EC 709: Synthetic Control and Extensions

Liang Zhong[1]

Boston University

samzl@bu.edu

Novemeber 2023

[1]This introduction to the topic of synthetic controls is based on the paper "Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects."

Overview

1. Introduction
2. Theoretical Justification
3. Implementation concerns in practice
4. Synthetic DID

Table of Contents

1 Introduction

2 Theoretical Justification

3 Implementation concerns in practice

4 Synthetic DID

Motivation

- 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).
  - Immigration Policy: Bohn et al., (2014); Borjas (2017); Peri and Yasenov (2017)
  - Minimum wages Policy: Allegretto et al., (2017); Jardim et al., (2017); Neumark and Wascher (2017); Reich et al., (2017)

$\Rightarrow$ Aim to estimate the effects of aggregate interventions

- DID is great if PT holds. However, we might often encounter the case that PT failed.

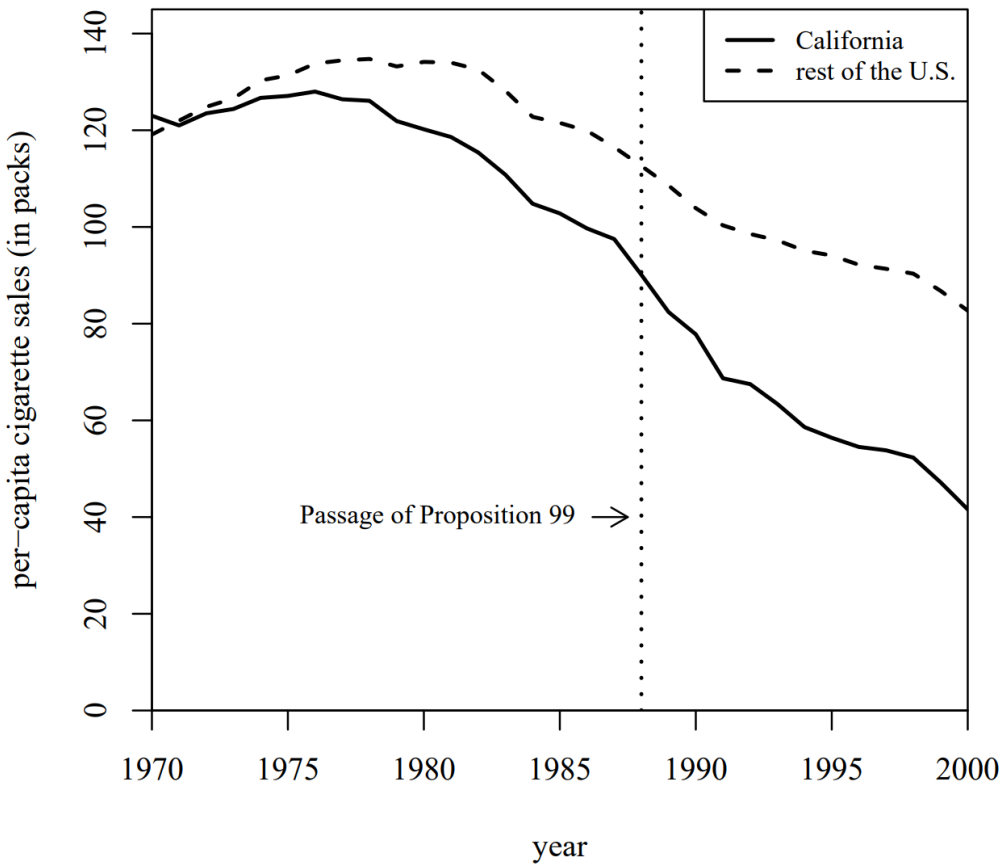
? What else can we do?

Example: Proposition 99 on Cigarette Consumption

- California increased the excise tax for Cigarette by 25 cents/pack in 1989
- What is the causal effect of the legislation on "smoking rates" in California afterward?
- Outcome of interest: per capita cigarette sales (packs/year)
- We have data (i) in the absence of smoking legislation in California in 1988 and prior, and (ii) for other states before, and (iii) for the other states after the change. (and other variables, but not of essence)

