

# Causal Inference III

## MIXTAPE SESSION

Prof. Scott Cunningham



# Roadmap

Introduction to course

- What is Mixtape Sessions?

- Potential outcomes review

Original synthetic control method

Imperfect fit

- Bounding the bias

- Demeaned Synthetic Control

- Augmented Synthetic Control

- Applying Machine Learning to Event studies from Finance

- Concluding remarks

Multiple Outcomes

Multiple Treated Units and Staggered

- Matrix completion with nuclear norm

- Synthetic difference-in-differences

- Synthetic Control with Staggered Adoption

# Welcome!

- I'm Scott Cunningham, professor of economics at Baylor University, author of Causal Inference: the Mixtape
- I teach and host workshops on causal inference all over the world because I believe many people do not have all the opportunities they want to have that they believe will help them be successful in their research and careers
- Workshops can be helpful ways to continue our education in methods that are otherwise inaccessible for whatever reason
  - Previous speakers include Alberto Abadie (MIT), Peter Hull (Brown), Jon Roth (Brown), Rocio Titunik (Princeton), Brigham Frandsen (BYU), Brantly Callaway (UGA), Jeff Gortmaker and Ariel Pakes (Harvard)

# What is Mixtape Sessions?

- Mixtape Sessions is my online platform started in November 2021 to “democratize causal inference” by helping connect people, from beginner to advanced, with material and teachers that for various reasons may not be accessible otherwise
- I tend to emphasize intuition, mechanics, narrow calculations, meaning, assumptions, code including actually taking time to code, advocate for data visualization – in other words the art and the science
- I decided to start teaching a workshop on synthetic control given it is considered one of the more important innovations in causal inference of the last twenty years (Athey and Imbens 2017)

# 2-day Causal Inference Workshop

- We workshop together for 2-days, 9am to 5pm CST with 15 min breaks on the hour and a 1-hour lunch break at noon CST
- I mix exposition, discussion of papers, coding exercises and discussion as best as I can
- I see this workshop as attempting to explain the method, explain some papers, but mainly towards providing guidance on projects using synthetic control, not simply the abstractions of the method

## What my pedagogy is like

- Long days that don't feel long because it's high energy, with regular breaks including lunch
- Move between the econometrics, applications, code, spreadsheets, exercises
- Ask questions at any point; I'll do my best to answer them and if I can't Kyle can

# Class goals

1. **Confidence:** You will feel like you have a good understanding of synthetic control so that by the end it doesn't feel all that mysterious or intimidating
2. **Comprehension:** You will have learned a lot both conceptually and in the specifics, particularly with regards to issues around identification and estimation
3. **Competency:** You will have more knowledge of programming syntax in Stata and R (and python!) so that later you can apply this in your own work

# Github repo

- We will communicate with one another regularly in the Discord channel and I will always be monitoring it
- Encourage you to talk to each other there, help one another, network with one another, coauthor with one another!
- I will be distributing things to you, like code and slides, via the github repo: <https://github.com/Mixtape-Sessions/Causal-Inference-3.git>
- Each lecture will be recorded and then uploaded to Vimeo as a password protected file that you'll have access to into perpetuity
- Kyle Butts and I are committed to over time making the Github Repository like an open public library where the only club goods are (a) recordings, (b) Discord and (c) live lectures

# Topics

1. Synthetic control traditionally without extrapolation
2. Synthetic control with extrapolation and bias adjustment
3. Aggregating across outcomes and noisy time series
4. Revisiting the event study from finance
5. Multiple treated units, staggered adoption, and synthetic diff-in-diff

# Causal Inference vs Prediction

**Figure 1:** Examples of popular data analysis algorithms in statistics and econometrics, as well as machine learning and artificial intelligence, classified according to prediction and causal inference methods. Causal inference methods are further differentiated according to observational (based on ex-post observed data) and experimental approaches.

Prediction		Causal Inference		Statistics/Econometrics	Machine Learning
		Observational			
ANOVA	Linear Regression	Difference-in-Differences	Instrumental Variables	A/B Testing	
Logistic Regression	Time Series Forecasting	Propensity Score Matching	Regression Discontinuity	Business Experimentation	
Boosting	Decision Trees & Random Forests	Additive Noise Models	Causal Forests	Randomized Controlled Trials	
Lasso, Ridge & Elastic Net	Neural Networks	Causal Structure Learning	Directed Acyclic Graphs	Causal Reinforcement Learning	
Support Vector Machines		Double/Debiased Machine Learning		Multiarmend Bandits	
				Reinforcement Learning	

# Causal Inference vs Prediction

## Traditional prediction

- Traditional prediction seeks to detect patterns in data and fit functional relationships between variables with a high degree of accuracy
- “Does this person have heart disease?”, “How many books will I sell?”
- It is not predictions of what effect a choice will have, though

## Causal inference

- Causal inference is also a type of prediction, but it's a prediction of a *counterfactual* associated with a particular *choice taken*
- Causal inference takes that predicted (or imputed) counterfactual and constructs a causal effect that we hope tells us about a future in the event of a similar choice taken

# Warranted Belief and Causal Inference

- The goal in causal inference is *psychological*
- When is a correlation *credibly* a causal effect?
- When is it a “warranted belief” to believe that one thing caused another when observed in quantitative data and analyzed with statistics?

# RCT as Thought Experiment

- It's helpful to start every project using the randomized controlled trial (RCT) as a thought experiment
  - Ask yourself if you had a million dollars and complete freedom, what *randomized* experiment would you run?
  - It helps you separate out pure description and pure prediction from causal identification
  - If you cannot articulate the randomization of something to a group of people, pause and start over

# Decomposition Formula

- If we compare the average outcome measured for different two groups – one treated, one not treated – then that "simple difference in mean outcomes" (SDO) is **always equal to**:

$$\begin{aligned} SDO &= \text{Average causal effect} \\ &\quad + \text{Selection Bias} \\ &\quad + \text{Heterogenous Treatment Effect Bias} \end{aligned}$$

- Selection bias means the two groups were always going to have different outcomes even if they hadn't been treated
- Heterogenous treatment effect bias means the two groups were also different with respect to their response to the treatment

# Decomposition Formula

- The decomposition formula is an identity, not a theorem – it is in other words **always true**
- One of the unique features of randomization is that since it is a data generating process independent of the features of the data, it makes the treatment and control group equal in large samples on all things
- That is, equal **except for the fact** that one group was treated and one wasn't
- This "equivalence in expectation" deletes selection bias and heterogenous treatment effect bias leaving only the first term – the average effect of the treatment

# RCT Framing Synthetic Control

- In many industry and policy applications, you simply cannot run the RCT you want
  - Maybe Uber wants to know the effect of some new self-driving program, but doesn't want to randomize it
  - It decides to roll it out in Austin, Texas only
  - Who is the control group? How will they be chosen?
- Synthetic control methods are usually candidates for causal inference in these situations because they work well with quasi-experimental policies at aggregate levels
- They also use data driven procedures to find optimal comparison groups

# Potential outcomes history

- Causality (metaphysics and philosophy) versus causal inference (epistemology and psychology)
- Should I believe that this one thing caused another thing – it's about credibility, evidence and beliefs
- We'll introduce briefly the notation we use throughout called potential outcomes which is a language rooted in a 1923 thesis by Jerzy Neyman, philosophical positions by JS Mill in the 1800s, and modern research agenda by Don Rubin
- Throughout the workshop, we will be focused on **binary treatment variables** applied to large populations

## Potential outcomes notation

- Let the treatment be a binary variable:

$$D_{i,t} = \begin{cases} 1 & \text{if exposed to autonomous vehicle fleet at time } t \\ 0 & \text{if not exposed to autonomous vehicle fleet at time } t \end{cases}$$

where  $i$  indexes an individual observation, such as a person, though for our workshop it'll be an aggregate region

## Potential outcomes notation

- Potential outcomes:

$$Y_{i,t}^j = \begin{cases} 1: \text{accidents at time } t \text{ if self-driving vehicles} \\ 0: \text{accidents at time } t \text{ if no self-driving vehicles} \end{cases}$$

where  $j$  indexes a hypothetical state of the world

# Treatment effect definitions

## Individual treatment effect

The individual treatment effect,  $\delta_i$ , equals  $Y_i^1 - Y_i^0$

Measures the effect of self-driving vehicles on accidents at the unit level.

Causal inference is a missing data problem in that no one knows the counterfactual because it is "missing"

Though here it is more "fictional" than simply "missing" as these missing data **do not exist anywhere**

# Conditional Average Treatment Effects

## Average Treatment Effect on the Treated (ATT)

The average treatment effect on the treatment group is equal to the average treatment effect conditional on being a treatment group member:

$$\begin{aligned} E[\delta | D = 1] &= E[Y^1 - Y^0 | D = 1] \\ &= E[Y^1 | D = 1] - \textcolor{red}{E[Y^0 | D = 1]} \end{aligned}$$

Example: The ATT is the averaging of the city's treatment effects who got the policy and it is the **only** causal effect both synthetic control *and* difference-in-differences can estimate

# Roadmap

Introduction to course

- What is Mixtape Sessions?

- Potential outcomes review

Original synthetic control method

Imperfect fit

- Bounding the bias

- Demeaned Synthetic Control

- Augmented Synthetic Control

- Applying Machine Learning to Event studies from Finance

- Concluding remarks

Multiple Outcomes

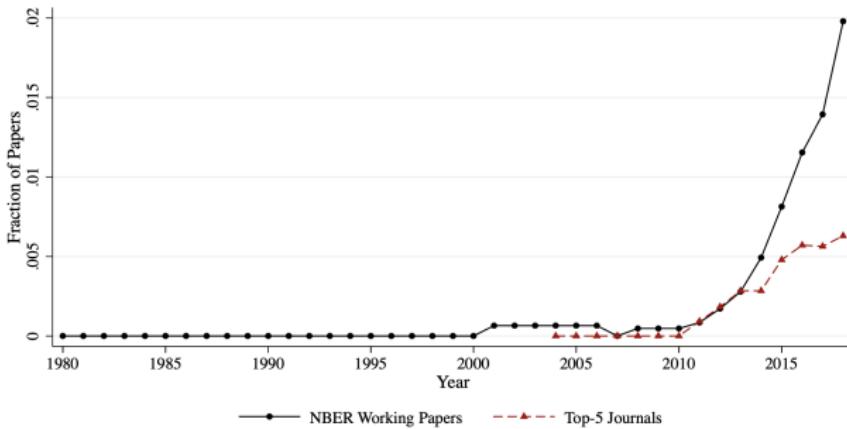
Multiple Treated Units and Staggered

- Matrix completion with nuclear norm

- Synthetic difference-in-differences

- Synthetic Control with Staggered Adoption

## D: Synthetic Control



# What is a comparative case study?

- Originally designed for comparative case studies, but now can accommodate multiple treated units, differential timing, bias adjustment, modifications to difference-in-differences
- Comparative case studies compare a single aggregate unit (like a country) to another unit (like another country) to make statements about causal effects from aggregate policies or events
- But the single treated unit can be any aggregate unit: a country, a state, school, firm, etc.

# What is a comparative case study?

Social scientists traditionally approached comparative case studies in one of two ways:

1. Qualitatively (political science)
2. Quantitatively (economics)

# First synthetic control paper

"The Economic Costs of Conflict: A Case Study of Basque Country"  
(2003) by Alberto Abadie and Javier Gardeazabal, *American Economic Review* (see /Readings at Github Repo)

*"About synthetic controls: I had played with related notions for some time, thinking about panel data methods to measure the effects of aggregate interventions. But the idea of synthetic controls shaped up in my mind with the Basque Country example. As always, taking vague concepts to data helps. There wasn't much one could do in terms of DD for the Basque example. When I saw what is now Figure 1 in the Basque paper appearing on my computer screen after the code finished running for the first time, the value of the method became immediately clear."*

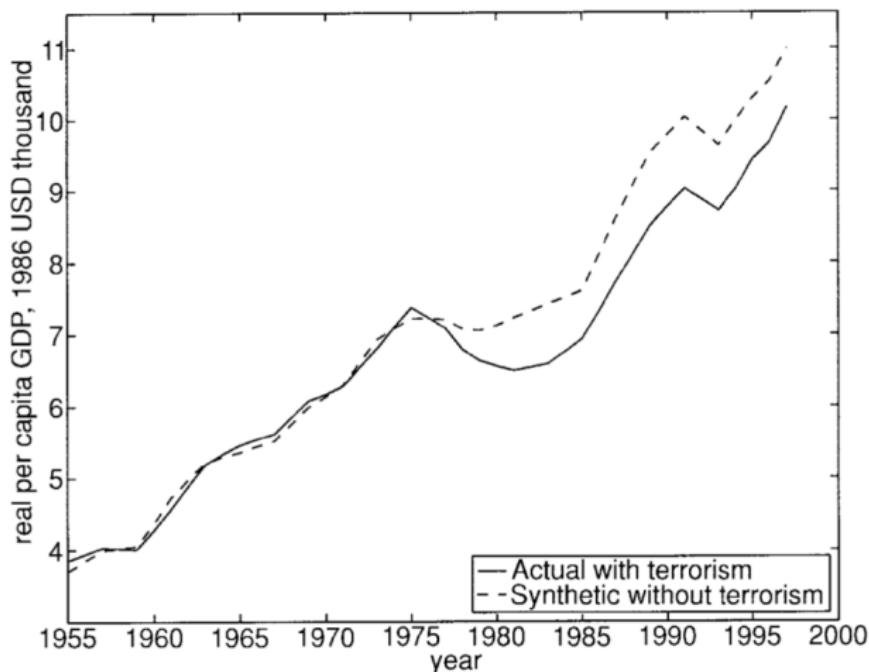


FIGURE 1. PER CAPITA GDP FOR THE BASQUE COUNTRY

## Synth in Abadie's own words

Approximately 5 minutes

<https://youtu.be/19cRm8aKF-I?si=dD13K60RjTt6E4EF&t=265>

## Qualitative comparative case studies

- In qualitative comparative case studies, the goal might be to reason *inductively* the causal effects of events or characteristics of a single unit on some outcome, oftentimes through logic and historical analysis.
  - Classic example of comparative case study approach is Alexis de Toqueville's Democracy in America (but he is regularly comparing the US to France)
- Sometimes there may not be an explicit counterfactual, or if there is, it's not principled (subjective researcher decision)
- Quantitative claims about causal effects are unlikely – de Toqueville's won't claim GDP per capita fell \$500 when compared against France

## Traditional **quantitative** comparative case studies

- Traditional **quantitative** comparative case studies are explicitly causal designs in that there is a treatment and control, usually involving natural experiment on a single aggregate unit
- Comparison focuses on the evolution of an aggregate outcome for the unit affected by the intervention to the evolution of the same *ad hoc* aggregate control group (Card 1990; Card and Krueger 1994)
- It'll essentially be diff-in-diff, but it may not use the event study, and the point is the choice of controls is a subset of all possible controls

## Pros of the traditional comparative case study

Pros:

- Takes advantage of policy interventions that take place at an aggregate level (which is common and so this is useful)
- Aggregate/macro data are often available (which may be all we have)

## Cons of the traditional comparative case study

Cons:

- Selection of control group is *ad hoc* – opens up researcher biases, even unconscious
- Standard errors do not reflect uncertainty about the ability of the control group to reproduce the counterfactual of interest

## Description of the Mariel Boatlift

- In 1980, Fidel Castro allowed anyone to leave Cuba so long as they did in the fall from the Mariel boat dock.
- The Mariel Boatlift brought 100,000 Cubans to Miami which increased the Miami labor force by 7%
- Card (1990) uses the Mariel Boatlift as a natural experiment to measure the effect of a sudden influx of immigrants on unemployment among less-skilled natives
- His question was how do inflows of immigrants affect the wages and employment of natives in local US labor markets?
- Individual-level data on unemployment from the Current Population Survey (CPS) for Miami and comparison cities







## Selecting control groups

- His treatment group was low skill workers in Miami since that's where Cubans went
- But which control group?
- He chose Atlanta, Los Angeles, Houston, Tampa-St. Petersburg

## Why these four?

Tables 3 and 4 present simple averages of wage rates and unemployment rates for whites, blacks, Cubans, and other Hispanics in the Miami labor market between 1979 and 1985. For comparative purposes, I have assembled similar data for whites, blacks, and Hispanics in four other cities: Atlanta, Los Angeles, Houston, and Tampa-St. Petersburg. These four cities were selected both because they had relatively large populations of blacks and Hispanics and because they exhibited a pattern of economic growth similar to that in Miami over the late 1970s and early 1980s. A comparison of employment growth rates (based on establishment-level data) suggests that economic conditions were very similar in Miami and the average of the four comparison cities between 1976 and 1984.

# Diff-in-diff

Differences-in-differences estimates of the effect of immigration on unemployment<sup>a</sup>

Group	Year			
	1979 (1)	1981 (2)	1981–1979 (3)	
Whites				
(1)	Miami	5.1 (1.1)	3.9 (0.9)	- 1.2 (1.4)
(2)	Comparison cities	4.4 (0.3)	4.3 (0.3)	- 0.1 (0.4)
(3)	Difference Miami-comparison	0.7 (1.1)	- 0.4 (0.95)	- 1.1 (1.5)
Blacks				
(4)	Miami	8.3 (1.7)	9.6 (1.8)	1.3 (2.5)
(5)	Comparison cities	10.3 (0.8)	12.6 (0.9)	2.3 (1.2)
(6)	Difference Miami-comparison	- 2.0 (1.9)	- 3.0 (2.0)	- 1.0 (2.8)

<sup>a</sup> Notes: Adapted from Card (1990, Tables 3 and 6). Standard errors are shown in parentheses.

