w}, \boldsymbol{v}) = \sum_{j=1}^{J} w_j Y_{jt} - \sum_{j=1}^{J} v_j Y_{jt},$$

  where Y_{jt} are observed outcomes.
- How do we choose $\boldsymbol{w}^*$ and $\boldsymbol{v}^*$?

Synthetic controls for experimental design

- The first goal of the experimenter is to choose $w_1, \ldots, w_J$ such that

$$\sum_{j=1}^{J} w_j Y_{jt}^I = \sum_{j=1}^{J} f_j Y_{jt}^I, \tag{1}$$

for $t = T_0 + 1, \ldots, T$.

- The second goal of the experimenter is to choose $v_1, \ldots, v_J$ such that

$$\sum_{j=1}^{J} v_j Y_{jt}^N = \sum_{j=1}^{J} f_j Y_{jt}^N, \tag{2}$$

or, alternatively,

$$\sum_{j=1}^{J} v_j Y_{jt}^N = \sum_{j=1}^{J} w_j Y_{jt}^N. \tag{3}$$

Synthetic controls for experimental design

- If (1) and (2) hold, then $\tau_t(\boldsymbol{w}, \boldsymbol{v})$ is equal to

$$\tau_t = \sum_{j=1}^{J} f_j(Y_{jt}^I - Y_{jt}^N),$$

  which is the **average treatment effect**.
- If (3) holds, then $\tau_t(\boldsymbol{w}, \boldsymbol{v})$ is equal to

$$\tau_t^T = \sum_{j=1}^{J} w_j(Y_{jt}^I - Y_{jt}^N),$$

  which is the **average effect of the treatment on the treated** ($\boldsymbol{w}$-weighted).
- We cannot directly fit potential outcomes because they are not directly observed. Instead, we will fit their predictors.

Estimation

- Define the **estimation periods** $\mathcal{E} \subseteq \{1, \ldots, T_0\}$, $T_{\mathcal{E}} = |\mathcal{E}|$, and let $\boldsymbol{Y}_j^{\mathcal{E}}$ be the $(T_{\mathcal{E}} \times 1)$ vector of $T_{\mathcal{E}}$ pre-intervention outcomes for unit j.
- For any $j \in \{1, 2, ..., J\}$, a vector of **predictors**, $\boldsymbol{X}_j$, is defined as

$$\boldsymbol{X}_j = \begin{pmatrix} \boldsymbol{Y}_j^{\mathcal{E}} \\ \boldsymbol{Z}_j \end{pmatrix},$$

  where $\boldsymbol{Z}_j$ are other pre-treatment covariates, aside from the outcomes, $\boldsymbol{Y}_j^{\mathcal{E}}$.
- The vector of **predictors population averages** is defined as