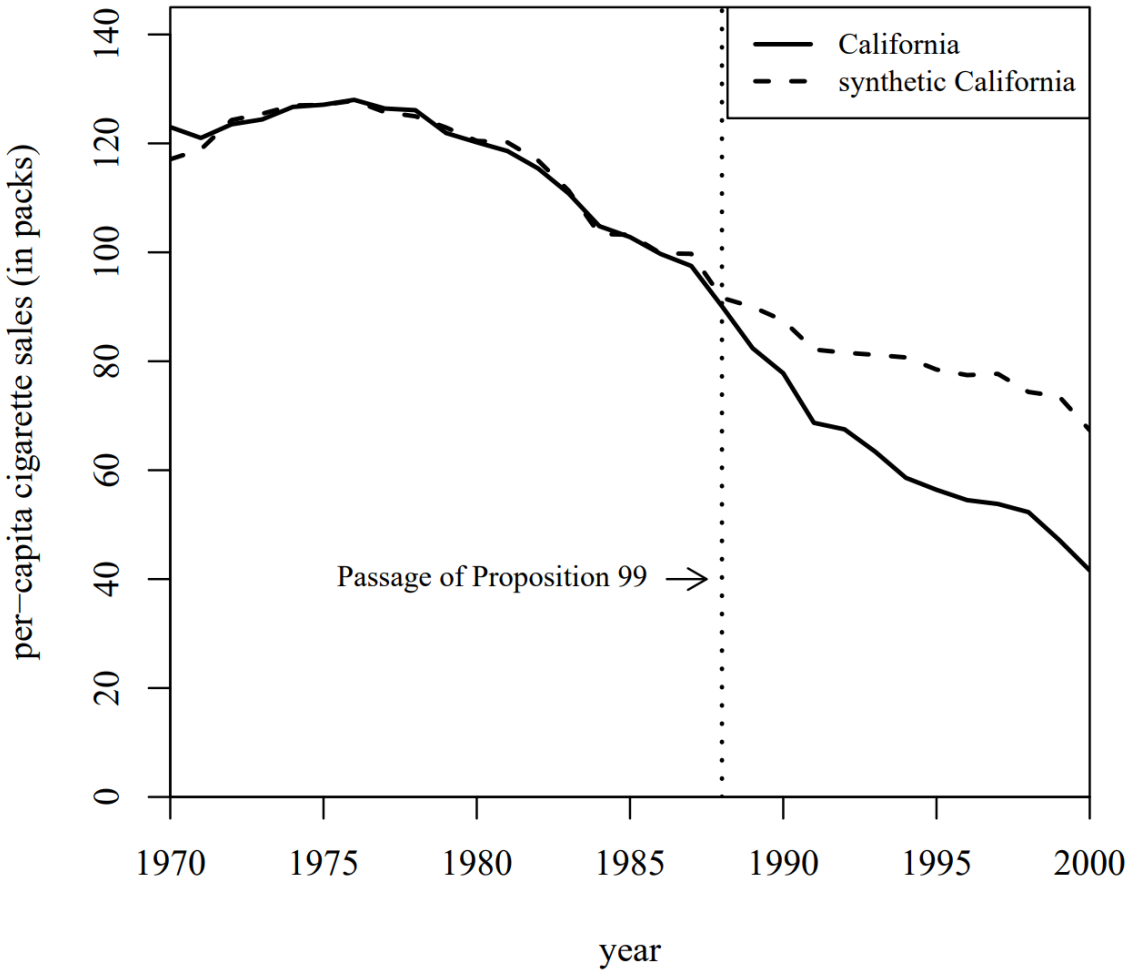
⋆ We observed smoking rates in California in 1989 and later given the legislation. We need to impute the counterfactual smoking rates in California during those years had the legislation not been enacted

Row Plot on the Outcome Variable



- Using 38 states that had never passed such programs as controls
- PT clearly failed
- What else can we do?

The Plot from SCM



Use a selected weighted average of all potential comparison units that best resemble the characteristics of the treated unit(s) as the comparison unit

Advantages of SCM

1. 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
2. Can estimate not only average treatment effect across time periods, but also the effect on each period
3. The SC method potentially allows for time-varying unobserved confounders

- Now, formally, why does it work?

Table of Contents

1. Introduction
2. Theoretical Justification
3. Implementation concerns in practice
4. Synthetic DID

Formal Setup

- Observe $J+1$ units in periods $1, 2, ..., T$
  - Unit "one" is treated during periods $T_0+1, ..., T$
  - The remaining J units are an untreated reservoir of potential controls (a "donor pool")

- Y_{it}^I: outcome that would be observed for unit i at time t if unit i is treated in periods T_0+1 to T
- Y_{it}^N: outcome that would be observed for unit i at time t if unit i is not treated
- Parameter of interest: $ATT(t) \equiv \tau_{1t} = Y_{1t}^I - Y_{1t}^N = Y_{1t} - Y_{1t}^N$ for $t > T_0$

⋆ How to construct Y_{1t}^N?