## Parallel trends

- Card's analysis used diff-in-diff to estimate the impact of the Mariel boatlift on domestic labor markets
- His estimate is unbiased if  $\Delta E[Y^0]$  for the comparison cities approximates the counterfactual  $\Delta E[Y^0]$  for the treatment group
- Card selected his controls based on a mixture of matching logic (e.g., covariates) and trend logic (e.g., employment growth) but does not report much
- Black result would have been positive, too, were it not that the comparison cities growth was twice as large
- Creates uncertainty about his null result – was it no effect or was it the control group?

# Synthetic Control

- Abadie and Gardeazabal (2003) introduced synthetic control in the AER in a study of a terrorist attack in Spain (Basque Country) on GDP
- Revisited again in a 2010 JASA with Diamond and Hainmueller, two political scientists who were PhD students at Harvard (more proofs and inference)
- Basic idea is to use a combination of comparison units as counterfactual for a treated unit where the units are chosen according to a data driven procedure

## Researcher's objectives

- Simple terms: Find a weighted average of donor pool units that "look like" the treatment group before the treatment happened on *outcomes*
- More technical way of saying it:
  - Synthetic control is a constrained minimization problem where the target goal is the minimization of a vector of squared gaps in pre-treatment characteristics
  - Choice vector is an endogenous weights that are constant per control group unit over time and range from [0,1).
  - Our goal here is to reproduce the counterfactual of a treated unit by finding the combination of untreated units that best resembles the treated unit *before* the intervention in terms of the values of  $k$  relevant covariates (predictors of the outcome of interest)
- Method selects *weighted average of all potential comparison units* that best resembles the characteristics of the treated unit(s) - called the "synthetic control"

## Synthetic control method: advantages

- “Convex hull” means synth is a non-negatively weighted average of donor pool units that closely resemble the treatment group over time.
- Constraints on the model use non-negative weights which does not allow for extrapolation
- Makes explicit the contribution of each comparison unit to the counterfactual
- Formalizing the way comparison units are chosen has direct implications for inference

## Notation and setup

Suppose that we observe  $J + 1$  units in periods  $1, 2, \dots, T$

- Unit “one” is exposed to the intervention of interest (that is, “treated” during periods  $T_0 + 1, \dots, T$ )
- The remaining  $J$  are an untreated reservoir of potential controls (a “donor pool”)

## Potential outcomes notation

- Let  $Y_{it}^0$  be the outcome that would be observed for unit  $i$  at time  $t$  in the absence of the intervention
- Let  $Y_{it}^1$  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  $T$ .

## Group-time ATT with only one treated group

Treatment effect parameter is defined as dynamic ATT where

$$\begin{aligned}\delta_{1t} &= Y_{1t}^1 - Y_{1t}^0 \\ &= Y_{1t} - Y_{1t}^0\end{aligned}$$

for each post-treatment period,  $t > T_0$  and  $Y_{1t}$  is the outcome for unit one at time  $t$ . We will estimate  $Y_{1t}^0$  using the  $J$  units in the donor pool

## Optimal weights

- Let  $W = (w_2, \dots, w_{J+1})'$  with  $w_j \geq 0$  for  $j = 2, \dots, J + 1$  and  $w_2 + \dots + w_{J+1} = 1$ . Each value of  $W$  represents a potential synthetic control
- Let  $X_1$  be a  $(k \times 1)$  vector of pre-intervention characteristics for the treated unit. Similarly, let  $X_0$  be a  $(k \times J)$  matrix which contains the same variables for the unaffected units.
- The vector  $W^* = (w_2^*, \dots, w_{J+1}^*)'$  is chosen to minimize  $\|X_1 - X_0 W\|$ , subject to our weight constraints

Optimal weights differ by another weighting matrix

Abadie, et al. consider

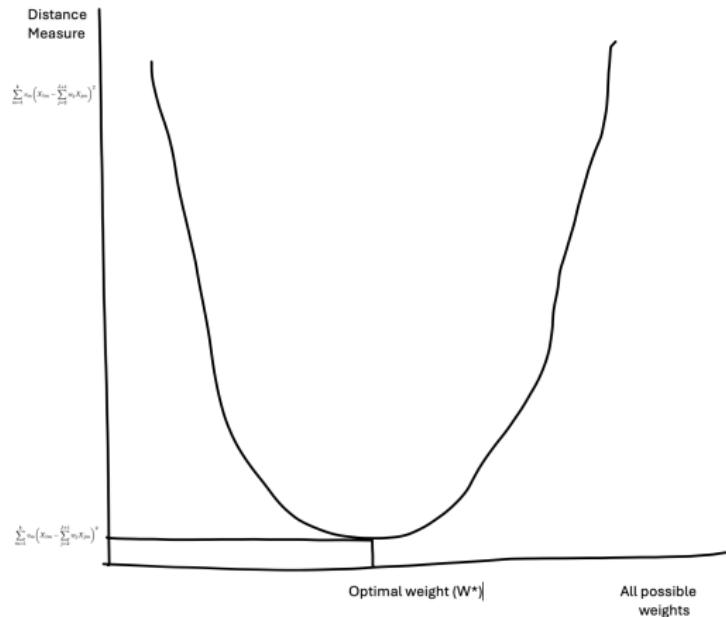
$$\|X_1 - X_0 W\| = \sqrt{(X_1 - X_0 W)' V (X_1 - X_0 W)}$$

where  $X_{jm}$  is the value of the  $m$ -th covariates for unit  $j$  and  $V$  is some  $(k \times k)$  symmetric and positive semidefinite matrix

## Similarity to distance minimization

- Bears some resemblance to nearest neighbor matching though I don't want to oversell that
- There is a unique solution that selects weighting minimizing the distance between the covariates comparison characteristics and treatment group
- Let's look at an example from nearest neighbor matching of minimizing Euclidean distance to help firm the ideas
- [https://docs.google.com/spreadsheets/d/1iro1Qzrr1eLDY\\_LJVz0YvnQZWmxY8JyTcDf6YcdhkwQ/edit?usp=sharing](https://docs.google.com/spreadsheets/d/1iro1Qzrr1eLDY_LJVz0YvnQZWmxY8JyTcDf6YcdhkwQ/edit?usp=sharing)

# Illustration of the "optimal weight"



## More on the V matrix

Typically,  $V$  is diagonal with main diagonal  $v_1, \dots, v_k$ . Then, the synthetic control weights  $w_2^*, \dots, w_{J+1}^*$  minimize:

$$\sum_{m=1}^k v_m \left( X_{1m} - \sum_{j=2}^{J+1} w_j X_{jm} \right)^2$$

where  $v_m$  is a weight that reflects the relative importance that we assign to the  $m$ -th variable when we measure the discrepancy between the treated unit and the synthetic controls

## Choice of $V$ is critical

- The synthetic control  $W^*(V^*)$  is meant to reproduce the behavior of the outcome variable for the treated unit in the absence of the treatment
- Therefore, the  $V^*$  weights directly shape  $W^*$

# Estimating the $V$ matrix

Choice of  $v_1, \dots, v_k$  can be based on

- Assess the predictive power of the covariates using regression
- Subjectively assess the predictive power of each of the covariates, or calibration inspecting how different values for  $v_1, \dots, v_k$  affect the discrepancies between the treated unit and the synthetic control
- Minimize mean square prediction error (MSPE) for the pre-treatment period (default):

$$\sum_{t=1}^{T_0} \left( Y_{1t} - \sum_{j=2}^J w_j^*(V^*) Y_{jt} \right)^2$$

## Cross validation

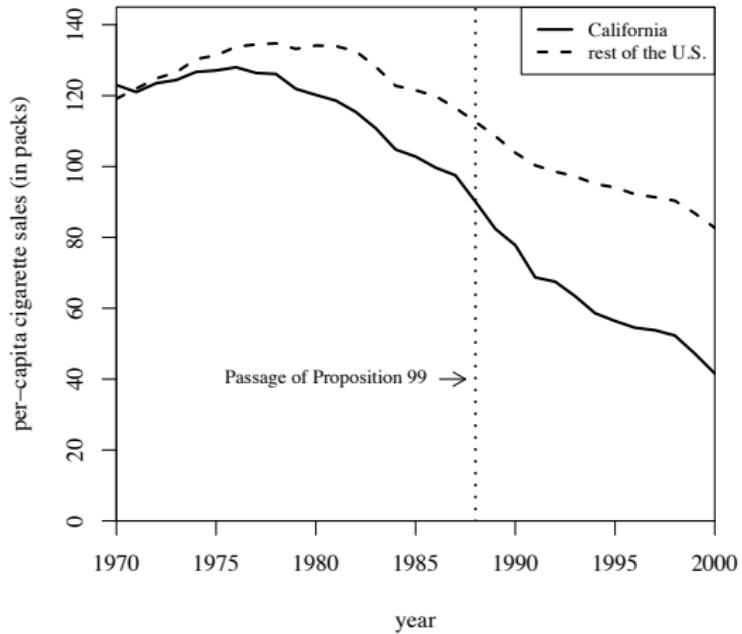
- Abadie recommends cross validation for selecting the covariates
- Divide the pre-treatment period into an initial **training** period and a subsequent **validation** period
- For any given  $V$ , calculate  $W^*(V)$  in the training period.
- Minimize the MSPE of  $W^*(V)$  in the validation period

Let's look at an example from the 2010 JASA paper with Hainmueller and Diamond

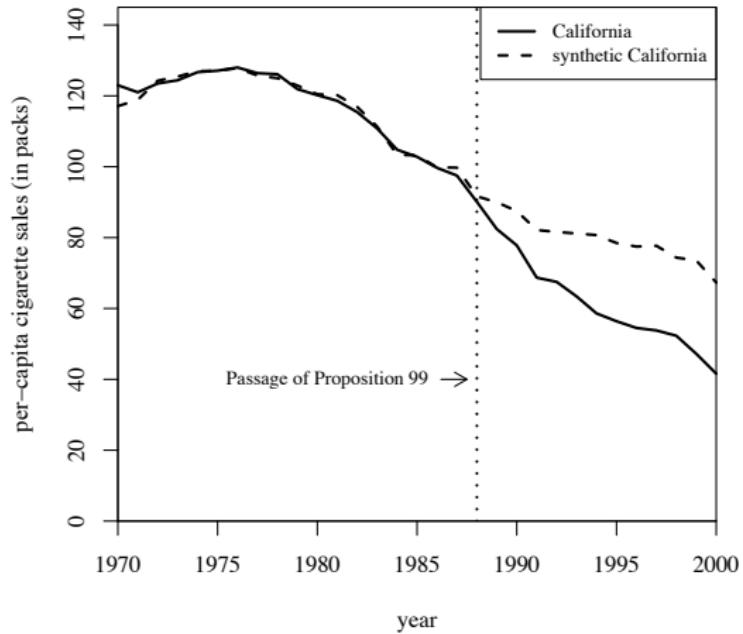
## Example: California's Proposition 99

- In 1988, California first passed comprehensive tobacco control legislation:
  - increased cigarette tax by 25 cents/pack
  - earmarked tax revenues to health and anti-smoking budgets
  - funded anti-smoking media campaigns
  - spurred clean-air ordinances throughout the state
  - produced more than \$100 million per year in anti-tobacco projects
- Other states that subsequently passed control programs are excluded from donor pool of controls (AK, AZ, FL, HI, MA, MD, MI, NJ, OR, WA, DC)

# Cigarette Consumption: CA and the Rest of the US



# Cigarette Consumption: CA and synthetic CA



# Sparsity and Synthetic Control Weights

Table 2. State weights in the synthetic California

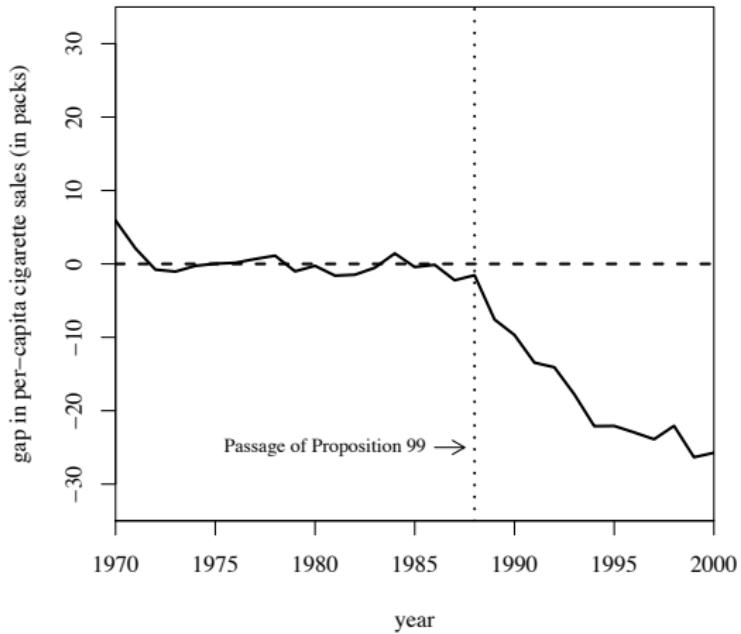
State	Weight	State	Weight
Alabama	0	Montana	0.199
Alaska	–	Nebraska	0
Arizona	–	Nevada	0.234
Arkansas	0	New Hampshire	0
Colorado	0.164	New Jersey	–
Connecticut	0.069	New Mexico	0
Delaware	0	New York	–
District of Columbia	–	North Carolina	0
Florida	–	North Dakota	0
Georgia	0	Ohio	0
Hawaii	–	Oklahoma	0
Idaho	0	Oregon	–
Illinois	0	Pennsylvania	0
Indiana	0	Rhode Island	0
Iowa	0	South Carolina	0
Kansas	0	South Dakota	0
Kentucky	0	Tennessee	0
Louisiana	0	Texas	0
Maine	0	Utah	0.334
Maryland	–	Vermont	0
Massachusetts	–	Virginia	0
Michigan	–	Washington	–
Minnesota	0	West Virginia	0
Mississippi	0	Wisconsin	0
Missouri	0	Wyoming	0

# Predictor Means: Actual vs. Synthetic California

Variables	Real	California Synthetic	Average of 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).

# Smoking Gap between CA and synthetic CA



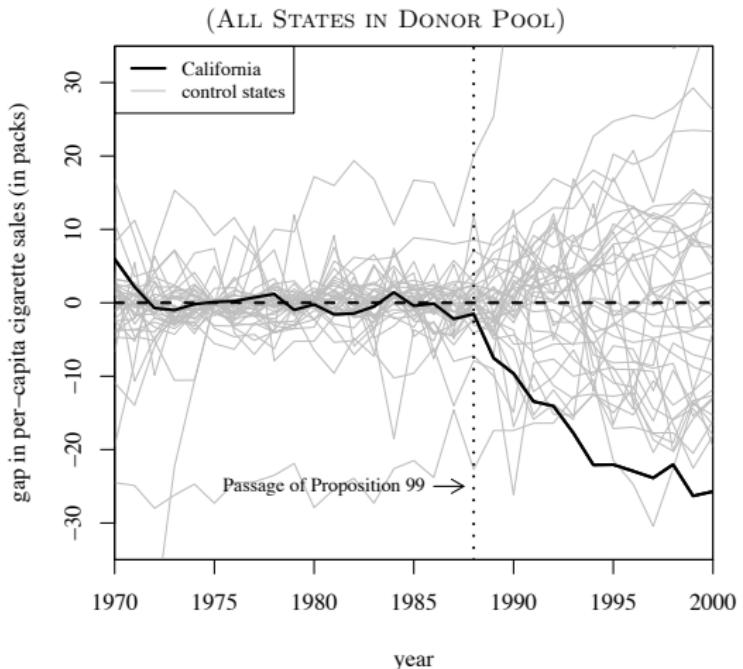
# Inference

- To assess significance, we calculate exact p-values under Fisher's sharp null using a test statistic equal to after to before ratio of RMSPE
- Exact p-value method
  - Iteratively apply the synthetic method to each country/state in the donor pool and obtain a distribution of placebo effects
  - Compare the gap (RMSPE) for California to the distribution of the placebo gaps. For example the post-Prop. 99 RMSPE is:

$$RMSPE = \left( \frac{1}{T - T_0} \sum_{t=T_0+1}^T \left( Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt} \right)^2 \right)^{\frac{1}{2}}$$

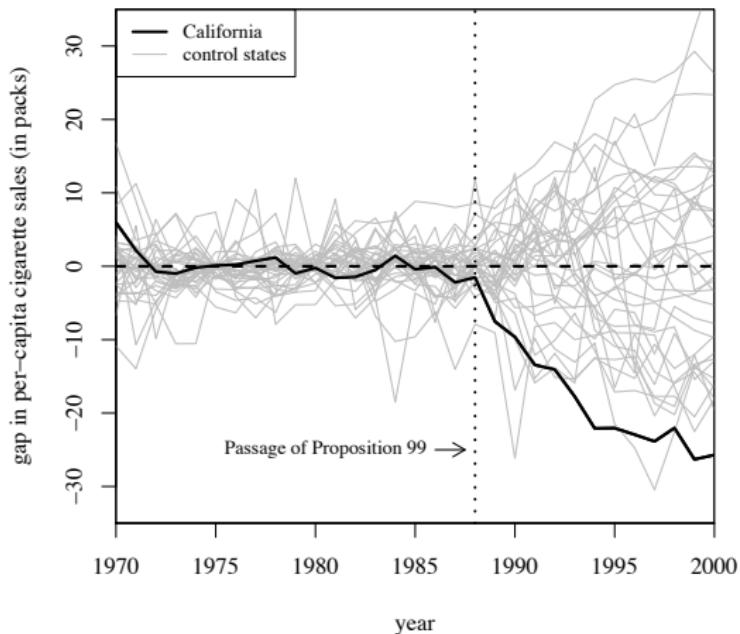
and the exact p-value is the treatment unit rank divided by  $J$