$$\overline{\boldsymbol{X}} = \sum_{j=1}^{J} f_j \boldsymbol{X}_j.$$

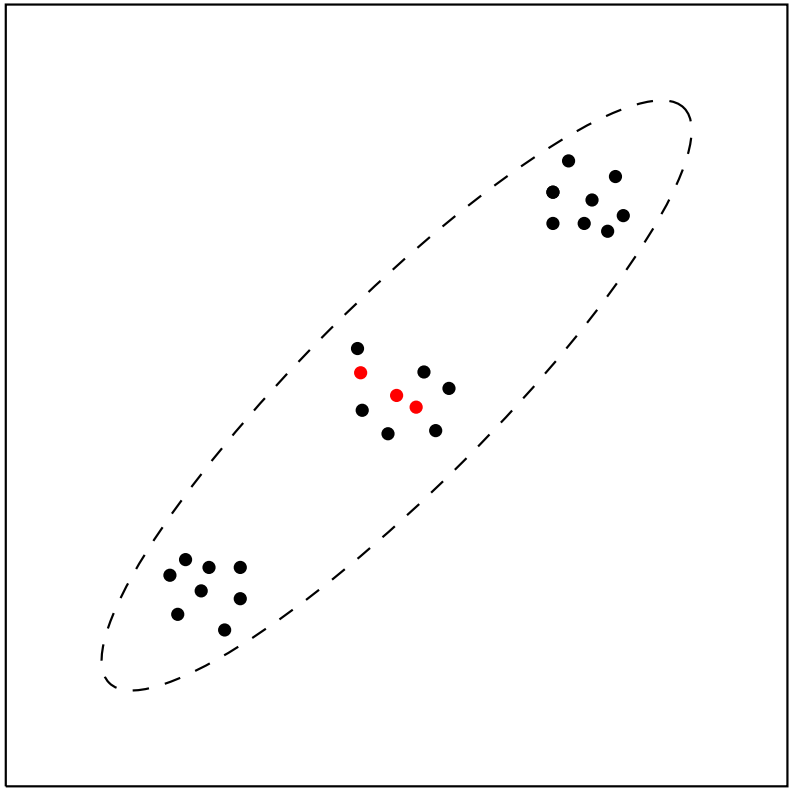
Estimation

$$\min_{\substack{w_1,\ldots,w_J,\\ v_1,\ldots,v_J}} \quad \left\| \overline{\boldsymbol{X}} - \sum_{j=1}^{J} w_j \boldsymbol{X}_j \right\|^2 + \left\| \overline{\boldsymbol{X}} - \sum_{j=1}^{J} v_j \boldsymbol{X}_j \right\|^2$$

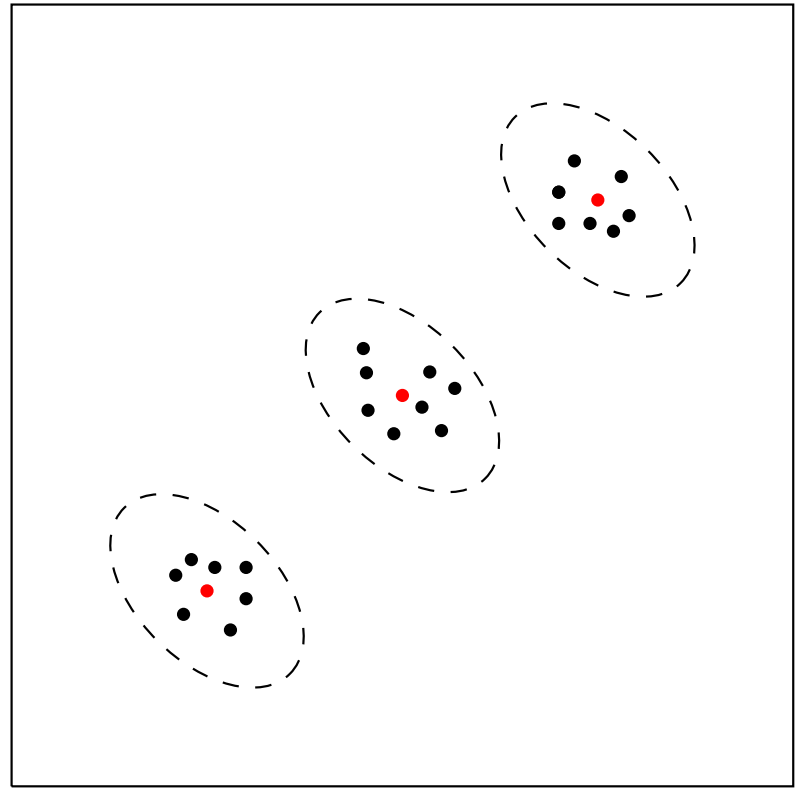
$$\begin{aligned}
\text{s.t.} \quad & \sum_{j=1}^{J} w_j = 1, \\
& \sum_{j=1}^{J} v_j = 1, \\
& w_j, v_j \geq 0, \quad j = 1, \ldots, J, \\
& w_j v_j = 0, \quad j = 1, \ldots, J, \\
& \underline{m} \leq \|\boldsymbol{w}\|_0 \leq \overline{m},
\end{aligned}$$

where $\|\boldsymbol{w}\|_0$ is the number of non-zero components of $\boldsymbol{w}$.