Assumption 1: Linear factor Model for Potential Outcomes

- $Y_{it}^I = \tau_{it} + Y_{it}^N$
- $Y_{it}^N = \theta_t' Z_i + \eta_t + \lambda_t \mu_i + \epsilon_{it}$
  - Z_i are observed features, the rest including μ_i are unobserved features
  - λ_t: a $1 \times F$ vector of time-varying factors
  - μ_i: a $F \times 1$ vector of factor loadings (treated as fixed)
  - ϵ_{it} is a unit-level transitory shock, modeled as random noise
- Example: μ_i is vector of state i's industry shares, while λ_t represents outcomes for each industry in time t

$\Rightarrow$ Allows time-varying confounders: nest TWFE model when λ_t is constant

Assumption 2: Treatment Not Related to Random Shock

- Conditional on treatment assignment $E[\epsilon_{it}] = 0$ for all i and t
- Important: it may be that treatment assignment is correlated with the linear factor structure $\lambda_t \mu_i$
  - States with similar (unobserved) industry mix μ_i and thus similar time trends may be more likely to be treated
- Can we still have an unbiased estimator for τ_{it}?

The Ideal Procedure

- For the moment, suppose we observe μ_i as well, and $M = (\mu_2, ...\mu_{J+1})'$
- Let $\Delta_J = \{W = (w_2, ..., w_{J+1})' \in R^J : \sum_{i=2}^{J+1} w_i = 1, w_i \geq 0\}$, the set of all potential synthetic controls
- Denote $Z = (Z_2, ...Z_{J+1})'$, $Y_t = (Y_{2,t}, ...Y_{J+1,t})'$
- If there are weights $W^S \in \Delta_J$ such that $\mu_1 = M'W^S$ and $Z_1 = Z'W^S$:

$\Rightarrow$ $Y_t'W^S$ is an unbiased estimator for Y_{1t}^N

  - $Y_t'W^S$ is affected by the common shocks (λ_t) in the same way as the treated unit
  - Linear combination of idiosyncratic shocks is uncorrelated with treatment assignment

$\Rightarrow$ $\hat{\tau}_{1t} = Y_{1t} - Y_t'W^S$ is unbiased

- However, we don't observe μ_i

The Feasible Procedure in Practice

- Key intuition: Since these μ_i are evidenced in the pre-treatment outcomes for both treated and controls, we can try to use this information to "balance" on μ_i

$\Rightarrow$ Let $X_1 = (Z_1, Y_{1,1}, ..., Y_{1,T_0})$ be a $(k \times 1)$ vector of pre-intervention characteristics for the treated unit; X_0 be a $(k \times J)$ matrix which contains the same variables for the unaffected units

- $W^\star = (w_2^\star, ..., w_{J+1}^\star)' = argmin_{W \in \Delta_J} \|X_1 - X_0 W\|$
- $\|X_1 - X_0 W\| = \sqrt{(X_1 - X_0 W)' V (X_1 - X_0 W)}$, where V is some $(k \times k)$ symmetric and positive semidefinite matrix
    - Various ways to choose V (subjective assessment of predictive power of X, minimize MSPE, cross-validation, etc.), See Section 3.2 of Abadie (2021) for details

$\Rightarrow$ $\hat{\tau}_{1t} = Y_{1t} - \sum_{j=2}^{J+1} w_j^\star Y_{jt}$

"Balance" of SCM

Variables	California		Average of
	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: All variables except lagged cigarette sales are averaged for the 1980-1988 period (beer consumption is averaged 1984-1988).

Figure 1: Predictor Means: Actual vs. Synthetic California

Bias of SCM

- If $X_1 = X_0 W^\star$: $\|E(\hat{\tau}_{1t} - \tau_{1t})\| < Bias(|\lambda_{tf}|, |\epsilon_{jt}|, F, J, T_0)$, where bias:
  - Increase with $|\epsilon_{jt}|$: Y_{it} provides noisy information to $\mu_i \Rightarrow W^\star \neq W^S$
  - Decrease with T_0: less affected by an individual error term ϵ_{jt}
  - Increase with F: Less likely to fit μ_i perfectly
  - Increase with J: Makes it easier to fit pretreatment outcomes even when there are substantial discrepancies in factor loadings between the treated unit and the synthetic control

$\Rightarrow$ A large T_0 cannot drive down the bias since $\mu_1 \neq M' W^\star$ even if $X_1 = X_0 W^\star$

- If $|\epsilon_{jt}| \to 0$, then $Bias \to 0$

$\star$ The Result above is only for the ideal scenario of close fit, and the bias is nonzero in general