# Smoking Gap for CA and 38 control states



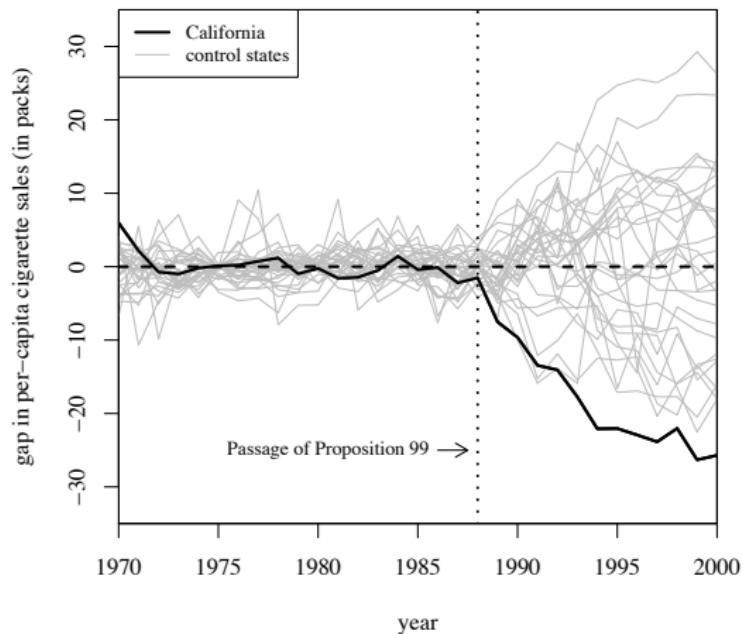
# Smoking Gap for CA and 34 control states

(PRE-PROP. 99 MSPE  $\leq$  20 TIMES PRE-PROP. 99 MSPE FOR CA)



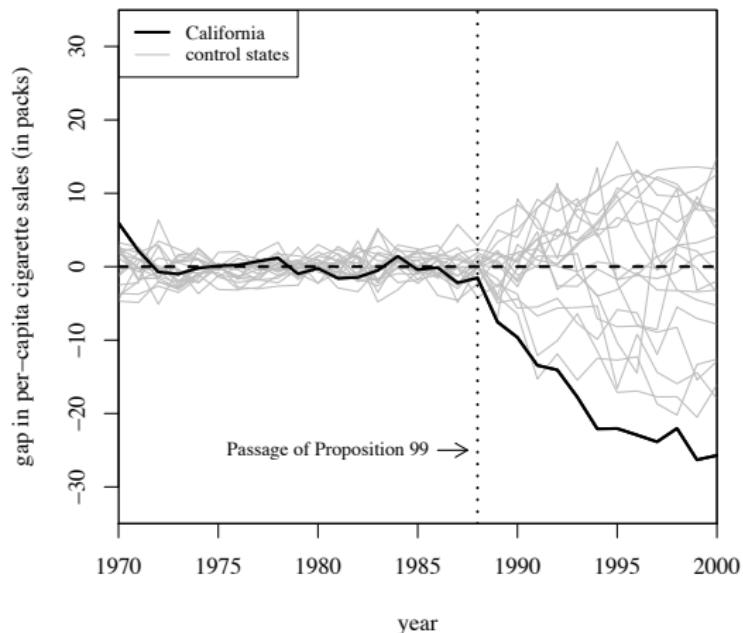
# Smoking Gap for CA and 29 control states

(PRE-PROP. 99 MSPE  $\leq$  5 TIMES PRE-PROP. 99 MSPE FOR CA)

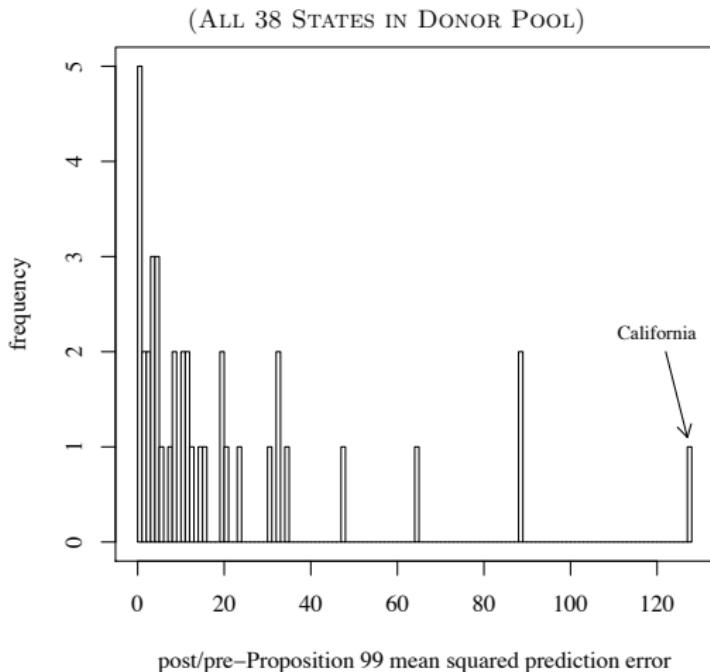


# Smoking Gap for CA and 19 control states

(PRE-PROP. 99 MSPE  $\leq$  2 TIMES PRE-PROP. 99 MSPE FOR CA)



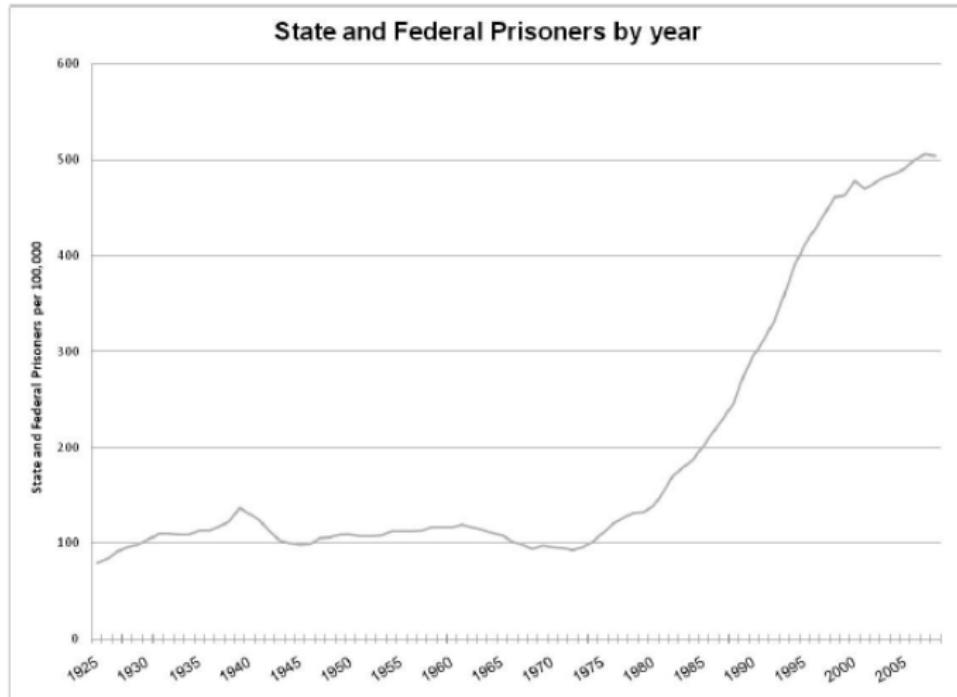
# Ratio Post-Prop. 99 RMSPE to Pre-Prop. 99 RMSPE



## Replication exercise

- The US has the highest prison population of any OECD country in the world
- 2.1 million are currently incarcerated in US federal and state prisons and county jails
- Another 4.75 million are on parole
- From the early 1970s to the present, incarceration and prison admission rates quintupled in size

Figure 1  
History of the imprisonment rate, 1925 - 2008



Source: [www.albany.edu/sourcebook/tost\\_6.html](http://www.albany.edu/sourcebook/tost_6.html)

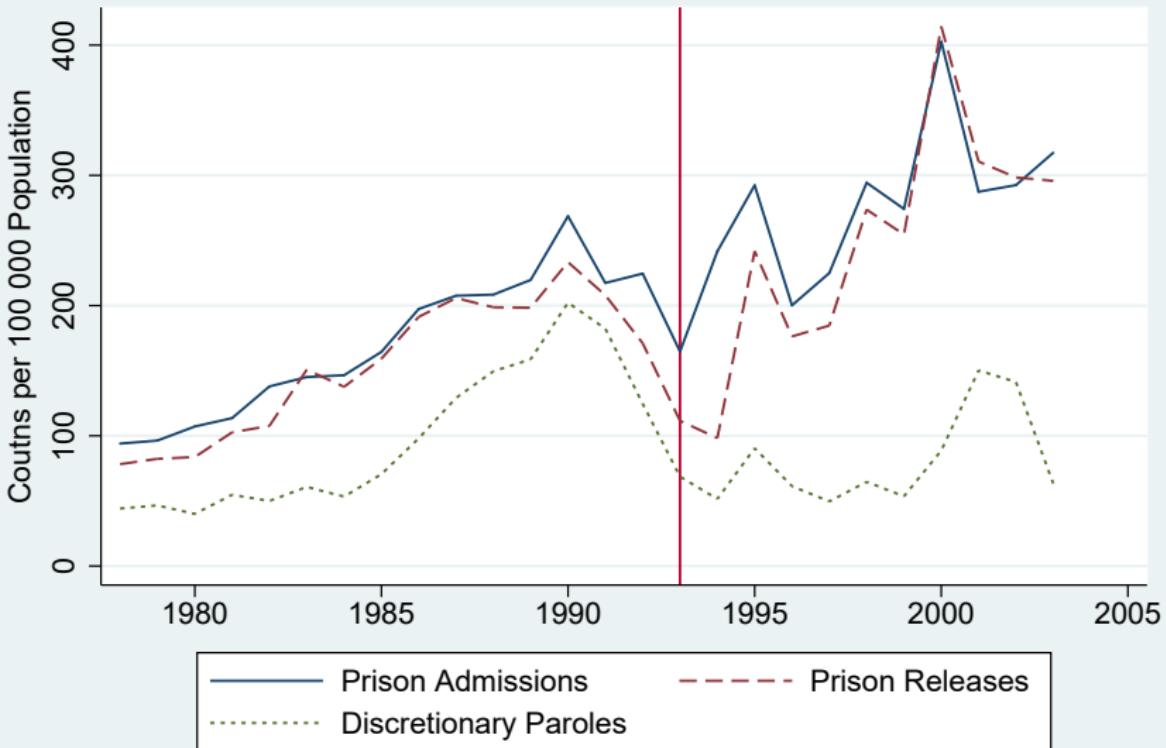
# Prison constraints

- Prisons are and have been at capacity for a long time so growth in imprisonment would bite on state corrections
- Managing increased flows can only be solved by the following:
  - Prison construction
  - Overcrowding
  - Paroles
- Texas chooses overcrowding

## Ruiz v. Estelle 1980

- Class action lawsuit against TX Dept of Corrections (Estelle, warden).
- TDC lost. Lengthy period of appeals and legal decrees.
- Lengthy period of time relying on paroles to manage flows

## Texas Prison Flows Measures per 100 000 Population

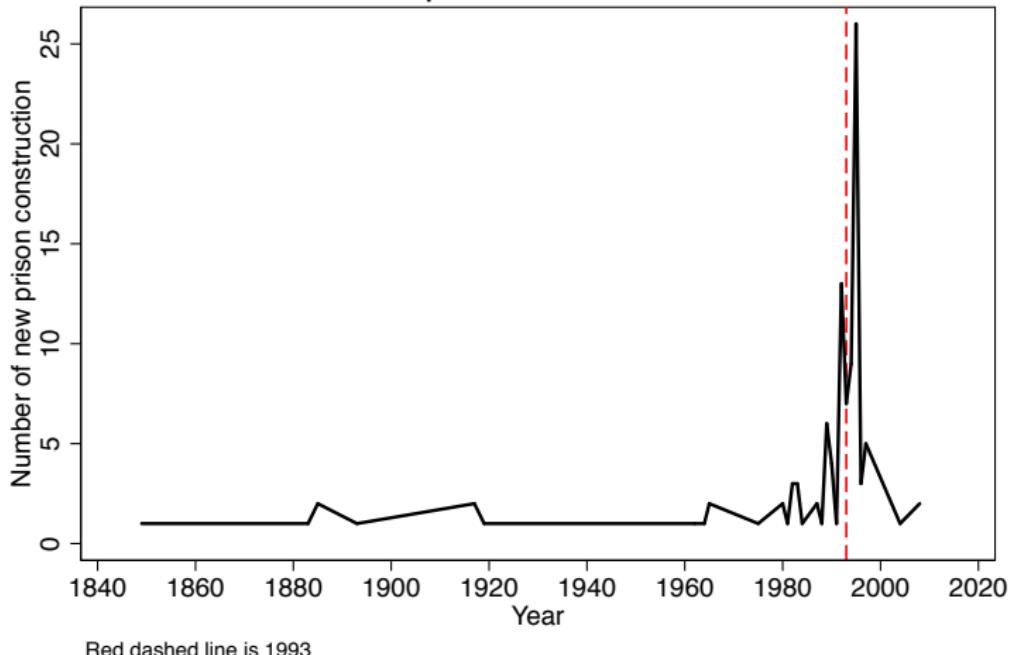


# Texas prison boom

Governor Ann Richards (D) 1991-1995

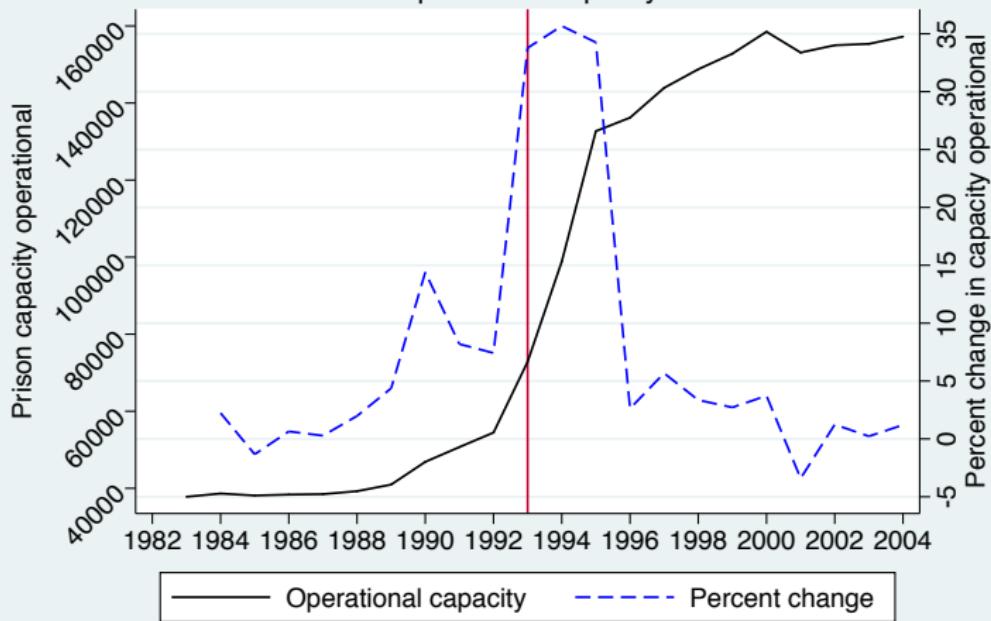
- Operation prison capacity increased 30-35% in 1993, 1994 and 1995.
- Prison capacity increased from 55,000 in 1992 to 130,000 in 1995.
- Building of new prisons (private and public)