Estimation



(a)



(b)

Panels (a) and (b) plot the same sample values of $\boldsymbol{X}_j$. Units assigned to treatment are drawn in red. In panel (a) we treat the entire sample as a single cluster. In panel (b) we divide the sample into three clusters and assign one unit in each cluster to the treatment.

Estimation

Assume that potential outcomes follow a linear factor model,

$$Y_{jt}^N = \delta_t + \boldsymbol{\theta}_t \boldsymbol{Z}_j + \boldsymbol{\lambda}_t \boldsymbol{\mu}_j + \epsilon_{jt},$$
$$Y_{jt}^I = \upsilon_t + \boldsymbol{\gamma}_t \boldsymbol{Z}_j + \boldsymbol{\eta}_t \boldsymbol{\mu}_j + \xi_{jt},$$

where $\boldsymbol{Z}_j$ is a $(r \times 1)$ vector of observed covariates; $\boldsymbol{\theta}_t$ and $\boldsymbol{\gamma}_t$ are $(1 \times r)$ vectors of unknown parameters; $\boldsymbol{\mu}_j$ is a $(F \times 1)$ vector of unobserved covariates; $\boldsymbol{\lambda}_t$ and $\boldsymbol{\eta}_t$ are $(1 \times F)$ vectors of unknown parameters; ϵ_{jt} and ξ_{jt} are unobserved mean-zero random shocks.

Estimation

Assume also that, with probability one,

$$\sum_{j=1}^{J} w_j^* \boldsymbol{Z}_j = \sum_{j=1}^{J} f_j \boldsymbol{Z}_j, \qquad \sum_{j=1}^{J} w_j^* Y_{jt} = \sum_{j=1}^{J} f_j Y_{jt}, \qquad \forall t \in \mathcal{E},$$

and

$$\sum_{j=1}^{J} v_j^* \boldsymbol{Z}_j = \sum_{j=1}^{J} f_j \boldsymbol{Z}_j, \qquad \sum_{j=1}^{J} v_j^* Y_{jt} = \sum_{j=1}^{J} f_j Y_{jt}, \qquad \forall t \in \mathcal{E}.$$

Estimation

Under the assumptions above (and additional regularity conditions), we obtain

$$|E\left[\widehat{\tau}_t - \tau_t\right]| \leq c\frac{\bar{\sigma}}{\sqrt{T_{\mathcal{E}}}},$$

where $\bar{\sigma}^2$ is a bound on the variance proxy of ε_{jt}.

Inference

- Recall

$$Y_{jt}^N = \delta_t + \boldsymbol{\theta}_t \boldsymbol{Z}_j + \boldsymbol{\lambda}_t \boldsymbol{\mu}_j + \epsilon_{jt},$$
$$Y_{jt}^I = \upsilon_t + \boldsymbol{\gamma}_t \boldsymbol{Z}_j + \boldsymbol{\eta}_t \boldsymbol{\mu}_j + \xi_{jt}.$$

- Null hypothesis: For $t = T_0 + 1, \ldots, T$, and $j = 1, \ldots, J$,

$$Y_{jt}^I = \delta_t + \boldsymbol{\theta}_t \boldsymbol{Z}_j + \boldsymbol{\lambda}_t \boldsymbol{\mu}_j + \xi_{jt},$$

  where ξ_{jt} has the same distribution as ϵ_{jt}.

- Under the null hypothesis, the distribution of Y_{jt}^I is the same as the distribution of Y_{jt}^N.