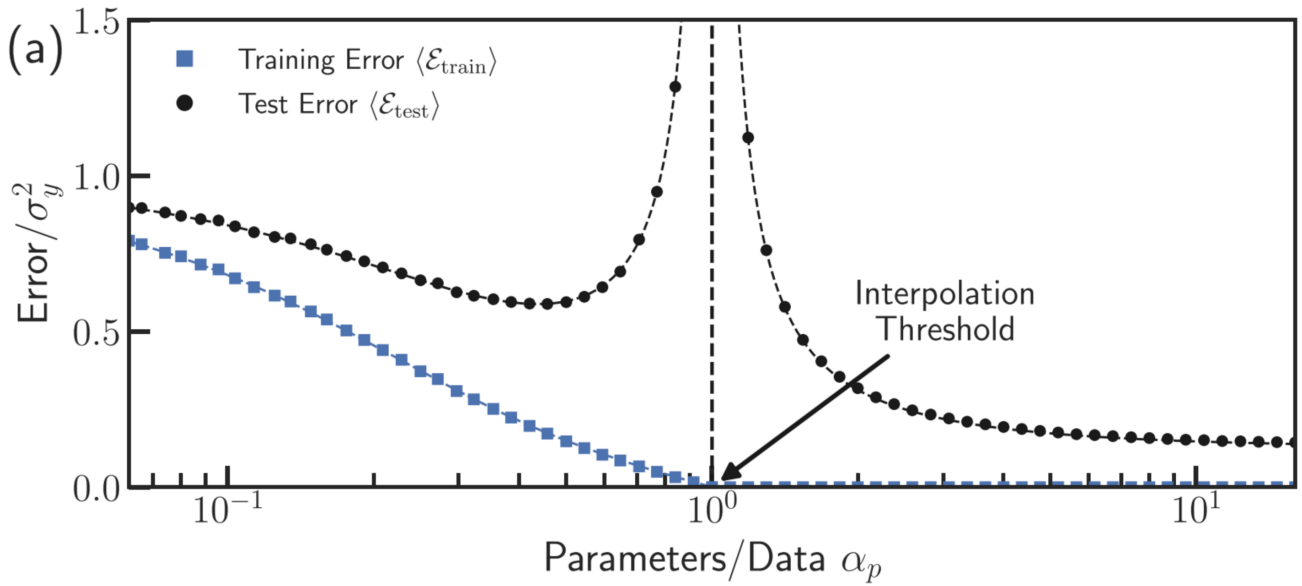
Takeaways from the bias formula

1. 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
    - 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
2. Settings with small T_0, large J, and large noise create substantial risk of overfitting

    $\Leftarrow$ To reduce interpolation biases and risk of overfitting (large J), restrict the donor pool to units that are similar to the treated unit
3. Even with $T_0 \to \infty$, bias is not guaranteed to be 0
    - Ferman and Pinto (2019) proposed a demeaning estimator for lower bias and variance

    $\Leftarrow$ Replace Y_{jt} with $Y_{jt} - \bar{Y}_j$, where $\bar{Y}_j$ is the pre-treatment average
    - Ferman (2020): Under additional assumptions, If both $J \to \infty$ and $T_0 \to \infty$, then $Bias \to 0$ (Similar to Double Descent)

A Side Note: Double Descent



- Small error with a small number or an **extremely** large number of parameters
- Large error if the number of parameters is about the same as the number of data points used to train the model
- Over-fitting is still a severe issue in practice

Table of Contents

1 Introduction

2 Theoretical Justification

3 Implementation concerns in practice

4 Synthetic DID

Inference procedure

- A lot of options available nowadays: Chernozhukov et.al (2021) proposed a robust permutation test; Chernozhukov et.al (2023) Proposed a t-test

- I will introduce the Randomization Inference framework from Abadie:

1. Iteratively reassigning the treatment to the units in the donor pool and estimating "placebo effects" in each iteration
2. The effect of the treatment on the unit affected by the intervention is deemed to be significant when the test statistics' magnitude is extreme relative to the permutation distribution