## New prison construction

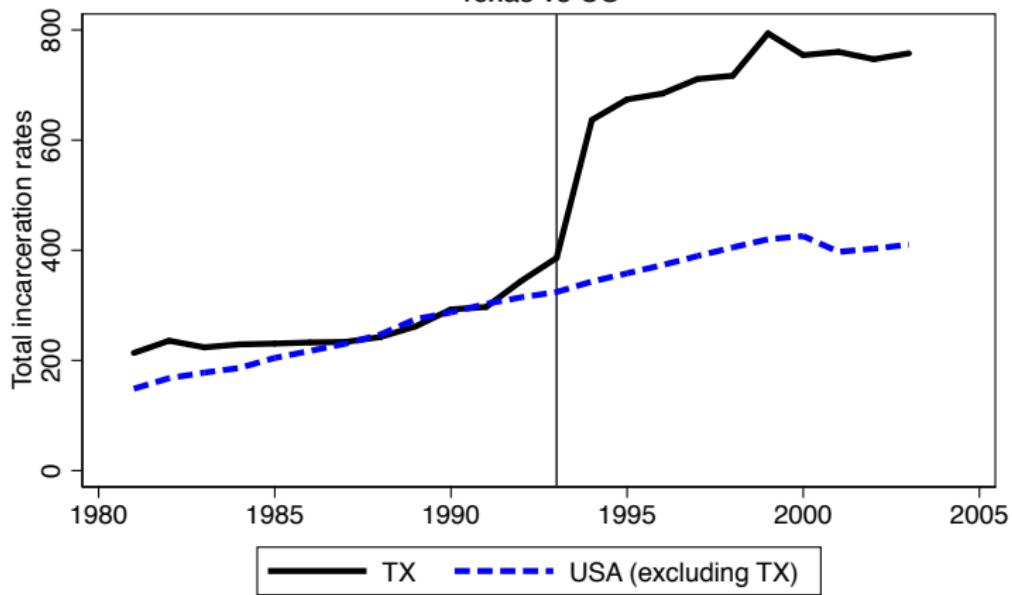


## Texas prison growth

### Operational capacity



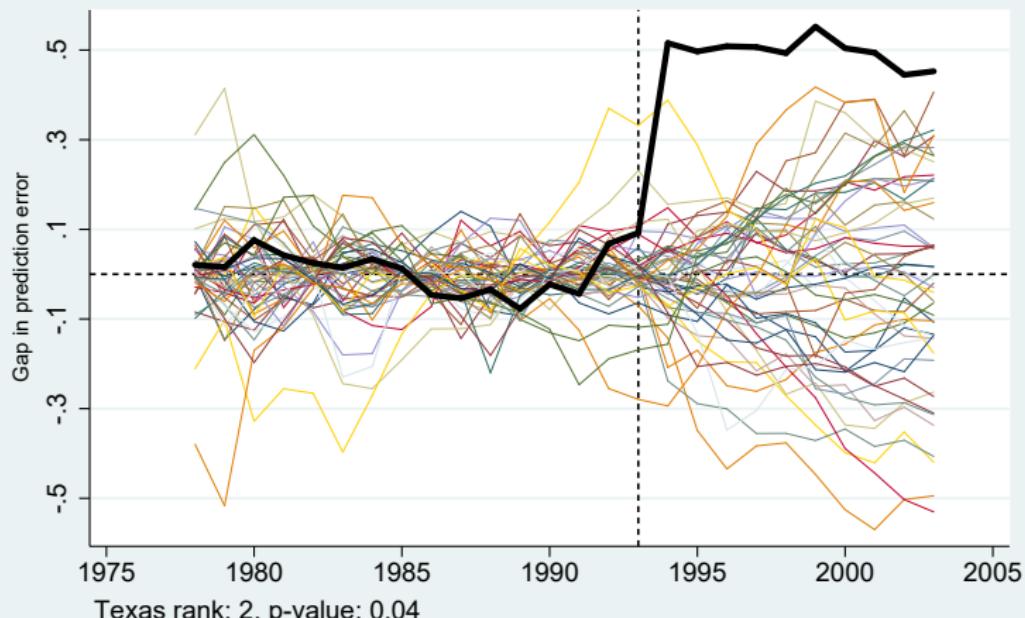
## Total incarceration per 100 000 Texas vs US



1993 starts the prison expansion

# Incarcerated persons per 100,000

1993 Treatment



## Coding together

- Let's go to Mixtape Sessions repository now into /Labs/Texas
- I'll walk us through the Stata and R code so you understand the syntax and underlying logic
- But then I have us a practice assignment

# Avoiding Cherry Picking: Subway franchise and Scandal

The Role of Repugnance in Markets: How the Jared Fogle Scandal Affected Patronage of Subway

John Cawley, Julia Eddelbuettel, Scott Cunningham, Matthew D. Eisenberg, Alan D. Mathios,  
and Rosemary J. Avery

NBER Working Paper No. 31782

October 2023

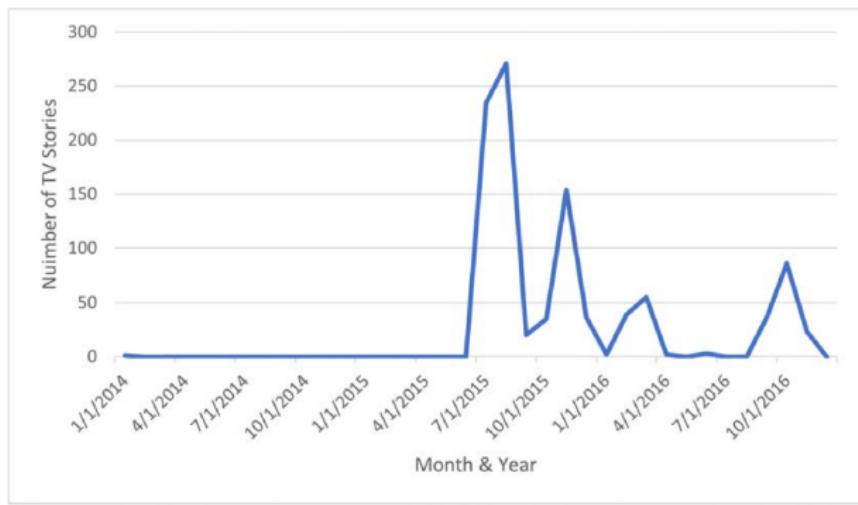
JEL No. D1,D22,D80,D83,I1,J1,K42,L2,L83,M2,M3,M5

## ABSTRACT

Economics has long studied how consumers respond to the disclosure of information about firms. We study a case in which the disclosed information is unrelated to the product or firm leadership, but which could still potentially affect consumer patronage through the mechanism of repugnance, as described in Roth (2007). The information in this case concerns the arrest of Jared Fogle, the advertising pitchman for the Subway sandwich franchise, who was arrested in 2015 on charges of sex with a minor and child pornography. We study how the disclosure of this information, which was widely covered in the media, affected patronage of Subway. We estimate synthetic control models using data from a large nationwide survey of consumers regarding the restaurants they patronize. Despite the close and long-standing association of Jared Fogle with Subway, and heavy publicity of his crimes, we consistently fail to detect any effect of the Jared Fogle scandal on the probability of visiting a Subway restaurant. These results contrast with past studies of negative information disclosure, which tend to find negative impacts on sales, revenue, or stock price of the relevant companies. The absence of an effect in this case suggests that repugnance did not drive demand, and that consumers largely separated the offenses of a symbol of the firm from the products of the firm.

# Subway franchise and Scandal

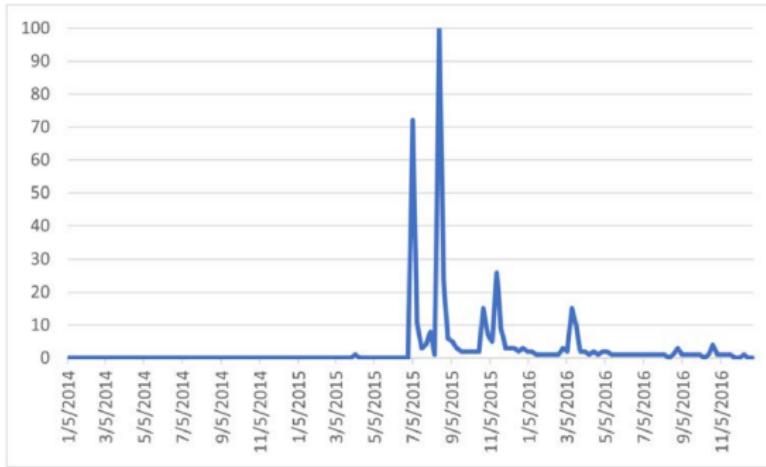
**Figure 1. Number of TV News Stories about “Jared Fogel”**



Note: Source: TV Archive database. Number of TV news stories matching search term “Jared Fogel” by month from January 2014 through December 2016.

# Subway franchise and Scandal

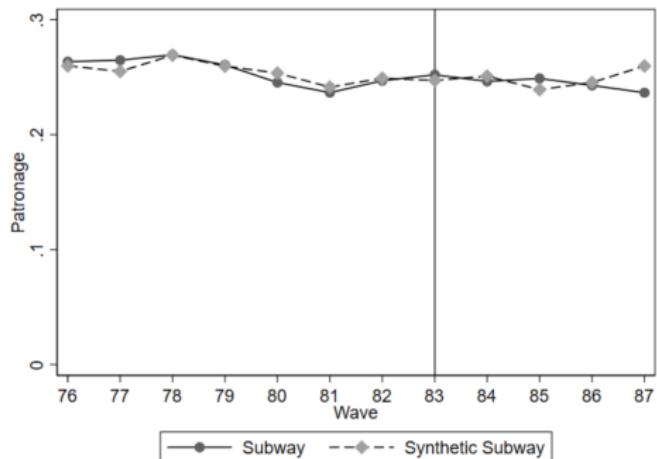
**Figure 2: Google Trends Searches of “Jared Fogle”**



Notes: Source: Google Trends. Number of Google searches for “Jared Fogel” by week from January 2014 through December 2016.

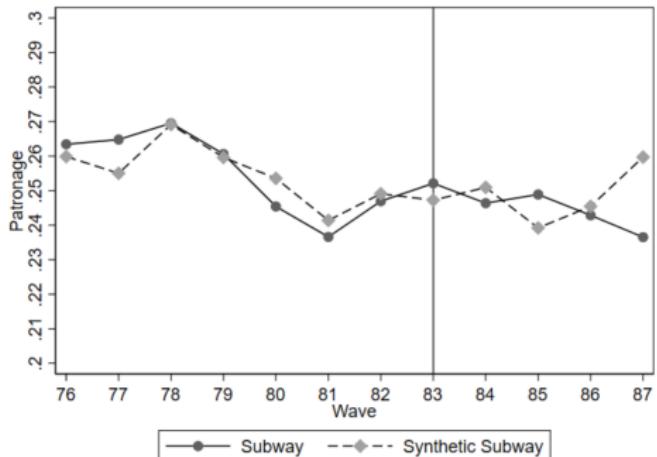
# Subway franchise and Scandal

**Figure 3a:** with Y axis (*Percent of Respondents Patronizing in the Past 30 Days*) ranging from 0 to 0.3



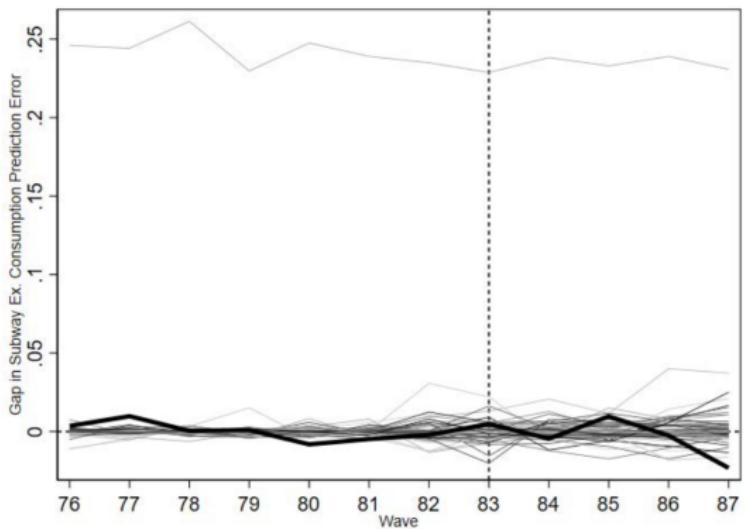
# Subway franchise and Scandal

**Figure 3b:** with Y axis (*Percent of Respondents Patronizing in the Past 30 Days*) ranging from 0.2 to 0.3



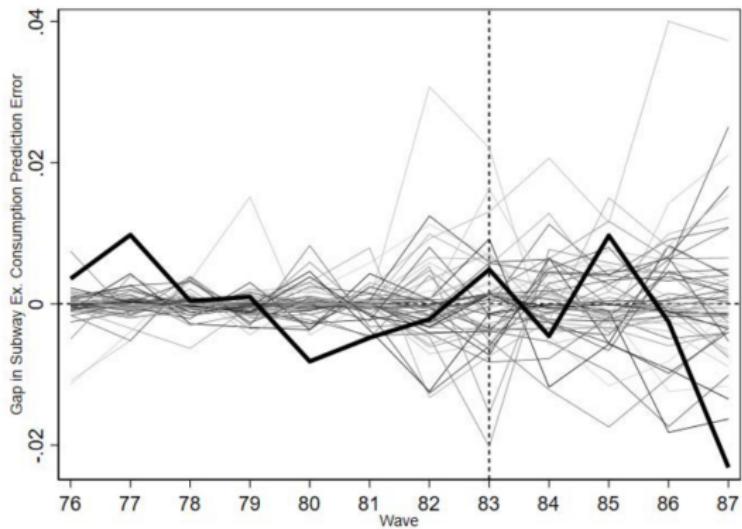
# Subway franchise and Scandal

*Figure 5a: RMSPE Gaps by Wave, Including McDonald's (Top Line)*



# Subway franchise and Scandal

**Figure 5b: RMSPE Gaps by Wave, Excluding McDonald's**



# Cherry picking synthetic controls

- Abadie, et al. recommended a nested minimization problem using the preintervention data to estimate the SC weights, but very little to no guidance was provided beyond that
- Historically, a broad range of selection of pre-treatment characteristics were used by early researchers which created researcher degrees of freedom or “cherry picking synthetic controls”
- Ferman, Pinto and Possbaum (2020) suggest specific specifications and report all of them with and without covariates
- Here’s an example of the kinds of tables and graphical evidence they suggested researchers follow (followed by what we did)

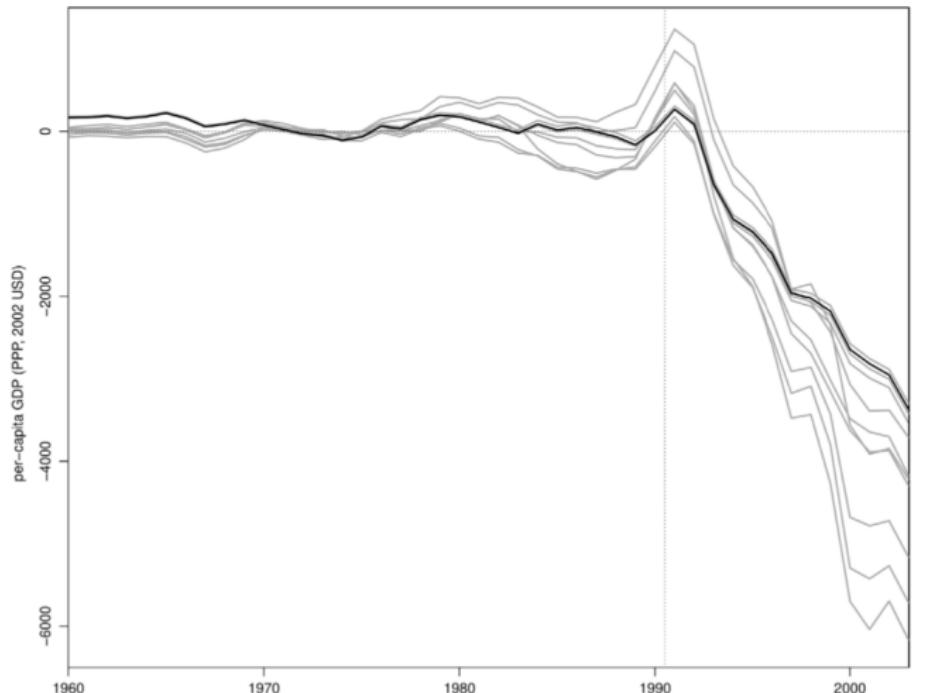
# Show p-values across all recommended specifications

**Table 3.** Specification searching—database from Abadie et al. (2015).

Specification	(1a)	(1b)	(2a)	(2b)	(3a)	(3b)	(4a)	(4b)
p-value	0.059	0.059	0.059	0.118	0.118	0.059	0.059	0.059
Specification	(5a)	(5b)	(6a)	(6b)	(7a)	(7b)		
p-value	0.118	0.059	0.588	0.059	0.353	0.059		

*Notes:* We analyze 14 different specifications. The number of the specifications refers to: (1) all pre-treatment outcome values, (2) the first three-fourths of the pre-treatment outcome values, (3) the first half of the pre-treatment outcome values, (4) odd pre-treatment outcome values, (5) even pre-treatment outcome values, (6) pre-treatment outcome mean (original specification by Abadie, Diamond, & Hainmueller, 2010), and (7) three outcome values. Specifications that end with an *a* do not include covariates, while specifications that end with a *b* include the covariates trade openness, inflation rate, industry share, schooling levels, and investment rate.