Inference

- Blank periods: $\mathcal{B} \subseteq \{1, \ldots, T_0\} \setminus \mathcal{E}$, which comprise pre-intervention periods whose outcomes Y_{jt} have not been used to calculate $\boldsymbol{w}^*$ or $\boldsymbol{v}^*$
- "Placebo" treatment effects estimated for the blank periods: for $t \in \mathcal{B}$,

$$\widehat{u}_t = \sum_{j=1}^{J} w_j^* Y_{jt} - \sum_{j=1}^{J} v_j^* Y_{jt}$$

- Post-intervention estimates of the treatment effects: for $t \in \{T_0 + 1, ..., T\}$,

$$\widehat{\tau}_t = \sum_{j=1}^{J} w_j^* Y_{jt} - \sum_{j=1}^{J} v_j^* Y_{jt}$$

- Define the vector

$$\begin{aligned} \widehat{\boldsymbol{r}} &= (\widehat{r}_1, \ldots, \widehat{r}_{T-T_{\mathcal{E}}}) \\ &= (\widehat{\tau}_{T_0+1}, \ldots, \widehat{\tau}_T, \widehat{u}_{t_1}, \ldots, \widehat{u}_{t_{T_{\mathcal{B}}}}). \end{aligned}$$

Inference

- Permutation test: let Π be the set of all T_1-combinations of $\{1, 2, ..., T - T_{\mathcal{E}}\}$; for each $\pi \in \Pi$, let $\pi(i)$ be the i^{th} smallest value in π.
- Define the $(T_1 \times 1)$-vector

$$\widehat{\boldsymbol{e}}_\pi = (\widehat{r}_{\pi(1)}, \widehat{r}_{\pi(2)}, ..., \widehat{r}_{\pi(T_1)}).$$

- Test statistic:

$$S(\boldsymbol{e}_\pi) = \frac{1}{T_1} \sum_{t=1}^{T_1} |e_t| .$$

- p-value:

$$\widehat{p} = \frac{1}{|\Pi|} \sum_{\pi \in \Pi} 1\{S(\widehat{\boldsymbol{e}}_\pi) \geq S(\widehat{\boldsymbol{e}})\},$$

where $\widehat{\boldsymbol{e}} = (\widehat{\tau}_{T_0+1}, \ldots, \widehat{\tau}_T)$.