Choice of Test Statistics

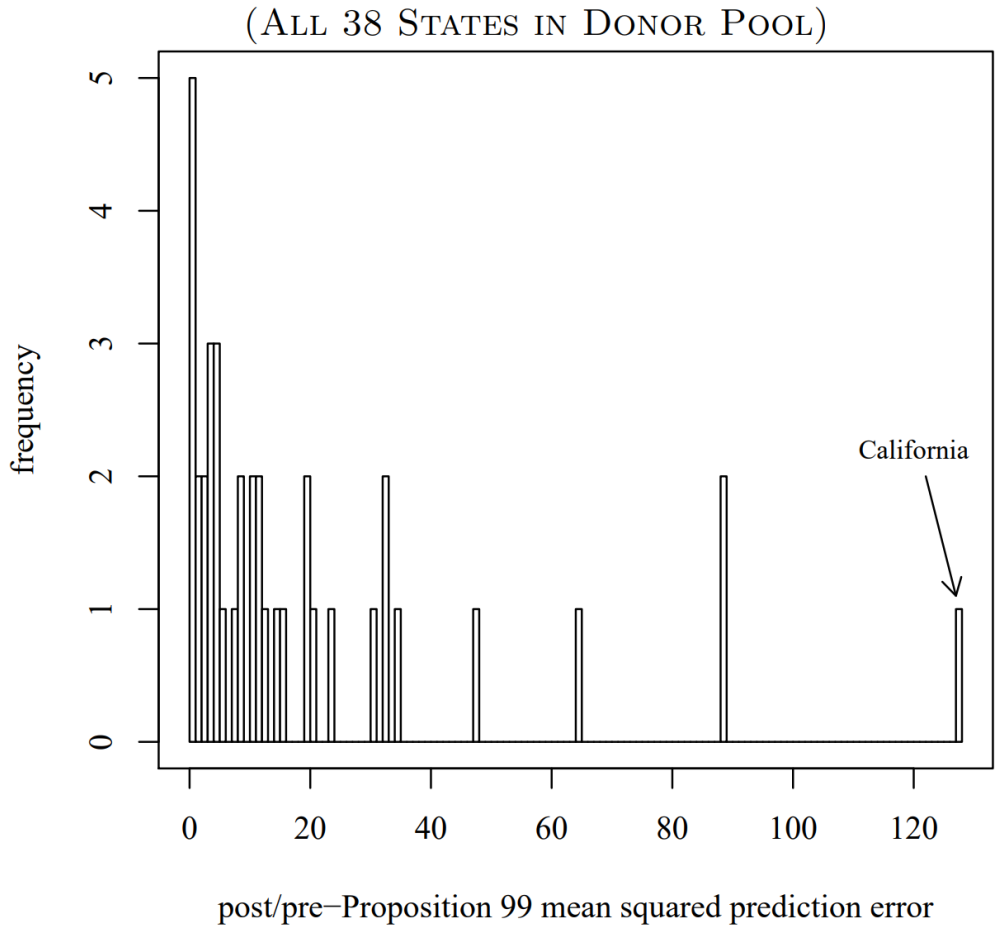
- The design of the test statistics for the "placebo effects" needs to take into account:
  - Estimation of interest is a vector: τ_{1t} for each t
  - Even if a synthetic control is able to closely fit the trajectory of the outcome variable for the treated unit before the intervention, the same may not be true for all the units in the donor pool

⋆ Abadie proposed to use the ratio between the post-intervention RMSPE and pre-intervention RMSPE

  - RMSPE: Root mean squared prediction error between the treated unit and the synthetic unit
  - post-intervention RMSPE is large with a large treatment effect
  - pre-intervention RMSPE is small with good-fit

Result from RI

Application: California tobacco control program



Link of the package: Code: synth (Matlab, Stata and R)

Concerns with Wrong Underlying Model

- Assumption 1 is very strong: Linear model is unlikely to hold in practice
  - If the underlying model is nonlinear, even a close fit by a synthetic control could potentially result in large interpolation biases
- Abadie (2021): "The factor model should be interpreted only as an approximation to a more general (nonlinear) process for Y_{it}^N."

$\star$ Practical implication: We shouldn't rely too much on the linear factor model for the validity of SCM

- Same spirit as DID, we shouldn't rely too much on the no pre-trend; Economics intuitions matter the most

$\Rightarrow$ Each of the units in the donor pool has to be chosen judiciously to provide reasonable control for the treated unit.

- Choose units that have similar observed attributes Z_j and no suspected large differences in the values of the unobserved attributes μ_j relative to the treated unit
- Including unsuitable controls is a recipe for bias

Some Recommended Tests for SCM

1. How robust is our result to the choice of units in the donor pool?

$\Rightarrow$ Leave-one-out re-analysis: taking out units that contribute to the synthetic control one-at-a-time

- If results have the same sign and centered around the result produced using the entire donor pool, then it is pretty robust
- Suspicious if the exclusion of a unit has a large effect on results without a discernible change in pre-intervention fit