Show event study (not placebo) across all recommended specifications



*Notes:* The solid black line is the original specification by Abadie, Diamond, and Hainmueller (2015) and gray lines are specifications 1 through 5. The vertical line denotes the beginning of the post-treatment

# Subway franchise and Scandal

## ***Robustness Checks and Extensions***

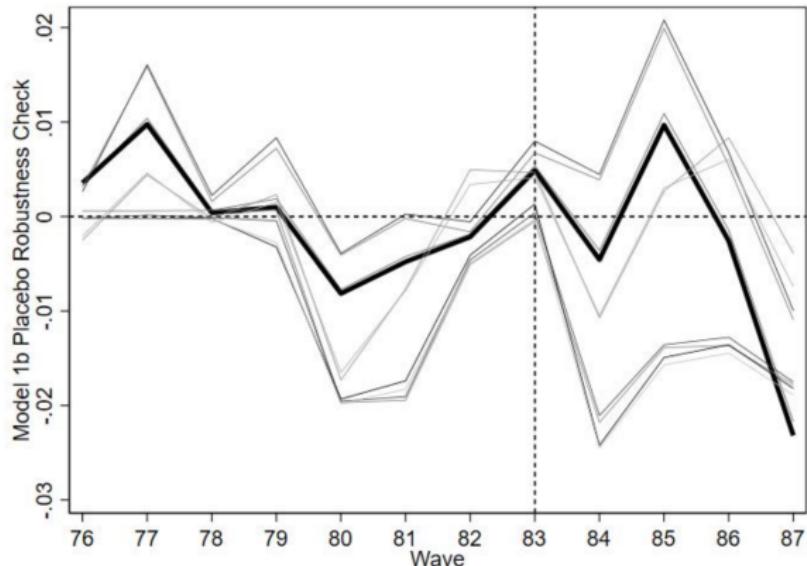
In order to examine the sensitivity of our results, we conduct several robustness checks. First, we examine the sensitivity of the results to the set of matching variables. Ferman, Pinto, and Possebom (2020) suggest estimating synthetic control models with specific sets of matching variables, in order to demonstrate that researchers didn't cherry-pick a set of matching variables in order to engineer a specific result. The sets that they use in their paper are:

1. All pre-treatment outcome values
  - a. With no other covariates
  - b. With other covariates
2. The first three-fourths of the pre-treatment outcome values
  - a. With no other covariates
  - b. With other covariates
3. The first half of pre-treatment outcome values
  - a. With no other covariates
  - b. With other covariates
4. Odd period pre-treatment outcome values
  - a. With no other covariates
  - b. With other covariates
5. Even period pre-treatment outcome values
  - a. With no other covariates
  - b. With other covariates

We follow Ferman et al. (2020) and re-estimate our synthetic control models with each of these

# Subway franchise and Scandal

**Figure 6: RMSPE Gaps by Wave, for Subway with Different Matching Variables**



Notes: Each line represents a different estimate for Subway, using the different specifications of matching variables in Ferman et al. (2020). The bold line corresponds to our primary model.

# Summarizing

- Method is simple: take donor pool units and find a combination of units, weight them, that minimize a distance function subject to two weight constraints  $W^*$  and  $V^*$ .
- But what is the bias and what can we do when we think the bias is severe?
- We now move towards that now by examining the nature of the bias in synthetic control

# Roadmap

Introduction to course

- What is Mixtape Sessions?

- Potential outcomes review

Original synthetic control method

Imperfect fit

- Bounding the bias

- Demeaned Synthetic Control

- Augmented Synthetic Control

- Applying Machine Learning to Event studies from Finance

- Concluding remarks

Multiple Outcomes

Multiple Treated Units and Staggered

- Matrix completion with nuclear norm

- Synthetic difference-in-differences

- Synthetic Control with Staggered Adoption

Suppose  $Y^0$  is given by a factor model

- What about unmeasured factors affecting the outcome variables as well as heterogeneity in the effect of observed and unobserved factors?
- Abadie, et al. assume that the missing potential outcome,  $Y_{it}^0$ , is generated using a factor model

$$Y_{it}^0 = \alpha_t + \theta_t Z_i + \lambda_t u_i + \varepsilon_{it}$$

# Factor Model Intuition

Untreated potential outcomes are given by a factor model:

$$Y_{it}^0 = \alpha_t + \theta_t Z_i + \lambda_t u_i + \varepsilon_{it} \quad (1)$$

- $Z_i$  is observed features,  $u_i$  is unobserved features,  $\varepsilon_{it}$  is unit-level transitory shock
- $\lambda_t$  is the set of factors (macroeconomic shock) at time  $t$ .
- $u_i$  is unit i's factor loading (exposure) to the shocks.

For example,  $\lambda_t$  is a shock to the returns to technological ability and  $u_i$  is unobserved ability.

# Factor Model Intuition

Untreated potential outcomes are given by a factor model:

$$Y_{it}^0 = \alpha_t + \theta_t Z_i + \lambda_t \textcolor{brown}{u}_i + \varepsilon_{it} \quad (2)$$

Suppose we found  $W^*$  such that:

$$\sum_{j=1}^{J+1} w_j^* Y_{jt} = Y_{1t}$$

$$\sum_{j=1}^{J+1} w_j^* Z_j = Z_1$$

and so on. This may only hold approximately in your sample. If you're able to do that, then you can establish a bound on the bias of the estimator.

# Synthetic Control Bias Under Factor Model

$$Y_{1t}^0 - \sum_{j=2}^{J+1} w_j^* Y_{jt} = \sum_{j=2}^{J+1} w_j^* \sum_{s=1}^{T_0} \lambda_t \left( \sum_{n=1}^{T_0} \lambda_n' \lambda_n \right)^{-1} \lambda_s' (\varepsilon_{js} - \varepsilon_{1s}) \\ - \sum_{j=2}^{J+1} w_j^* (\varepsilon_{jt} - \varepsilon_{1t})$$

- If  $\sum_{t=1}^{T_0} \lambda_t' \lambda_t$  is nonsingular, then RHS will be close to zero if number of preintervention periods is “large” relative to size of transitory shocks
- Only units that are alike in observables and unobservables should produce similar trajectories of the outcome variable over extended periods of time

## Synthetic control bounds

- Abadie, et al. 2010 shows the bounds on the bias and it is based on how closely you're able to fit the synthetic control to the treatment group over time
- Bias bound is controlled by a ratio between the scale of the transitory shocks,  $\varepsilon_{it}$ , and the time length
- Your credibility of a synthetic control depends on the extent to which it is able to fit a trajectory  $Y_{1t}$  for an extended pre-intervention period
- The idea is that on the shorter series, you're fitting noise ( $\varepsilon$ ) not the factor model

# Example of long $T_0$ and transitory shocks

Review of Economic Studies (2018) 85, 1683–1715

© The Author 2017. Published by Oxford University Press on behalf of The Review of Economic Studies Limited.

Advance access publication 20 December 2017

doi:10.1093/restud/rdx065

## Decriminalizing Indoor Prostitution: Implications for Sexual Violence and Public Health

SCOTT CUNNINGHAM

*Baylor University*

and

MANISHA SHAH

*University of California, Los Angeles & NBER*

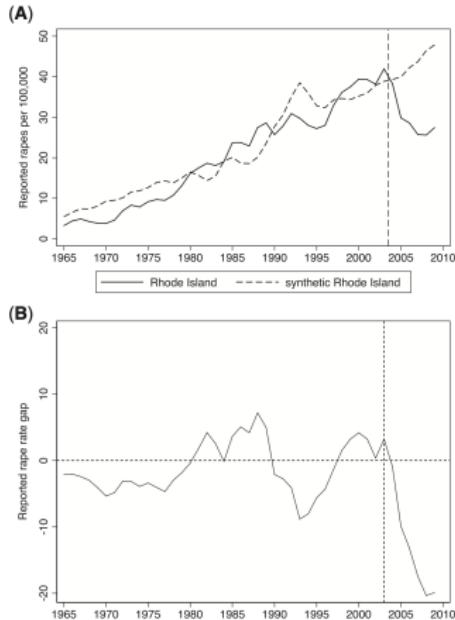
*First version received November 2015; Editorial decision August 2017; Accepted November 2017 (Eds.)*

Most governments in the world, including the U.S., prohibit sex work. Given these types of laws rarely change and are fairly uniform across regions, our knowledge about the impact of decriminalizing sex work is largely conjectural. We exploit the fact that a Rhode Island District Court judge unexpectedly decriminalized indoor sex work to provide causal estimates of the impact of decriminalization on the composition of the sex market, reported rape offences, and sexually transmitted infections. While decriminalization increases the size of the indoor sex market, reported rape offences fall by 30% and female gonorrhoea incidence declines by over 40%.

# Example of long $T_0$ and transitory shocks

1704

REVIEW OF ECONOMIC STUDIES



# Synthetic Control Bias Under Factor Model

Synthetic control will do well when you have many time-periods because:

1. You will not fit on the error term  $\varepsilon_{it}$
2. Instead, you will find a synthetic control with the same exposure to aggregate shocks (matching on factor-loadings)

But what about when we don't have good fit?

Equation (2) can hold exactly only if  $(Y_{11}, \dots, Y_{1T_0}, \mathbf{Z}'_1)$  belongs to the convex hull of  $\{(Y_{21}, \dots, Y_{2T_0}, \mathbf{Z}'_2), \dots, (Y_{J+11}, \dots, Y_{J+1T_0}, \mathbf{Z}'_{J+1})\}$ . In practice, it is often the case that no set of weights exists such that Equation (2) holds exactly in the data. Then, the synthetic control region is selected so that Equation (2) holds approximately. In some cases, it may not even be possible to obtain a weighted combination of untreated units such that Equation (3) holds approximately. This would be the case if  $(Y_{11}, \dots, Y_{1T_0}, \mathbf{Z}'_1)$  falls far from the convex hull of  $\{(Y_{21}, \dots, Y_{2T_0}, \mathbf{Z}'_2), \dots, (Y_{J+11}, \dots, Y_{J+1T_0}, \mathbf{Z}'_{J+1})\}$ . Notice, however, that the magnitude of such discrepancy can be calculated for each particular application. So for each particular application, the analyst can decide if the characteristics of the treated unit are sufficiently matched by the synthetic control. In some instances, the fit may be poor and then we would not recommend using a synthetic control.

## Synthetic controls with imperfect pretreatment fit

BRUNO FERMAN

Sao Paulo School of Economics-FGV

CRISTINE PINTO

Sao Paulo School of Economics-FGV

We analyze the properties of the Synthetic Control (SC) and related estimators when the pre-treatment fit is imperfect. In this framework, we show that these estimators are generally biased if treatment assignment is correlated with unobserved confounders, even when the number of pre-treatment periods goes to infinity. Still, we show that a demeaned version of the SC method can improve in terms of bias and variance relative to the difference-in-difference estimator. We also derive a specification test for the demeaned SC estimator in this setting with imperfect pre-treatment fit. Given our theoretical results, we provide practical guidance for applied researchers on how to justify the use of such estimators in empirical applications.

**KEYWORDS.** Synthetic control, difference-in-differences, policy evaluation, linear factor model.

**JEL CLASSIFICATION.** C13, C21, C23.

## Imperfect pre-treatment fit

- Ferman and Pinto (2021) show that synthetic control will generally be biased if treatment assignment is correlated with an unobserved confounder even in very long pre-treatment periods
- "This happens because, in this setting, the SC weights converge to weights that simultaneously attempt to match the factor loadings of the treated unit and to minimize the variance of a linear combination of the idiosyncratic shocks." (Ferman and Pinto 2021)
- So even if you reconstructed the pre-trends with your synth, you may not have recovered the true weights because you'd need a comparison group that was affected by the dynamic confounder in exactly the same way, just not treated

## Demeaning to address bias

- Ferman and Pinto (2021) show that the synthetic control estimator converges to the parameter we want *plus linear* combinations of contemporaneous idiosyncratic shocks and common factors
- If you assume the idiosyncratic factors cancel out in large samples, then the bias will depend on the differences in how the treated and the synthetic control units are affected by the common shocks
- They show that it is generally unbiased under the same conditions as diff-in-diff, and can have lower asymptotic variance, but they recommend showing both diff-in-diff and demeaned synthetic control

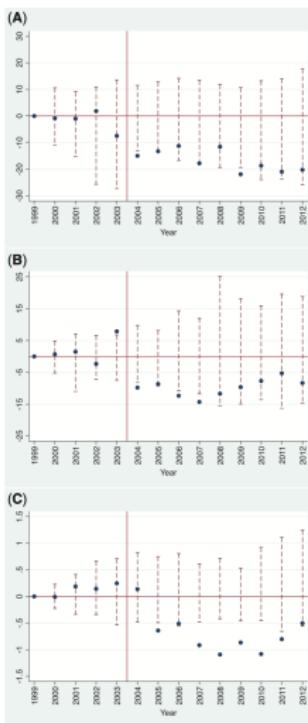


FIGURE 2

This figure plots the coefficients on Rhode Island-specific year effects ( $\beta_t$  from equation 2) for each outcome of interest. The solid vertical line denotes decriminalization. The dashed vertical lines are the sampling distribution for the placebo estimates from the 5th–95th percentile for each year. (A) Prostitution Arrests per 100,000; (B) Rape Offenses per 100,000; (C) Ln Gonorrhoea Cases per 100,000.

## Demeaning to address bias

- Cannot compare the original synthetic control with the demeaned synthetic control in terms of bias and variance – if units with similar factor loadings have similar fixed effects, then matching on levels would be better than demeaning
- Demeaning can increase the variance in finite pre-treatment
- Demeaning also implies extrapolating which is often considered to be one of the hallmark strengths of the original synthetic control estimator – it does not extrapolate

## What did we do?

- Our Rhode Island paper began in 2009 and was published in 2018, with working papers in circulation in 2014, well before this imperfect literature
- One of our main outcomes was reported female rape offenses per 100,000 as measured in the Uniform Crime Reports
- Rhode Island is a small state so even relatively modest changes year to year in reported rape offenses caused meaningful swings in the outcome series
- Our solution was to “smooth” using a moving average of each two years – which was not a demeaning, but rather simply  $\frac{Y_t+Y_{t+1}}{2}$

## More on imperfect fit

- Ferman and Pinto start this literature on imperfect fit, but it continued
- Next we examine relaxing the convexity requirement
- There had been an earlier paper by Doudchenko and Imbens that was never published that tried to add a constant so you can be off the convex hull
- But this next one will use negative weights

# Augmented Synthetic Control

- Synthetic control has built in constraints forcing weights to be non-negative
- Convex hull constraint ensures that synth is a feasible counterfactual in that it is formed by a combination of control units similar on pre-intervention characteristics
- Improves the validity of the estimated effect as there exists interpolated comparison group; similar to common support concept
- But, the convex hull constraint reduces extrapolation bias from comparing dissimilar units, but at the cost of failing to find matches at all

# What is augmented synthetic control?

- Eli Ben-Michael, Avi Feller and Jesse Rothstein present a modification to ADH in which they allow for negative weights, but only minimally so
- Model will “augment” the original synthetic control model by adjusting for pre-treatment imbalance using doubly robust bias adjustment (Abadie and Imbens 2011)
- Augmentation is conservative negative re-weighting using **penalized ridge regression** with constraints such that the negative weighting is only to the convex hull, not to the center of the convex hull (does not over-weight)

## Summarizing the argument

1. Original synthetic control needs perfect fit and so will be biased in practical settings as it won't be the case we get weights constrained to be on the convex hull
2. Augmentation of the synthetic control estimator uses an outcome model to estimate the bias caused by covariate imbalance
3. Outcome model is a penalized ridge regression which will provide new weights we use to reweight the original synthetic control (bias adjustment in the spirit of Abadie and Imbens (2011))

## Summarizing the argument

4. When synth is imbalanced, augmented synth will reduce bias reweighting and bias correction, and when synth is balanced, they are the same
5. When synth is balanced, the augmented and original synth are identical (but in practice, they won't be identical)
6. They argue synth DiD can be seen as a special case of augmented synth

# Notation

- Observe  $J + 1$  units over  $T$  time periods
- Unit 1 will be treated at time period  $T_0 = T - 1$  (we allow for unit 1 to be an average over treated units)
- Units  $j = 2$  to  $J + 1$  (using ADH original notation) are “never treated”
- $D_j$  is the treatment indicator

## Pre-treatment outcomes

$$\begin{pmatrix} Y_{11} & Y_{12} & Y_{13} & \dots & Y_{1T}^1 \\ Y_{21} & Y_{22} & Y_{23} & \dots & Y_{2T}^0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_{N1} & Y_{i2} & Y_{i3} & \dots & Y_{NT}^0 \end{pmatrix} \equiv \begin{pmatrix} X_{11} & X_{12} & X_{13} & \dots & Y_1 \\ X_{21} & X_{22} & X_{23} & \dots & Y_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ X_{N1} & X_{i2} & X_{i3} & \dots & Y_N \end{pmatrix} \equiv \begin{pmatrix} X_1 & Y_1 \\ X_0 & Y_0 \end{pmatrix}$$

This is a model of 2x2 (i.e., single last period block structure, not staggered roll out)

The last column is always post-treatment and switches from  $Y^1$  to  $Y$ .

The last column is just showing a top row of the treated unit 1 and the bottom row of all the donor pool (i.e., we will use  $X_0$  and  $Y_0$  to represent all the donor pool units)

## Optimal weights

Synth minimizes the following norm:

$$\begin{aligned} \min_w = & ||V_X^{1/2}(X_1 - X'_0 w)||_2^2 + \psi \sum_{D_j=0} f(w_j) \\ \text{s.t. } & \sum_{j=2}^N w_j = 1 \text{ and } w_j \geq 0 \end{aligned}$$

$Y'_0 w^*$  (i.e., optimally weighted donor pool) is the unit 1 “synthetic control”

# Predicting counterfactuals

Synth minimizes the following norm:

$$\begin{aligned} \min_w = & ||V_X^{1/2}(X_1 - X'_0 w)||_2^2 + \psi \sum_{D_j=0} f(w_j) \\ \text{s.t. } & \sum_{j=2}^N w_j = 1 \text{ and } w_j \geq 0 \end{aligned}$$

We are hoping that  $\widehat{Y}_1^0$  with  $Y'_0 w^*$  based on “perfect fit” pre-treatment

## $V_X$ matrix

Synth minimizes the following norm:

$$\begin{aligned} \min_w = & ||V_X^{1/2}(X_1 - X'_0 w)||_2^2 + \psi \sum_{D_j=0} f(w_j) \\ \text{s.t. } & \sum_{j=2}^N w_j = 1 \text{ and } w_j \geq 0 \end{aligned}$$

$V_x$  is the “importance” matrix on  $X_0$  (Stata default chooses  $V_x$  that min pre-treatment MSE).