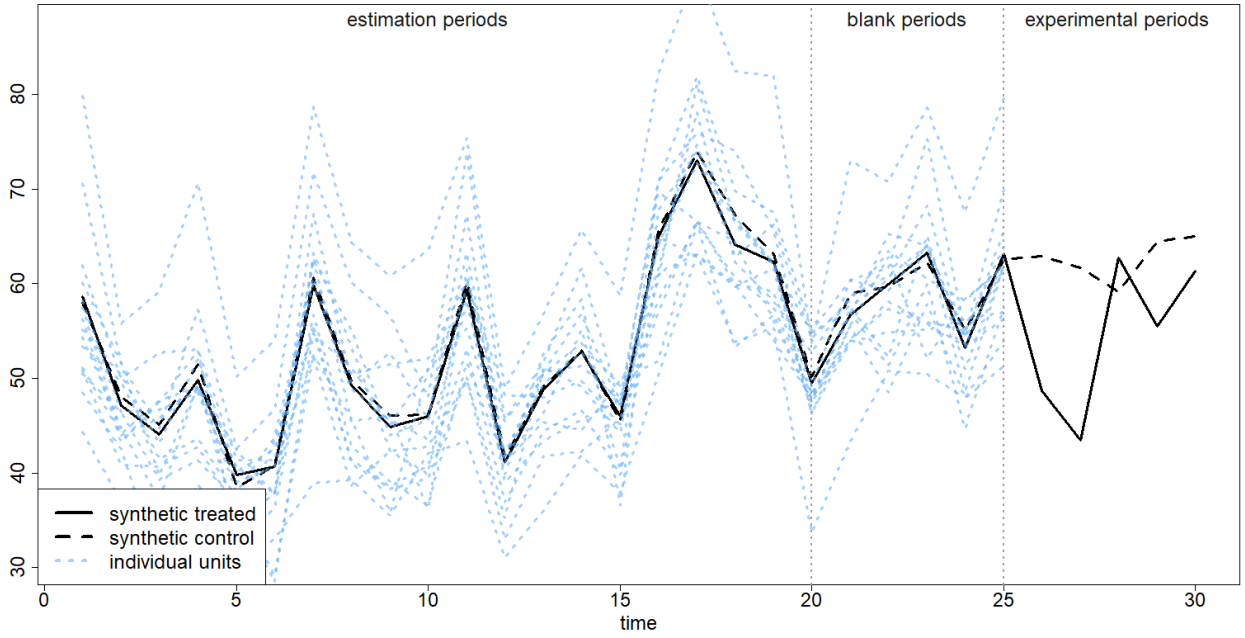
Inference

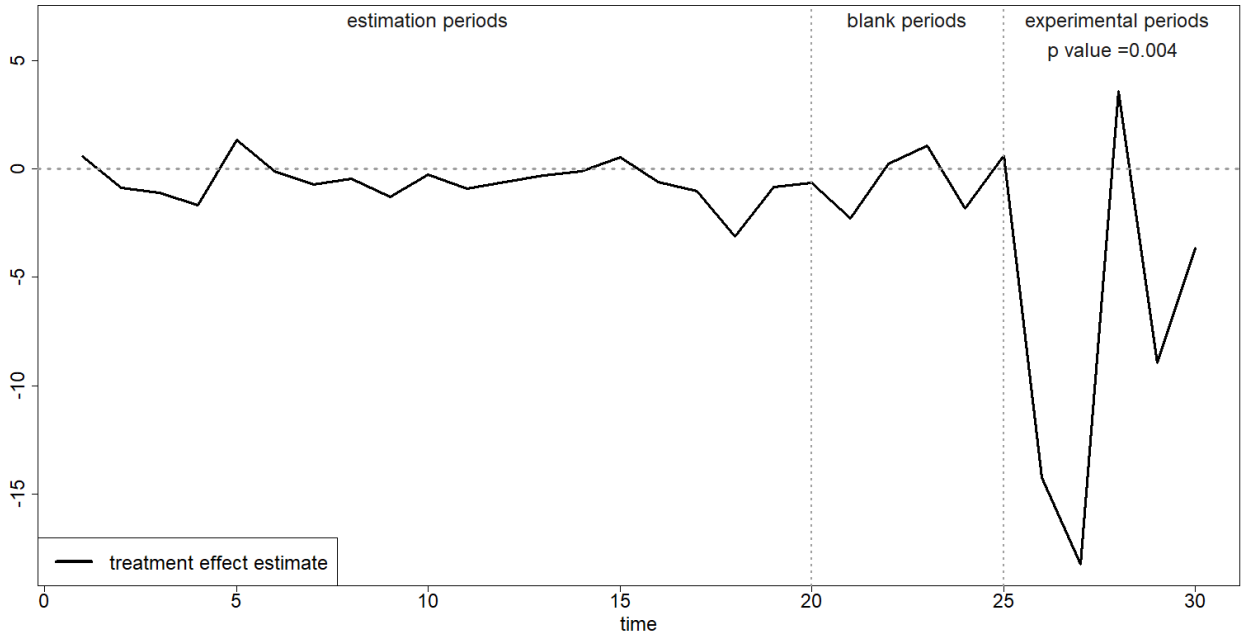
Under the assumptions above (and additional regularity conditions), we obtain that, for any $\alpha \in (0, 1]$,

$$\alpha - \frac{1}{|\Pi|} \leq \Pr(\widehat{p} \leq \alpha) \leq \alpha,$$

under the null (cf. Chernozhukov et al., 2019).

Synthetic controls for experimental design





$B = 1000$ simulations: Synthetic design

		$\frac{1}{B}\sum_{b=1}^{B}\tau_t^{(b)}$					
		$t = 26$	$t = 27$	$t = 28$	$t = 29$	$t = 30$	
		-13.05	-10.92	-7.65	-5.17	-2.22	
		$\frac{1}{B}\sum_{b=1}^{B}\widehat{\tau}_t^{(b)}$					*MAE*
		$t = 26$	$t = 27$	$t = 28$	$t = 29$	$t = 30$	
Unconstrained		-13.06	-10.89	-7.63	-5.14	-2.22	0.85
Constrained	$\bar{m} = 1$	-12.99	-10.83	-7.74	-5.19	-2.34	2.84
	$\bar{m} = 2$	-12.99	-10.93	-7.58	-5.07	-2.17	1.66
	$\bar{m} = 3$	-13.04	-10.97	-7.58	-5.12	-2.13	1.22
	$\bar{m} = 4$	-13.06	-10.91	-7.59	-5.19	-2.19	1.02
	$\bar{m} = 5$	-13.06	-10.90	-7.63	-5.14	-2.20	0.92
	$\bar{m} = 6$	-13.07	-10.96	-7.64	-5.15	-2.22	0.88
	$\bar{m} = 7$	-13.06	-10.89	-7.63	-5.14	-2.21	0.85