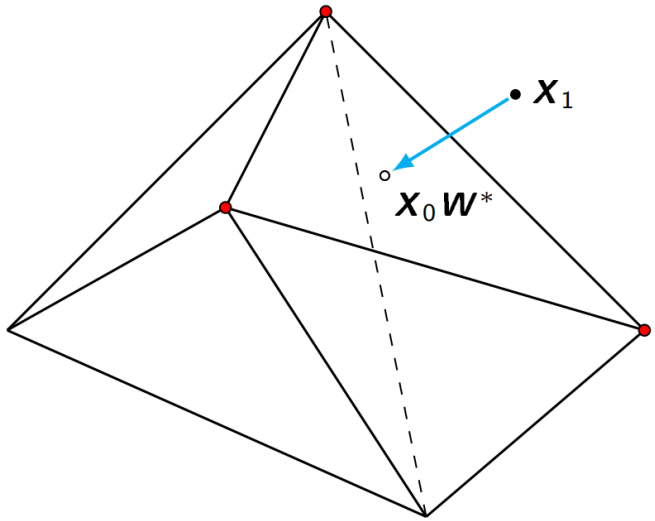
2. In practice, synthetic controls may not perfectly fit the characteristics of the treated units, how do we evaluate the validity of SCM?

$\Rightarrow$ Backdating exercise: backdate the intervention in the data set to a period with no intervention

- Conducting placebo SCM with the pre-treatment periods as the intervention
- Obtain an indication of the size and direction of the bias arising from imperfect fit

How to Choose $W^\star$?

Sparsity: Geometric interpretation



- In practice, X_1 does not belong to the convex hull of the columns of X_0 (the case in the figure), the synthetic control $X_0 W^\star$ is unique and sparse
- Otherwise, the weights may not be unique and sparse (Often the case with multiple treated units)
  - Different weights would offer different estimation results
  - How to choose $W^\star$? Sparsity is important for the interpretation

Penalized synthetic control (Abadie and L'Hour, 2020)

- How to force the sparse weights? Solve $W^\star(\lambda) \in \Delta_J$ for:

$$min_W \left\| X_1 - \sum_{j=2}^{J} W_j X_j \right\|^2 + \lambda \sum_{j=2}^{J} W_j \left\| X_1 - X_j \right\|^2$$

- $\lambda > 0$ 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
  - $\lambda \to \infty$: nearest neighbor matching
- For any $\lambda > 0$, the solution is unique and sparse:
  - The regularization term reduces the interpolation bias that occurs when averaging units that are far away from each other
- Same computational complexity as the unpenalized estimator

Issue with the benchmark X

- The benchmark choice is to use all the covariates for Z_i, and all previous outcomes for μ_i
- However, Kaul et al(2022): using all pre-treatment outcomes is more biased than using only one outcome-related predictor
  - Covariates are ignored when V is chosen to minimize the MSPE of the pre-treatment fit
- Moreover, the sparsity of the weights is controlled by the number of predictors k in X_1:
  - The number of nonzero weights bounded by k: it is the projection of X_1 on the convex hull of the columns of X_0

$\Rightarrow$ For Y_{jt}: use a summary measure or linear combination of multiple periods' outcome

$\Rightarrow$ For Z: Do not exclude them! Otherwise, increase the bias by increasing F: the number of unobserved factor loading

How to choose X?

- In practice, we might find a lot of potential combinations for X that have similar pre-treatment RMSPE
- Ferman et al (2020): you might find different X with a similar RMSPE gives totally different treatment effect!
  - Conduct robustness check for the choice of predictors of the outcome variable
  - You don't need outcomes from the post-treatment period to calculate weights, so don't conduct specification searches and p-hacking!
- How to choose X in practice? $\Rightarrow$ Evaluate the predictive power of sets of predictors
  1. Divides the pre-intervention periods into an initial training period and a subsequent validation period
  2. Synthetic control weights are computed using data from the training period only
  3. The validation period can then be used to evaluate the predictive power of the resulting synthetic control

Table of Contents

1. Introduction
2. Theoretical Justification
3. Implementation concerns in practice
4. Synthetic DID

Issue with a large number of pre-intervention periods

- Although we want $T_0 \to \infty$, it increase the possibility of structural breaks
- The linear model would be a good approximation only for a relatively short time scale
- To alleviate structural instability concerns, we might want to "up-weight" the most recent periods
- What weight should we use?
  - Intuitively, we can use the control group to find a weighted average of pre-treatment time periods that predicts average treatment-period outcomes well
- How to combine these weights into SCM? How to formalize and automate a process to prevent p-hacking?