## Penalizing the weights with ridge

Synth minimizes the following norm:

$$\begin{aligned} \min_w = & ||V_X^{1/2}(X_1 - X'_0 w)||_2^2 + \psi \sum_{D_j=0} f(w_j) \\ \text{s.t. } & \sum_{j=2}^N w_j = 1 \text{ and } w_j \geq 0 \end{aligned}$$

Modification to the original synthetic control model is the inclusion of the penalty term. “The choice of penalty is less central when weights are constrained to be on the simplex, but becomes more important when we relax this constraint.”

## Convex hull

Synth minimizes the following norm:

$$\begin{aligned} \min_w = & ||V_X^{1/2}(X_1 - X'_0 w)||_2^2 + \psi \sum_{D_j=0} f(w_j) \\ \text{s.t. } & \sum_{j=2}^N w_j = 1 \text{ and } w_j \geq 0 \end{aligned}$$

These weights will be used to address imbalance, not so much the control units, bc this method is for when the weighted controls are still outside the convex hull

## Original ADH factor model and bias

$$Y_{it}^0 = \alpha_t + \theta_t Z_i + \lambda_t u_i + \varepsilon_{it}$$

Original synth factor model (with ADH notation)

$$\begin{aligned} Y_{1t}^0 - \sum_{j=2}^{J+1} w_j^* Y_{jt} &= \sum_{j=2}^{J+1} w_j^* \sum_{s=1}^{T_0} \lambda_t \left( \sum_{n=1}^{T_0} \lambda'_n \lambda_n \right)^{-1} \lambda'_s (\varepsilon_{js} - \varepsilon_{1s}) \\ &\quad - \sum_{j=2}^{J+1} w_j^* (\varepsilon_{jt} - \varepsilon_{1t}) \end{aligned}$$

The bias of ADH synthetic control

## Perfect fit is necessary

$$\begin{aligned} Y_{1t}^0 - \sum_{j=2}^{J+1} w_j^* Y_{jt} &= \sum_{j=2}^{J+1} w_j^* \sum_{s=1}^{T_0} \lambda_t \left( \sum_{n=1}^{T_0} \lambda_n' \lambda_n \right)^{-1} \lambda_s' (\varepsilon_{js} - \varepsilon_{1s}) \\ &\quad - \sum_{j=2}^{J+1} w_j^* (\varepsilon_{jt} - \varepsilon_{1t}) \end{aligned}$$

Recall that the bias of ADH required “perfect fit” using their factor model  
(I’ll change  $\lambda$  factor loadings in a minute)

## Perfect fit models heterogeneity

$$\begin{aligned} Y_{1t}^0 - \sum_{j=2}^{J+1} w_j^* Y_{jt} &= \sum_{j=2}^{J+1} w_j^* \sum_{s=1}^{T_0} \lambda_t \left( \sum_{n=1}^{T_0} \lambda_n' \lambda_n \right)^{-1} \lambda_s' (\varepsilon_{js} - \varepsilon_{1s}) \\ &\quad - \sum_{j=2}^{J+1} w_j^* (\varepsilon_{jt} - \varepsilon_{1t}) \end{aligned}$$

Only units that are alike in observables and unobservables should produce similar trajectories of the outcome variable over extended periods of time

Remember that ADH15 quote

"The applicability of the [ADH2010] method requires a sizable number of pre-intervention periods. The reason is that the credibility of a synthetic control depends upon how well it tracks the treated unit's characteristics and outcomes over an extended period of time prior to the treatment. **We do not recommend using this method when the pretreatment fit is poor or the number of pretreatment periods is small.** A sizable number of post-intervention periods may also be required in cases when the effect of the intervention emerges gradually after the intervention or changes over time." (my emphasis, Abadie, et al. 2015)

## Slight change in synth notation

- Assume that our outcome,  $Y_{jt}^0$ , follows a factor model where  $m(\cdot)$  are pre-treatment potential outcomes:

$$Y_{jt}^0 = m_{jt} + \varepsilon_{jt}$$

- Since  $\widehat{m}(\cdot)$  estimates the post-treatment outcome, it can be viewed as an estimate of matching bias
- Procedure then becomes analogous to bias correction for inexact matching (Abadie and Imbens 2011)

## Bias correction

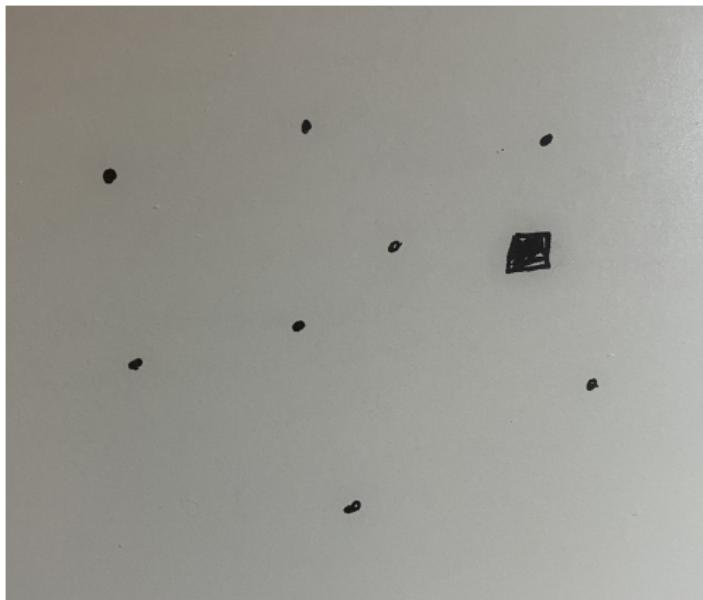
$$Y_{jt}^0 = m_{jt} + \varepsilon_{jt}$$

- Recall from earlier by Abadie, et al. (2010) and Ferman and Pinto (2021) the same point made which is that as  $T$  grows, the synthetic control achieves balance, not by fitting on the idiosyncratic noise (which is on average zero in large samples), but on the unobserved heterogeneity in the factor model
- Thus when the weights do achieve exact balance, the bias of synthetic control decreases with  $T$

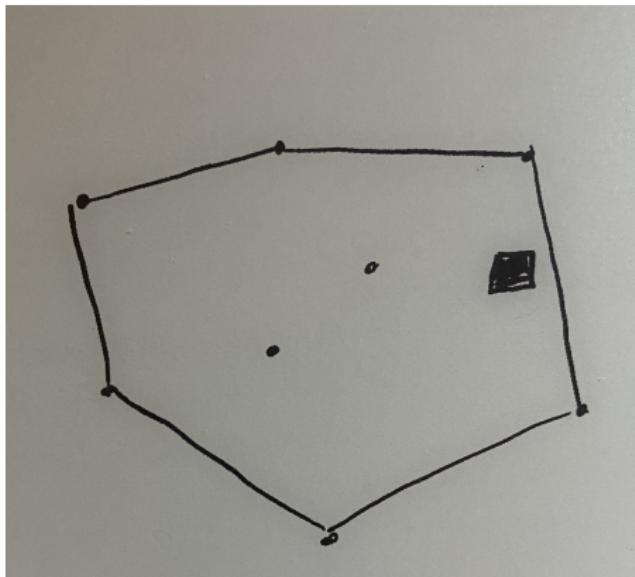
## Common practice

- Usually the number of time periods isn't much larger than the number of units
- And exact balance rarely holds, which if it doesn't hold, then the unobserved heterogeneity also doesn't get deleted even with large  $T$

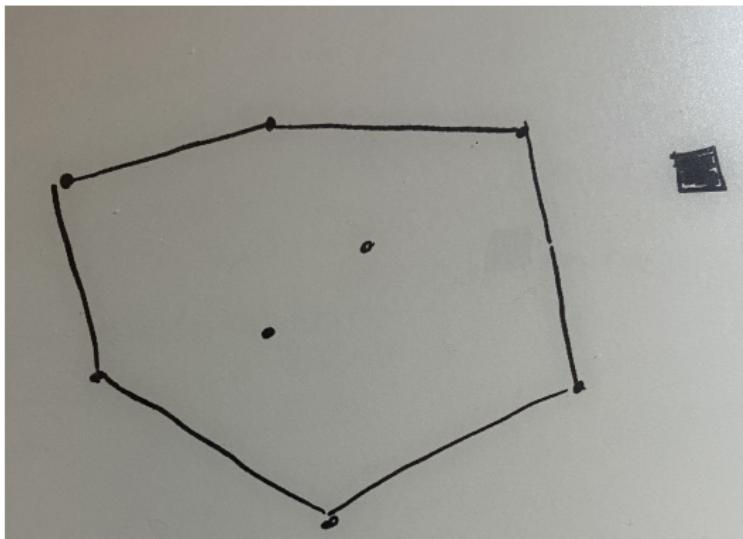
## Treatment and control units



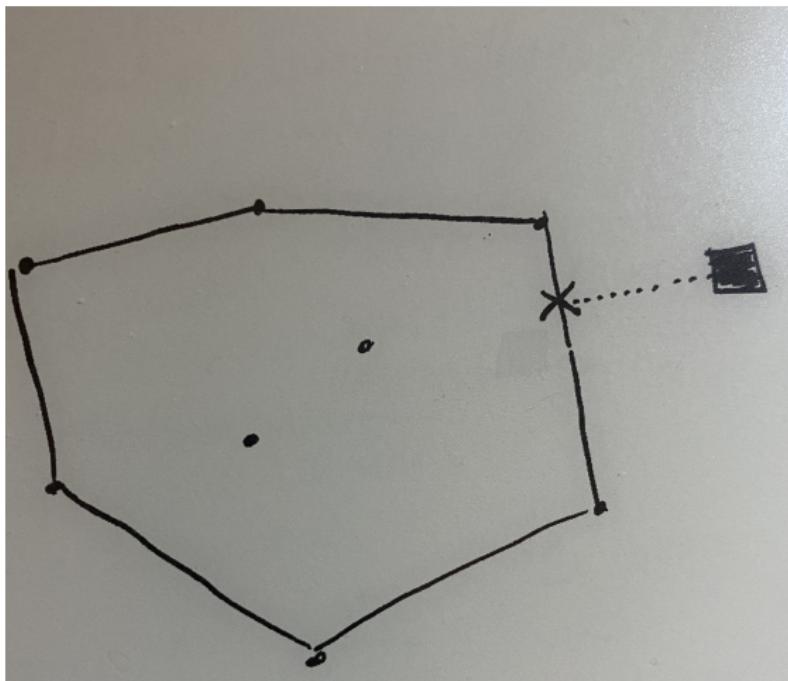
Convex hull – ideal for synth



Outside the convex hull bc of dimensionality



Outside the convex hull bc of dimensionality



## Estimating the bias

- Adjust the synthetic control to adjust for poor fit pre-treatment with an estimate of the “matching bias” for both the treatment group and the weighted average donor pools (Abadie and Imbens 2011)
- We will use our outcome regression model,  $\hat{m}_{jT}$ , to estimate the post-treatment potential outcome  $Y_{jT}^0$  which is recall unobserved for treatment group
- So there is in other words two steps involved: estimate the synthetic control finding optimal donor pool weights, then estimate the matching bias using ridge regression and adjust, similar to bias correction in Abadie and Imbens 2011

## Setup of the estimator

$Y_1^{aug,0}$  is the augmented potential outcome based on synthetic control (first term) and adjustments for matching imbalance (second and third terms):

$$\begin{aligned} Y_1^{aug,0} &= \sum_{D_j=0} \hat{w}_j^{synth} Y_j + \hat{m}(X_1) - \sum_{D_j=0} \hat{w}_j \hat{m}(X_j) \\ &= \hat{m}(X_1) + \sum_{D_j=0} \hat{w}_j (Y_j - \hat{m}(X_j)) \end{aligned}$$

## Interpreting line 1

$$\begin{aligned} Y_1^{aug,0} &= \sum_{D_j=0} \hat{w}_j^{synth} Y_{jT} + \left( \hat{m}_{1T} - \sum_{D_j=0} \hat{w}_j^{synth} \hat{m}_{jT} \right) \\ &= \hat{m}_{1T} + \sum_{D_j=0} \hat{w}_j^{synth} (Y_{jT} - \hat{m}_{jT}) \end{aligned}$$

- (1) Note how in the first line the traditional synthetic control weighted outcomes are corrected by the imbalance in a particular function of the pre-treatment outcomes  $\hat{m}$ .

## Interpreting line 1

$$\begin{aligned} Y_1^{aug,0} &= \sum_{D_j=0} \hat{w}_j^{synth} Y_{jT} + \left( \hat{m}_{1T} - \sum_{D_j=0} \hat{w}_j^{synth} \hat{m}_{jT} \right) \\ &= \hat{m}_{1T} + \sum_{D_j=0} \hat{w}_j^{synth} (Y_{jT} - \hat{m}_{jT}) \end{aligned}$$

- (1) Since  $\hat{m}$  estimates the post-treatment outcome, we can view this as an estimate of the bias due to imbalance, which is similar to how you address imbalance in matching with a bias correction formula (Abadie and Imbens 2011).

## Interpreting line 1

$$\begin{aligned} Y_1^{aug,0} &= \sum_{D_j=0} \hat{w}_j^{synth} Y_{jT} + \left( \hat{m}_{1T} - \sum_{D_j=0} \hat{w}_j^{synth} \hat{m}_{jT} \right) \\ &= \hat{m}_{1T} + \sum_{D_j=0} \hat{w}_j^{synth} (Y_{jT} - \hat{m}_{jT}) \end{aligned}$$

- (1) So if the bias is small, then synthetic control and augmented synthetic control will be similar because that interior term will be zero.

## Interpreting line 2

$$\begin{aligned} Y_1^{aug,0} &= \sum_{D_j=0} \hat{w}_j^{synth} Y_{jT} + \left( \hat{m}_{1T} - \sum_{D_j=0} \hat{w}_j^{synth} \hat{m}_{jT} \right) \\ &= \hat{m}_{1T} + \sum_{D_j=0} \hat{w}_j^{synth} (Y_{jT} - \hat{m}_{jT}) \end{aligned}$$

- (2) The second line is a double robust representation of the estimator equal to a regression based outcome model plus weighted residuals

## Form of outcome model

- To estimate  $\hat{m}$ , recall it is an extrapolation of  $Y^0$  based on covariates  $X$  (ignoring subscripts)
- But this can have overfitting problems, so they introduce a ridge regression
- Issue will be with regards to the hyperparameter so they'll suggest cross validation
- It's inside this second "reweighting" stage (or bias adjustment) that the negative weighting comes and it comes from the outcome model itself extrapolating

# Ridge Augmented SCM

$$\arg \min_{\eta_0, \eta} \frac{1}{2} \sum_{D_j=0} (Y_j - (\eta_0 + X'_j \eta))^2 + \lambda^{ridge} \|\eta\|_2^2$$

Here we estimate  $\hat{m}(X_j)$  with ridge regularized linear model and penalty hyper parameter  $\lambda^{ridge}$ . Sorry – this is not the same  $\lambda$ . I didn't create this notation though! Once we have those, we adjust for imbalance using the  $\hat{\eta}^{ridge}$  parameter as a weight on the outcome model itself.

# Ridge Augmented SCM

$$\arg \min_{\eta_0, \eta} \frac{1}{2} \sum_{D_j=0} (Y_j - (\eta_0 + X'_j \eta))^2 + \lambda^{ridge} \|\eta\|_2^2$$

Once we have those, we adjust for imbalance using the  $\hat{\eta}^{ridge}$  parameter as a weight on the outcome model itself.

Go back to that weighting but use the ridge parameters

$$\begin{aligned} Y_1^{aug,0} &= \sum_{D_j=0} \hat{w}_j^{synth} Y_j + \left( X_1 - \sum_{D_j=0} \hat{w}_j^{synth} X_j \right) \hat{\eta}^{ridge} \\ &= \sum_{D_j=0} \hat{w}_j^{aug} Y_j \end{aligned}$$

What you're trying to do is adjust with the  $\hat{w}_j^{aug}$  weights to improve balance.

The ridge weights are key to the augmentation

$$\hat{w}_j^{aug} = \hat{w}_j^{synth} + (X_j - X_0' \hat{w}_j^{synth})' (X_0' X_0 + \lambda I_{T_0})^{-1} X_i$$

The second term is adjusting the original synthetic control weights,  $w_j^{synth}$  for better balance. Again remember – we are trying to address the bias due to imbalance. You can achieve better balance, but at higher variance and can introduce negative weights.

Ridge will allow negative weights via extrapolation

$$\hat{w}_j^{aug} = \hat{w}_j^{synth} + (X_j - X_0' \hat{w}_j^{synth})' (X_0' X_0 + \lambda I_{T_0})^{-1} X_i$$

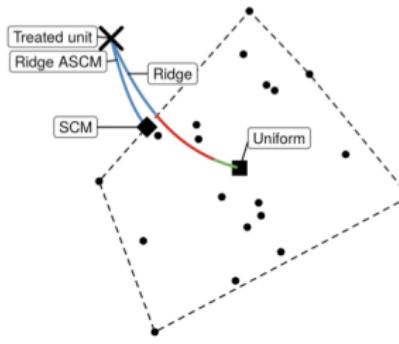
Relaxing the constraint from synth that weights be non-negative, as non-negative weights prohibit extrapolation. But we don't have synthetic control on the simplex, so we *must* extrapolate, otherwise synth will be biased.