$B = 1000$ simulations: Difference in means

$$\frac{1}{B}\sum_{b=1}^{B}\tau_t^{(b)}$$

$t = 26$	$t = 27$	$t = 28$	$t = 29$	$t = 30$
-13.05	-10.92	-7.65	-5.17	-2.22

$$\frac{1}{B}\sum_{b=1}^{B}\widehat{\tau}_t^{(b)}$$

	$t = 26$	$t = 27$	$t = 28$	$t = 29$	$t = 30$	*MAE*
$\overline{m} = 1$	-12.90	-10.65	-7.17	-4.90	-2.08	5.73
$\overline{m} = 2$	-12.87	-10.67	-7.33	-5.05	-1.98	4.54
$\overline{m} = 3$	-12.68	-10.51	-7.22	-4.87	-1.85	3.83
$\overline{m} = 4$	-12.99	-10.94	-7.78	-5.13	-2.16	3.38
$\overline{m} = 5$	-12.93	-10.75	-7.45	-5.14	-2.17	3.10
$\overline{m} = 6$	-12.91	-10.85	-7.53	-5.09	-2.12	3.08
$\overline{m} = 7$	-12.91	-10.66	-7.53	-5.02	-2.07	2.87

Synthetic controls for experimental design

- Experimental design methods have largely been concerned with settings where a large number of experimental units are randomly assigned to treatment and control.
- This focus on large samples and randomization has proven to be enormously useful in large classes of problems.
- However, it becomes inadequate when treating more than a few units is unfeasible, which is often the case in experimental studies with large aggregate units (e.g., cities).
- We have applied synthetic control techniques, widely used in observational studies, to the design of experiments when treatment can only be applied to a small number of experimental units.

Synthetic controls for experimental design

- The synthetic control design optimizes jointly over the identities of the units assigned to the treatment and the control arms, and over the weights that determine the relative contribution of those units to reproduce the counterfactuals of interest.
- Corporate research units and academic investigators are often confronted with settings where interventions at the level of micro-units (i.e., customers, workers, or families) are unfeasible, unethical, impractical or ineffective.
- There is, in consequence, a wide range of potential applications of experimental design methods for large aggregate entities.

Closing remarks

- Synthetic controls provide many practical advantages for the estimation of the effects of policy interventions and other events of interest.
- Like for any other statistical procedure (and especially for those aiming to estimate causal effects), the credibility of the results depends crucially on the level of diligence exerted in the application of the method and on whether contextual and data requirements are met in the empirical application at hand.

Closing remarks

- Some open areas of research: sampling-based inference, external validity, sensitivity to model restrictions, estimation with multiple interventions, data driven selectors of v_h, mediation analysis ...
- Results on robust and efficient computation of synthetic controls are scarce, and more research is needed on the computational aspects of this methodology.
- On the empirical side, many of the events and the policy interventions economists care about take place at an aggregate level, affecting entire aggregate units.

Resources

The material in this presentation comes from:

- Abadie, A. and J. Gardeazabal. 2003. "The Economic Costs of Conflict: A Case Study of the Basque Country." American Economic Review, 93(1), 112-132.
- 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(490), 493-505.
- Abadie, A., A. Diamond, and J. Hainmueller. 2015. "Comparative Politics and the Synthetic Control Method". American Journal of Political Science, 59(2), 495-510.
- Abadie, A. 2021. "Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects." Journal of Economic Literature, 59(2), 391-425.
- Abadie, A. and J. L'Hour. 2021. "A Penalized Synthetic Control Estimator for Disaggregated Data." Journal of the American Statistical Association, 116(536), 1817-1834.
- Abadie, A. and J. Zhao. 2021. Synthetic Controls for Experimental Design.