$\Rightarrow$ Motivate Synthetic DID (Arkhangelsky et al, 2021)

- STATA package **SDID** is available

Combining SCM with DID

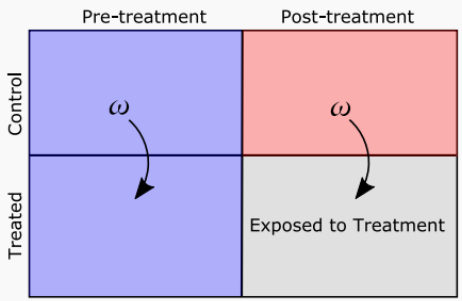
- Since we are now considering using the **average post-treatment outcome**, why don't we further extend it into a DID scenario?
  - Only interested in the ATT
  - Allowing more treated groups
- Need a modified version of synth. control weights: find a weighted average of control units with a pre-treatment trend **parallel** to the treated unit average:

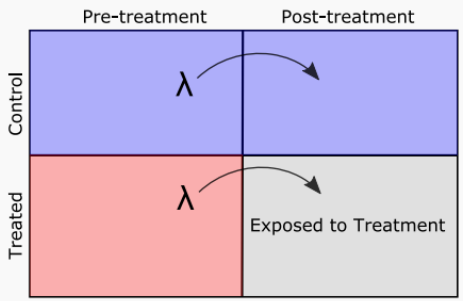
$$(\hat{\omega}_0, \hat{\omega}) = \underset{\substack{\omega_0 \in \mathbb{R} \\ \omega_1 \ldots \omega_{N_0} \geq 0 \\ \sum_{i \leq N_0} \omega_i = 1}}{\operatorname{argmin}} \frac{1}{T_0} \sum_{j \leq T_0} \left(\bar{Y}_{N_0+1:N,j} - \omega_0 - \sum_{i \leq N_0} \omega_i Y_{ij} \right)^2 + \zeta^2 \|\omega\|^2$$

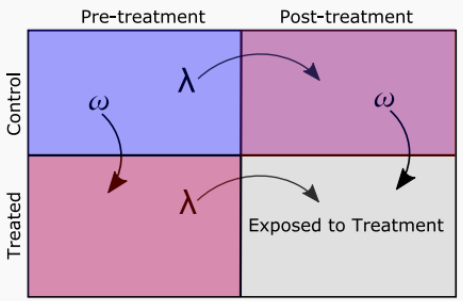
- Construct analogous time weights $\hat{\lambda}$

Intuition behind Synthetic DID By David Hirshberg

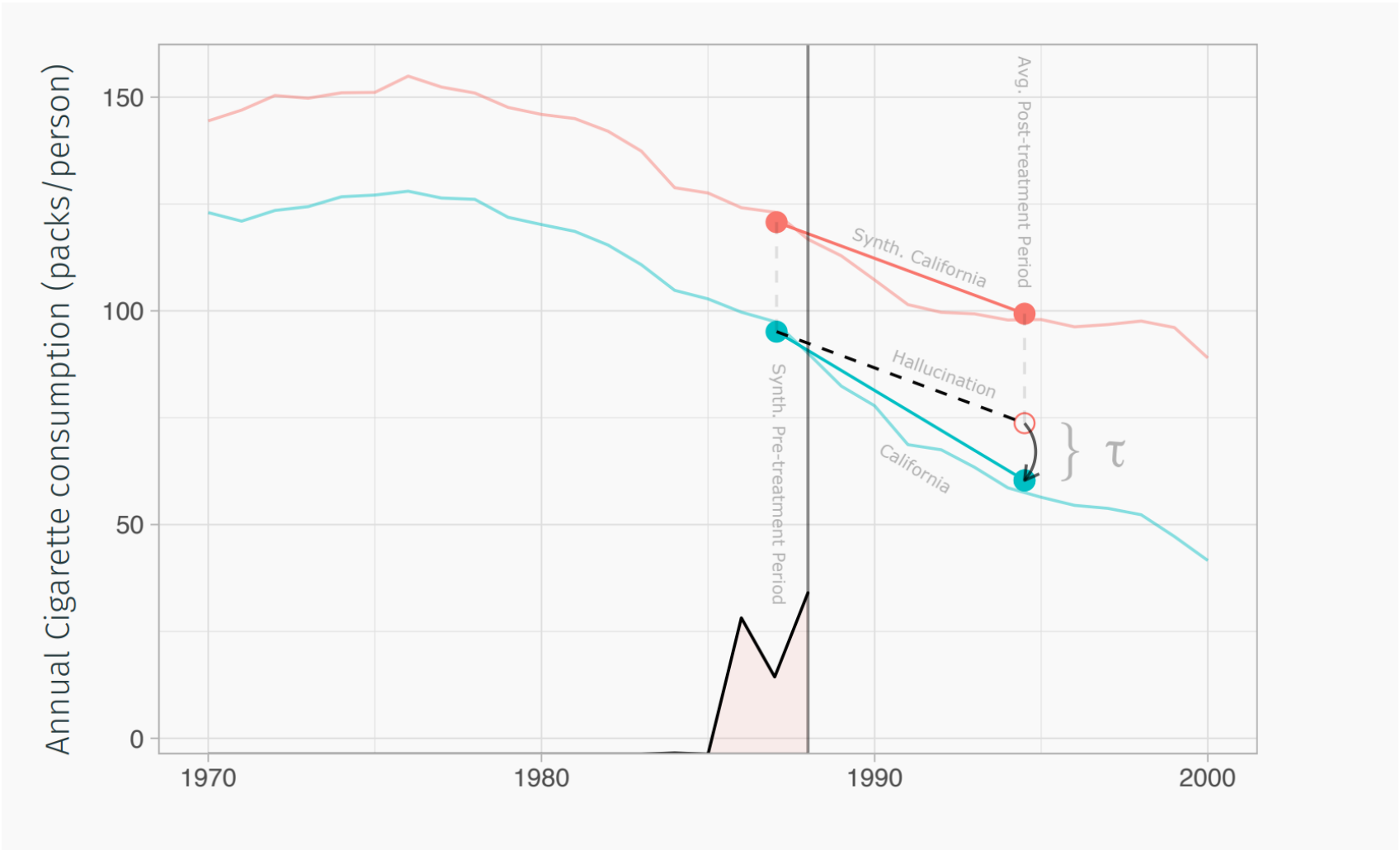
- Synthetic Control
  1. Using pre-treatment data, we learn an average of controls that's predictive of California.
  2. Assuming this relationship remain valid post-treatment, we use the same average of controls to impute treatment-free observations for California.
- Forecasting
  1. Using controls, we learn an average of periods forecasting what we see post-treatment.
  2. Assuming this relationship remain valid for the treated, we use the same average of periods to impute treatment-free observations for California.
- Synthetic Diff-in-Diff
  1. We do both synthetic control and forecasting and combine via diff-in-diff.
  2. Only one of these relationships has to remain valid.
  3. Constant offsets get *differenced out*: our synthetic control can be *parallel to* California.