## Summarizing and some comments

- When the treated unit lies in the convex hull of the control units so that the synth weights exactly balance lagged outcomes, then SCM and Ridge ASCM are the same
- When synth weights do not achieve exact balance, Ridge ASCM will use negative weights to extrapolate from the convex hull to the control units
- The amount of extrapolation will be determined by how much imbalance we're talking about and the estimated hyperparameter  $\hat{\lambda}^{ridge}$
- When synth has good pre-treatment fit or when  $\lambda^{ridge}$  is large, then adjustment will be small and the augmented weights will be close to the SCM weights

# Intuition

Ridge begins at the center of control units, while Ridge ASCM begins at the synth solution. Both move towards an exact fit solution as the hyperparameter is reduced. It is possible to achieve the same level of balance with non-negative weights. Both ridge and Ridge ASCM extrapolate from the support of the data to improve pre-treatment fit relative to synth alone. Let's look at a picture!



- In convex hull
- Out of convex hull
- Weights in simplex

(a) Treated and control units with the convex hull marked as a dashed line. Ridge and Ridge ASCM estimates in solid.

# Conformal Inference

Inference will be based on “conformal inference” method by Chernozhukov et al. (2019). We will get 95% point-wide confidence intervals. They also outline a jackknife method by Barber et al (2019).

# Steps of conformal Inference

- 1 Choose a sharp null (i.e., no unit-level treatment effects,  $\delta_0 = 0$ )
  - Enforce the null by creating an adjusted post-treatment outcome for the treated unit equal to  $Y_{1T} - \delta_0$  (in other words, we get CI on the post-treatment outcomes, not the pre-treatment)
  - Augment the original dataset to include the post-treatment time period  $T$  with the adjusted outcome and use the estimator to obtain the adjusted weights  $\widehat{w}(\delta_0)$
  - Compute a p-value by assessing whether the adjusted residual conforms with the pre-treatment residuals (see Appendix A for the exact formula)

# Steps of conformal Inference

- 2 Compute a level  $\alpha$  for  $\delta$  by inverting the hypothesis test (see Appendix A for the exact formula)
  - Chernozhukov et al. (2019) provide several conditions for which approximate or exact finite-sample validity of the  $p$ -values (and hence coverage of the predicted confidence intervals) can be achieved)

See Appendix A for more details

## Simulations (summarized)

- They examine the performance of synth against ridge, Augmented synth with ridge regularization, demeaned synth, and fixed effects under four DGP
- Augmenting synth with a ridge outcome regression reduces bias relative to synth alone in all four simulations
- This underscores the importance of the recommendation Abadie, et al. (2015) make which is that synth should be used in settings with excellent pre-treatment fit
- They also examine a real situation involving Kansas tax cuts in 2012

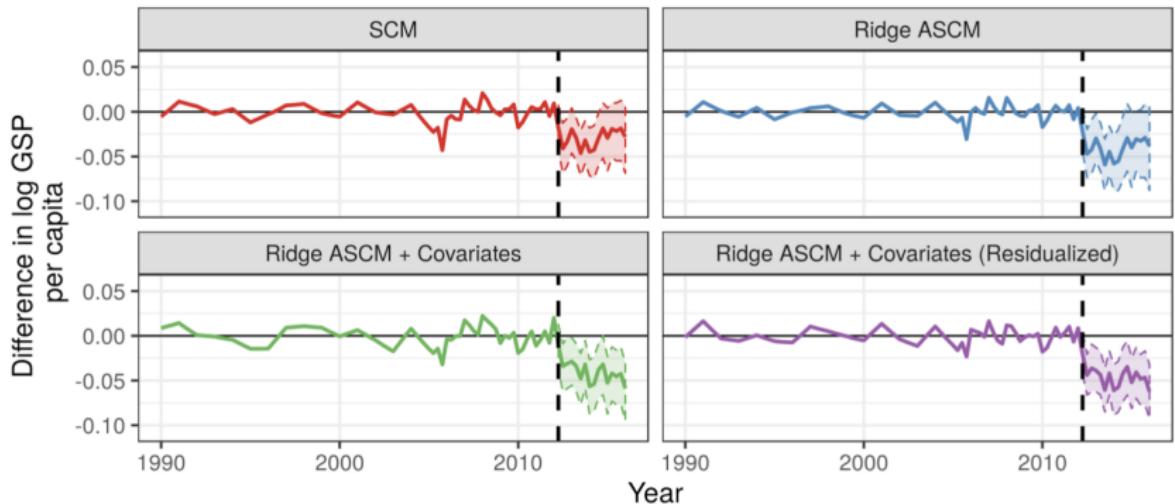
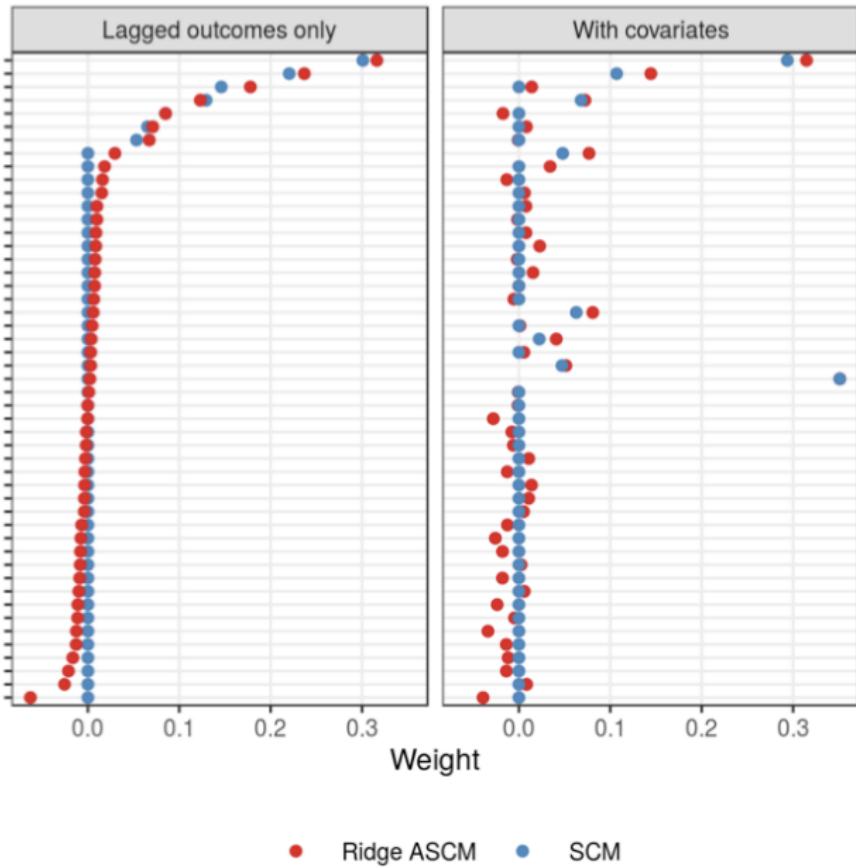


Figure 6: Point estimates along with point-wise 95% conformal confidence intervals for the effect of the tax cuts on log GSP per capita using SCM, Ridge ASCM, and Ridge ASCM with covariates.



## Couple of minor points

- Hyper parameter chosen using cross validation
- This can be extended to auxiliary covariates as opposed to just lagged outcomes (section 6)

## Augmented synth vs original synthetic control

- In conclusion, synthetic control is best when pre-treatment fit is excellent, otherwise it is biased
- Synthetic control avoids extrapolation by restricting weights to be non-negative and sum to one
- Ridge regression augmentation will allow for a degree of extrapolation to achieve pre-treatment balance and that creates negative weights
- Augmented synth will dominate synth in those instances by extrapolating outside the convex hull

# Replication

- We will now show code that does it in R and Stata
  - R: "augsynth" by the authors
  - Stata: "allsynth" that does several (including augsynth)
- Two examples to hopefully illustrate the bias and improvements and the lack of bias and no change in another
  - Smoking: augmented synth improves it
  - Prisons: augmented synth does nothing

# Possibilities for detecting corruption

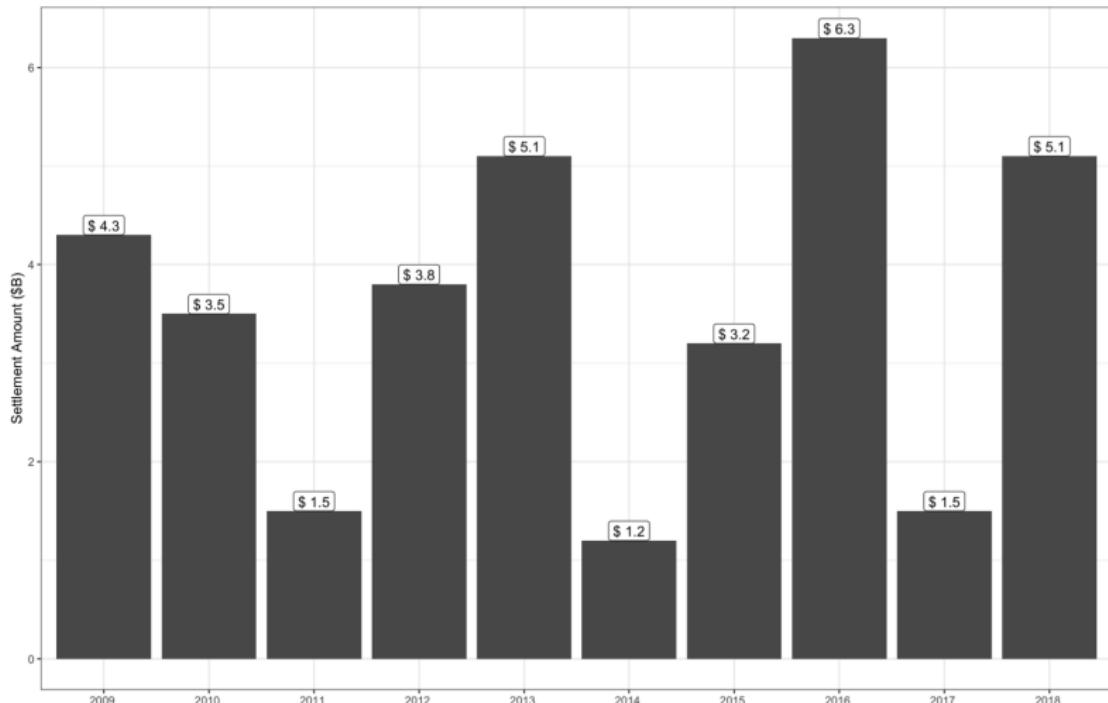
- Event studies in finance have been used to detect abnormal patterns around “events” involving single firms
- Baker and Gelbach (2020) proposes a type of synthetic control estimator that uses machine learning to estimate a counterfactual, as opposed to imposing strong parametric assumptions
- Examples of its use have been applied to disruptions with the Elon Musk Twitter deal which while not corruption does involve estimating potential damages from stock price movements

## Largest Securities Class Action Settlements

1. Enron: \$7.2b
2. WorldCom Inc: \$6.1b
3. Tyco International Ltd.: \$3.2b
4. Cendant Corporation: \$3.2b

Over time

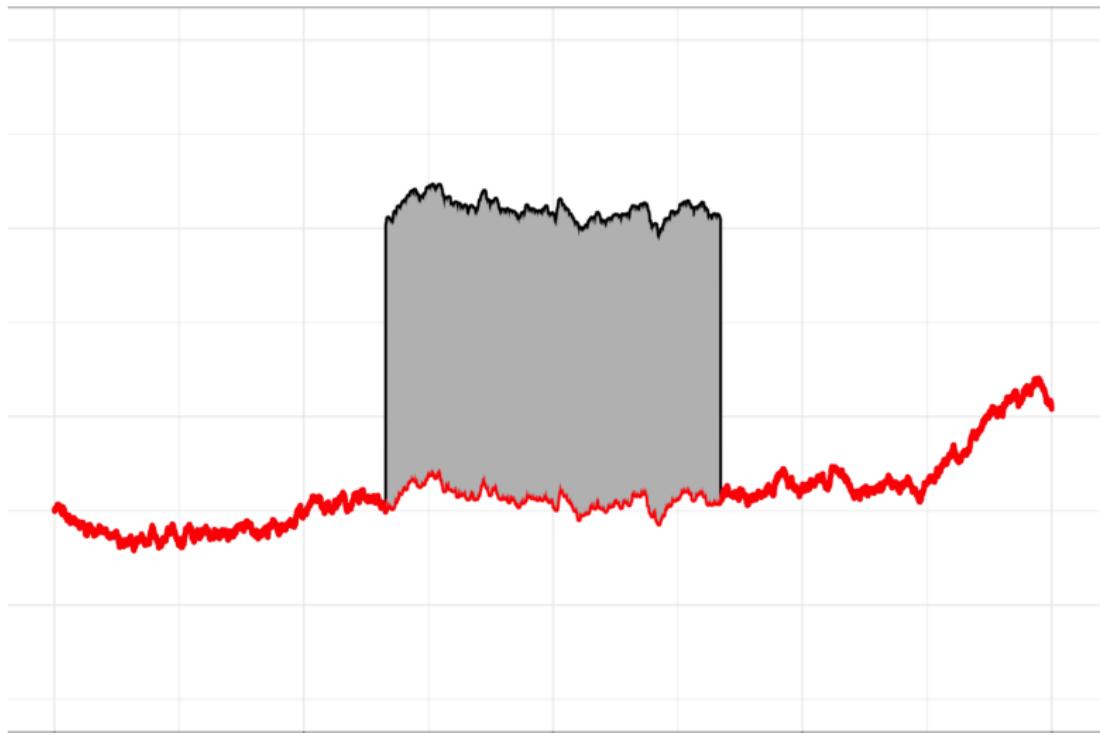
# Aggregate Settlement Value By Year



# Event studies and securities litigation

- Historically, the “event study” estimated “abnormal” returns under strong parametric assumptions (e.g., normality), but non-normal returns are normal  
*“The abnormal returns are the parameters that determine the damage estimates in securities suits, it is worthwhile to explore whether methods exist that can provide more accurate estimates of the abnormal return itself.”*
- They argue that the event study is an out-of-sample prediction problem, which ML is used for, but it is also an extension of the synth modeling framework

## Basic idea



## Event studies as a prediction problem

- Let the daily return for firm  $i$  on date  $t$  be  $r_{i,t}$  and variables used for prediction is  $X_{i,t}$  (e.g., market return, Fama-French and Carhart factors, a 1 for intercept, etc.)
- Suppose an event reveals fraud. Its effect on daily return is  $r_{i,t}^1 - r_{i,t}^0$  and we want to estimate  $r_{i,t}^0$  with  $\hat{r}_{i,t}^0$
- Construct a predicted residual as  $\hat{\varepsilon}_{i,t} = r_{i,t} - \hat{r}_{i,t}^0$
- Typically people would estimate this with OLS

$$r_{i,t} = \alpha + \beta_1 X_{i,t} + \varepsilon_{i,t}$$

# OLS, ML, MSE, Bias, Variance

- MSE of predicted abnormal return for  $\hat{\varepsilon}_{i,t} = r_{i,t} - \hat{\beta}X_{i,t}$  is the sum of a squared bias term and a variance term
- It's possible that the variance of one specification is lower enough than another to make up for a difference in bias
- OLS also suffers because it overfits data when used for prediction – it is best unbiased linear predictor but at the price of greater out-of-sample variance linear prediction
- Since MSE is the basis for measuring prediction accuracy, ML estimators may outperform conventional OLS as we can explore increasing bias and reducing variance
- ML methods accept bias in exchange for reduced variance out-of-sample accomplished through “training”

## Paper's punchline

*"Using real stock return data, we demonstrate that a number of out-of-the-box statistical approaches that are relatively easy to interpret perform better than the standard, OLS-based event study specifications used in court proceedings.*

*We find that specifications using penalized regression generally perform well. Specifications that adjust for daily market performance using data-driven peer indexes also generally perform well.*

*Finally, we obtain generally good performance from specifications that use a cross-validation technique that is robust to otherwise unmodeled time-series properties of the DGP. The best specifications provide noticeable improvements over event study approaches conventionally used in securities litigation.*

## Peer index

- They note that the best-performing specification makes use of both penalized regression and data-driven peer firm choice.
- They call this the “reasonable peer index”, and they show that ML methods can usefully serve as a basis for choosing *which* peer firms to include in an event study (again, making this a synth-like method) which can mitigate the subjective researcher bias that synth is meant to overcome
- Rather than subjectively picking which firms represent the counterfactual (over which there can be debate clearly, some disingenuous given the amount of money at stake), they propose letting the data say who the best peer is
- But using *any* peer index appears to mitigate this too

# Ranking all the ML methods

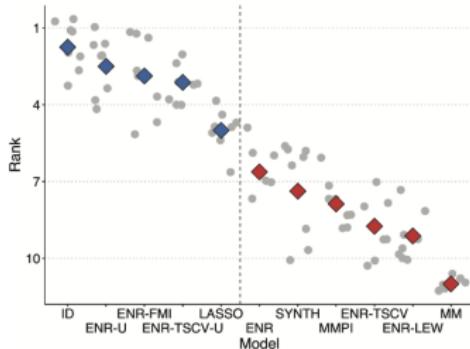


Figure 1: Distribution of Specification Ranks Across Models and Tests.