Code: `synth` (Matlab, Stata and R)
`http://web.stanford.edu/~jhain/synthpage.html`
`pensynth` (R)
`https://github.com/jeremylhour/pensynth`

Resources

While this talk has mostly focused on my work, many have contributed to the literature on synthetic control estimators and related methods. Some references:

- Acemoglu, D., S. Johnson, A. Kermani, J. Kwak, and T. Mitton. 2016. "The Value of Connections in Turbulent Times: Evidence from the United States. Journal of Financial Economics, 121, 368–391.
- Amjad, M., D. Shah, and D. Shen. 2018. "Robust Synthetic Control." Journal of Machine Learning Research, 19(22), 1–51.
- Amjad, M. J., V. Misra, D. Shah, and D. Shen. 2019. "mRSC: Multidimensional Robust Synthetic Control." In Proc. ACM Meas. Anal. Comput. Syst., Volume 3.
- Agarwal, A., D. Shah, and D. Shen. 2020. "Synthetic Interventions." arXiv:2006.07691.
- Arkhangelsky, D., S. Athey, D. A. Hirshberg, G. W. Imbens, and S. Wager. 2019. "Synthetic Difference in Differences." American Economic Review, 111(12), 4088-4118.
- Athey, S., M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi. 2021. "Matrix Completion Methods for Causal Panel Data Models." Journal of the American Statistical Association, 116(536), 1716-1730.

Resources

- Botosaru, I. and B. Ferman. 2019. "On the Role of Covariates in the Synthetic Control Method." The Econometrics Journal, 22(2): 117–130.
- Cattaneo, M. D., Y. Feng, and R. Titiunik. 2021. "Prediction Intervals for Synthetic Control Methods." Journal of the American Statistical Association, 116(536), 1865-1880.
- Chen, J. 2023. "Synthetic Control As Online Linear Regression." Econometrica, 91(2), 465-491.
- Chernozhukov, V., K. Wüthrich, and Y. Zhu. 2019a. "An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls." Journal of the American Statistical Association, 116(536), 1849-1864.
- Chernozhukov, V., K. Wüthrich, and Y. Zhu. 2019b. "Practical and Robust *t*-test Based Inference for Synthetic Control and Related Methods." arXiv:1812.10820v2.
- Doudchenko, N., K. Khosravi, J. Pouget-Abadie, S. Lahaie, M. Lubin, V. Mirrokni, and J. Spiess, et al. 2021. "Synthetic Design: An Optimization Approach to Experimental Design with Synthetic Controls." Advances in Neural Information Processing Systems, 34.

Resources

- Doudchenko, N. and G. W. Imbens. 2016. "Balancing, Regression, Difference-in-Differences and Synthetic Control Methods: A Synthesis." arXiv:1610.07748.
- Ferman, B. 2021. "On the Properties of the Synthetic Control Estimator with Many Periods and Many Controls." Journal of the American Statistical Association, 116(536), 1764-1772.
- Firpo, S. and V. Possebom. 2018. "Synthetic Control Method: Inference, Sensitivity Analysis and Confidence Sets." Journal of Causal Inference, 6(2).
- Gunsilius, F. 2020. "Distributional Synthetic Controls." Econometrica (forthcoming).
- Li, K. T. 2020. "Statistical Inference for Average Treatment Effects Estimated by Synthetic Control Methods." Journal of the American Statistical Association 115, 2068-2083.
- 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.
- Samartsidis, P., S. R. Seaman, A. M. Presanis, M. Hickman, and D. De Angelis. 2019. "Assessing the Causal Effect of Binary Interventions from Observational Panel Data with Few Treated Units. Statistical Science, 34(3), 486-503 .

Thank you!

abadie@mit.edu