Event Study plot from the SDID



SDID generalizes DID and SC:

2×2 diff-in-diff with synth. control unit and synth. pre-treatment pd.

	synthetic pre-treatment	average post-treatment
synthetic control	$\sum_{i \le N_0} \sum_{j \le T_0} \hat{\omega}_i \hat{\lambda}_j Y_{ij}$	$\sum_{i \le N_0} \sum_{j > T_0} \hat{\omega}_i T_1^{-1} Y_{ij}$
average treated	$\sum_{i > N_0} \sum_{j \le T_0} N_1^{-1} \hat{\lambda}_j Y_{ij}$	$\sum_{i > N_0} \sum_{j > T_0} N_1^{-1} T_1^{-1} Y_{ij}$

- DID uses equal weights $\omega_i = 1/N_0, \lambda_j = 1/T_0$
- SC takes only one difference (uses zero time weights $\lambda_j = 0$)

Comparison of weights

state	weight_sdid	weight_sc
Colorado	0.07	0.07
Connecticut	0.13	0.07
Delaware	0.07	–
Idaho	–	0.05
Illinois	0.09	–
Kansas	0.03	–
Minnesota	0.04	–
Montana	0.04	0.14
Nebraska	0.06	0.01
Nevada	0.18	0.17
New Hampshire	0.10	0.03
New Mexico	0.04	0.12
Utah	0.04	0.28
Wisconsin	0.03	–

- 24 of 37 states have zero weights for both SDID and SC
- All but the last three pre-treatment periods have zero weight

Comparison of the results

DID	SC	SDID
-27.35	-20.14	-13.36

- Even though SC fit in before period is good, there are some deviations in last three periods that make SDID estimate deviate from SC estimate
- DID implicitly has poor fit in before period.

Practical Requirement for SC related method

1. Aggregate level data:
   - Limited Volatility of the Outcome: $|\epsilon_{jt}|$ cannot be too large

   $\Leftarrow$ If substantial volatility is present, need to remove it via filtering
2. "Clean" comparison group: similar to the treated group; Not affected by other intervention
3. No anticipation and No interference: Similar to the assumptions in DID
4. "Balance" between the treated group and the synthetic treated group: need a perfect fit
5. Long enough Panel Time Horizon:
   - For the pre-treatment fit: Sufficient Pre-intervention Information

   $\Leftarrow$ Synthetic DID up-weight the most recent measures to bias from structural breaks
   - Also for the post-treatment effect analysis: Some effects might kick in slowly

Thank You!