**Note:** Figure 1 plots specification ranks. Each specification has 8 MSE performance values: two time periods (1999–2009 vs. 2009–2019), with and without the FFC factors, and two MSE normalization approaches ( $\hat{R}_{oos}$  and  $\hat{R}_{het}$ , described below). Each gray dot represents a rank from 1 to 11, and each rank is represented once for each of the eight time-period/FFC-factor/MSE-metric combinations. The diamonds plot the specifications' average ranks. Blue diamonds signify models that allow firms to enter the regression function individually and use cross-validation and penalized regression; red diamonds represent specifications that do not.

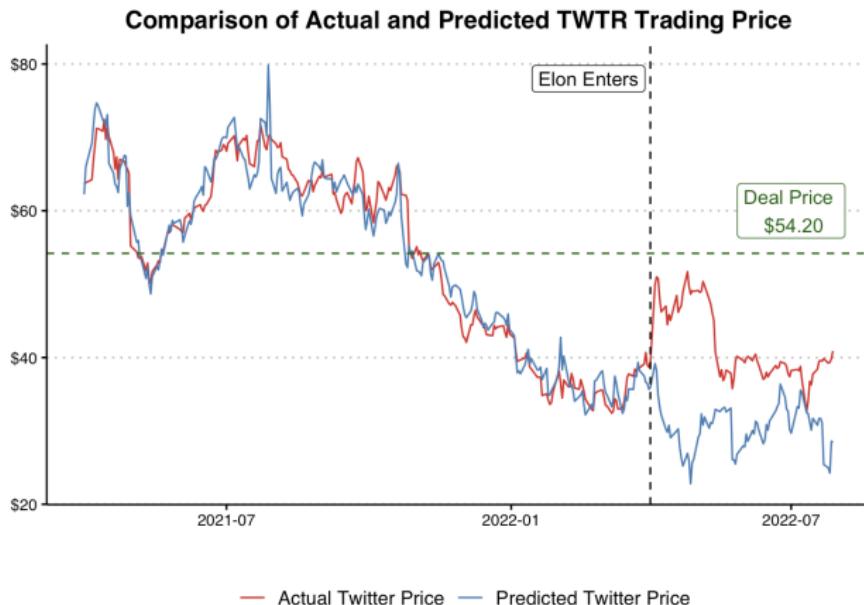
## Elon Musk example

- In an unpublished analysis, Baker examined Elon Musk's attempt to buy Twitter on Twitter's stock price
- Unlike his published paper, he's only going to use one form of "penalized" machine learning called ridge regression (which constrains what the coefficients can be in his model)
- He will use peer index and the S&P500 for prediction purposes

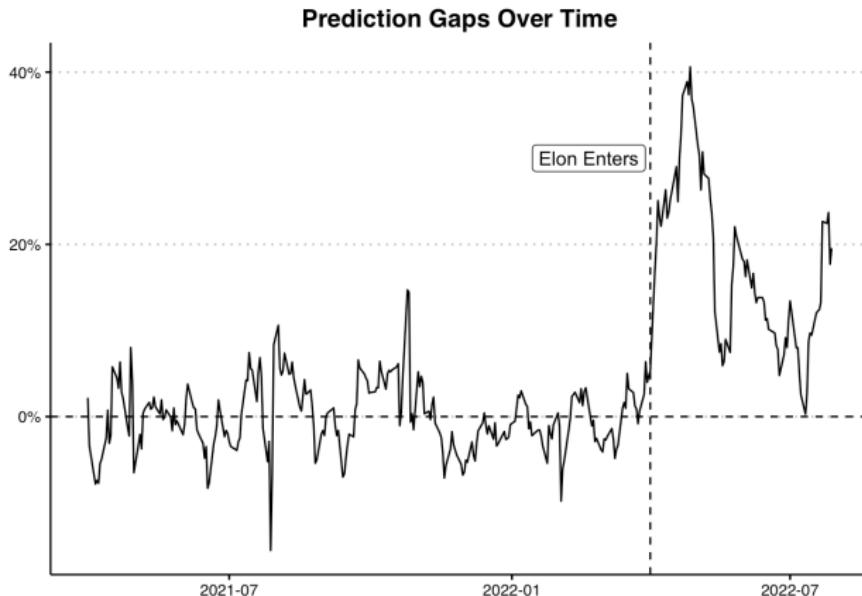
## Purpose of the exercise

*"The goal here is to get a rough estimate of what TWTR would be trading at had Elon never put the stock in play. Note, this does not mean that the prediction is equivalent to what TWTR would trade at were the deal to not go through (without any damage payments), as Elon has likely destroyed value in the process. This prediction could in fact be used as a baseline price in any tort-type damages claim that the company would want to bring against Elon after the process is over."*

# Basic idea



# Basic idea



## Abadie on the value and use of synthetic control

Synthetic controls provide many practical advantages for the estimation of the effects of policy interventions and other events of interest. However, like for any other statistical procedure (and especially for those aimed at estimating 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. In this article, I

# Roadmap

Introduction to course

- What is Mixtape Sessions?

- Potential outcomes review

Original synthetic control method

Imperfect fit

- Bounding the bias

- Demeaned Synthetic Control

- Augmented Synthetic Control

- Applying Machine Learning to Event studies from Finance

- Concluding remarks

Multiple Outcomes

Multiple Treated Units and Staggered

- Matrix completion with nuclear norm

- Synthetic difference-in-differences

- Synthetic Control with Staggered Adoption

## Motivation: A Tale of Two Californias

- Recall from earlier the example wherein California in 1988 implemented strong anti-smoking laws—taxes, advertising bans, public campaigns.
- In Abadie et al. (2010), a "Synthetic California" was built to estimate the law's effect on smoking.
- Suppose we then asked: What about alcohol consumption?
  - We'd construct a new synthetic California matched on pre-treatment alcohol trends.
  - But we'd likely get different weights—different donor states.

# Interpretability Challenge

- Now we have two synthetic Californias:
  - One built to predict smoking trends
  - Another to predict alcohol trends
- This breaks the narrative: which one is "the" counterfactual California?
- SCM aims to construct a unit-level counterfactual—not a new one for every outcome.

## Why This Problem Matters More in SCM

- Matching on covariates (e.g., Abadie & Imbens) allows unit-specific matches—it's fine if donor sets vary.
- SCM differs: it promises to build one synthetic unit.
- If outcomes are driven by the same unobserved factors, we gain both interpretability and statistical efficiency by using common weights.

## Motivation: SCM with Multiple Outcomes

- Many SCM applications involve multiple outcomes, like arrests, violence and STIs (Cunningham and Shah 2018), employment and wages (Jardim, et al. 2022), etc.
- Standard practice is to estimate separate SCM weights for each outcome.
- Problem is that separate weights can (will?) yield very different donor pools & make interpretation difficult.

## Value-Add of This Paper

- Sun, Ben-Michael, and Feller (2025) proposes methods to estimate a **common set of SCM weights** across outcomes.
- They'll present two approaches:
  - Concatenated outcomes
  - Averaged outcomes
- Their goal is to lower bias & get more interpretable synthetic units.

# A Tension: Fit vs. Interpretability

## Audience Thought

"If I separately optimized each outcome, won't any common set of weights necessarily fit **worse**?"

- Forcing a common set of weights improves interpretability—but may worsen pre-treatment fit.
- Separate SCM fits each outcome individually better—but sacrifices coherence.
- The real challenge: can we pool structure *without* making both sides worse?

## Separate SCM: Optimal for Each Outcome Individually

- In separate SCM for outcome  $k$ , we solve:

$$\min_{\gamma} \sum_{t=1}^{T_0} \left( Y_{1tk} - \sum_{i=2}^N \gamma_i Y_{itk} \right)^2$$

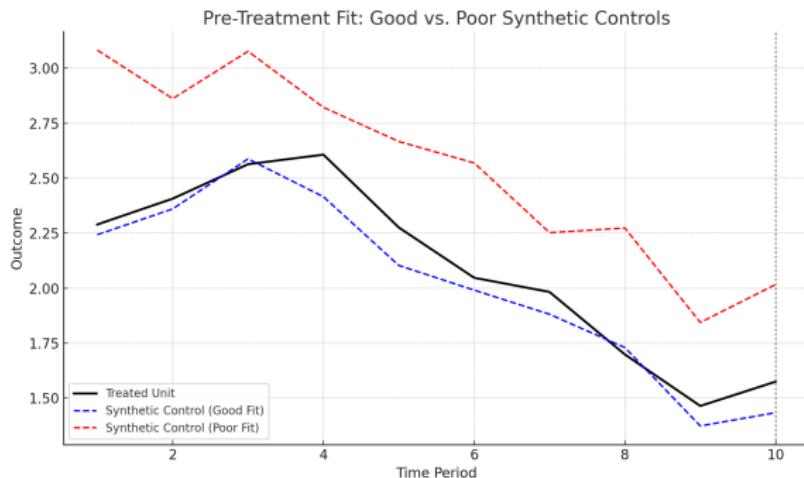
subject to:

- $\sum_{i=2}^N \gamma_i = 1$  (weights sum to one)
- $\gamma_i \geq 0$  for all  $i$  (non-negative weights)

- Thus, the resulting weights are **optimal for outcome  $k$**  given these constraints.
- Without the non-negativity constraint, weights might sometimes be negative (as in some extensions like augmented SCM).

# Reminder: Pre-Treatment Fit

- In SCM, even the best possible weights usually leave some **residual discrepancy**.
- $q(\gamma)$  measures the pre-treatment gap between the treated unit and the synthetic control.
- Smaller  $q(\gamma)$   $\Rightarrow$  better fit, but perfect match is rare given real-world donor pools.



# A New Optimization Problem: Global Coherence

- Sun, Ben-Michael, and Feller (2025) propose not just a new estimator, but a **new objective**.
- Minimize pre-treatment discrepancy **across outcomes simultaneously** by optimizing a *pooled* loss:

$$\min_{\gamma} \{\nu \cdot q_{\text{avg}}(\gamma) + (1 - \nu) \cdot q_{\text{cat}}(\gamma)\}$$

- Where:
  - $q_{\text{avg}}(\gamma)$  minimizes discrepancy for the average of outcomes.
  - $q_{\text{cat}}(\gamma)$  minimizes discrepancy across all outcomes stacked together.
- The resulting common weights are **optimal for this new pooled objective**—not for any single outcome's individual fit.

## Setup and Target Parameter

- Recall that Abadie used  $J$  for the number of donor units by here Sun, Ben-Michael, and Feller (2025) use  $N$  for total units and  $N_0$  for donor units.
- One treated unit,  $N - 1$  control units,  $K$  outcomes.
- Interested in post-treatment effects  $\tau_k = Y_{1T,k}(1) - Y_{1T,k}(0)$ .
- SCM estimates  $Y_{1T,k}(0)$  using weighted average of donor pool units.

## Demeaned estimator

*"We follow the potential outcomes framework ... and throughout focus on de-meansed or intercept-shifted weighting estimators, which were introduced in the single outcome setting (Doudchenko and Imbens, 2017; Ferman and Pinto, 2021) and were adapted to multiple outcomes by Tian et al. (2023), who argue that outcome-specific demeaning is useful for comparing across outcomes." – Sun, Ben-Michael and Feller (2025)*

# Estimator Form

- Demeaned estimator:

$$\hat{Y}_{1T,k}(0) = \bar{Y}_{1\cdot k} + \sum_{i=2}^N \gamma_i (Y_{iT,k} - \bar{Y}_{i\cdot k})$$

- Where:
  - $\bar{Y}_{1\cdot k}$  is the treated unit's pre-treatment average for outcome  $k$ .
  - $\bar{Y}_{i\cdot k}$  is control unit  $i$ 's pre-treatment average for outcome  $k$ .
  - $(Y_{iT,k} - \bar{Y}_{i\cdot k})$  is the deviation of control unit  $i$ 's post-treatment outcome from its pre-treatment mean.
  - $\gamma_i$  are the donor pool weights chosen to best match pre-treatment trajectories.
- We choose weights  $\gamma$  to minimize pre-treatment imbalance.
- **Key difference:** How we define imbalance when  $K > 1$ .

# Problems with Separate Weights

- Each outcome gets its own set of SCM weights.
- Can lead to overfitting, especially in short panels.
- Different donors per outcome  $\Rightarrow$  hard to interpret one synthetic unit.

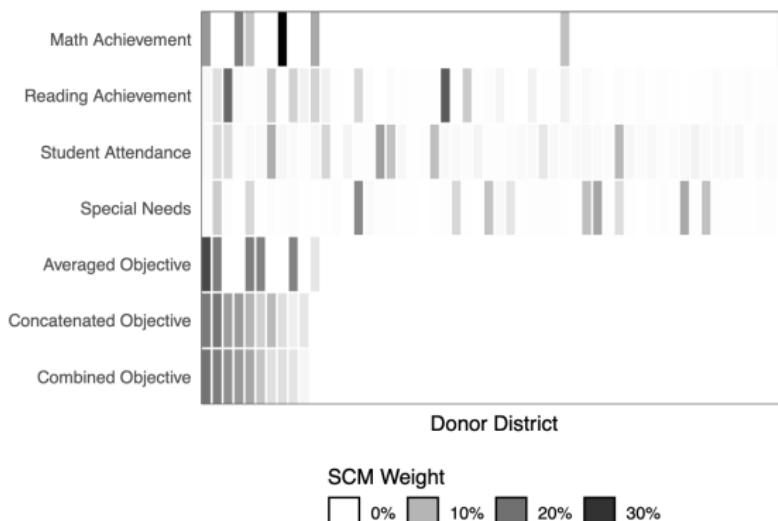


Figure 1: Separate SCM weight placed on each donor unit using (i) Student Attendance, (ii) Special Needs, (iii) Reading Achievement, and (iv) Math Achievement separately as outcomes, along with the weights solving the concatenated and averaged objectives. The top 5 districts accounting for over 75% of the weights solving the combined objective are (i)

## Proposed Solution: Shared Weights

- Estimate a **common set of weights** across outcomes.
- Two strategies:
  - **Concatenated:** stack all outcome-time pairs.
  - **Averaged:** average outcomes before computing imbalance.

# Formalizing the Objective Functions

- Concatenated weights minimize:

$$q_{\text{cat}}(\gamma)^2 = \frac{1}{T_0 K} \sum_{k=1}^K \sum_{t=1}^{T_0} \left( Y_{1tk} - \sum_{i=2}^N \gamma_i Y_{itk} \right)^2$$

- Averaged weights minimize:

$$q_{\text{avg}}(\gamma)^2 = \frac{1}{T_0} \sum_{t=1}^{T_0} \left( \frac{1}{K} \sum_{k=1}^K Y_{1tk} - \sum_{i=2}^N \gamma_i \frac{1}{K} \sum_{k=1}^K Y_{itk} \right)^2$$

# Why Averaging Helps

- Averaging outcomes smooths out idiosyncratic noise.
- Leads to tighter fit and lower bias bounds as  $K$  grows.
- Bias bound:  $O(1/\sqrt{K})$  under low-rank factor model.

## Key Assumption: Low-Rank Structure

- Outcomes are driven by common latent factors:

$$L_{itk} = \phi_i^\top \mu_{tk}$$

- Oracle weights exist iff the data matrix is low rank.
- Factor model aligns donor units via shared structure.

# Bias Decomposition

- Total estimation error = bias + noise
- Bias: failure to balance latent structure
- Noise: post-treatment idiosyncratic error

# Theorem 1: Bias Bounds

- Separate SCM:  $O(1)$  bias
- Concatenated SCM:  $O(1)$ , better overfitting control
- Averaged SCM:  $O(1/\sqrt{K})$  bias

	Bias due to imperfect fit	Bias due to overfitting
$\hat{\gamma}^{sep}$	$O(1)$	$O\left(\frac{1}{\sqrt{T_0}}\right)$
$\hat{\gamma}^{cat}$	$O(1)$	$O\left(\frac{1}{\sqrt{T_0 K}}\right)$
$\hat{\gamma}^{avg}$	$O\left(\frac{1}{\sqrt{K}}\right)$	$O\left(\frac{1}{\sqrt{T_0 K}}\right)$

Table 1: Leading terms in high probability bounds on the bias due to imperfect fit and overfitting in Theorem 1, with  $N$  fixed.

# Practical Diagnostics

- Singular value decomposition (SVD): do a few components explain most variation?
- Hold-out fit: leave one outcome out and check fit.
- Combine averaging and concatenation (weighted frontier).

# Application: Flint Water Crisis

- Flint is treated unit; 54 MI school districts as controls.
- 4 outcomes: math, reading, attendance, special needs.
- Estimate SCM using separate, cat., avg., and combined weights.

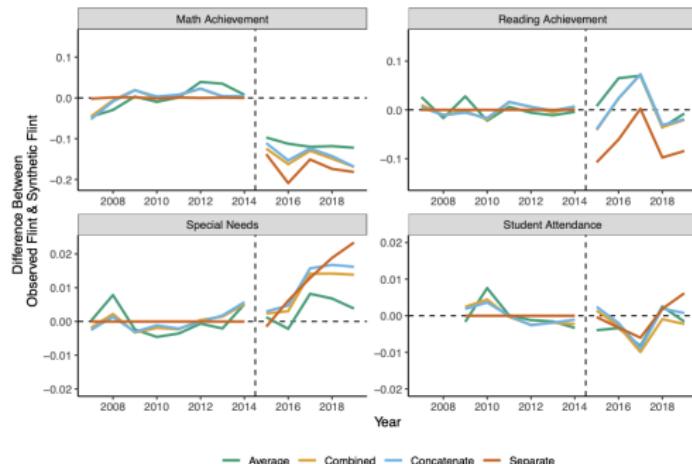


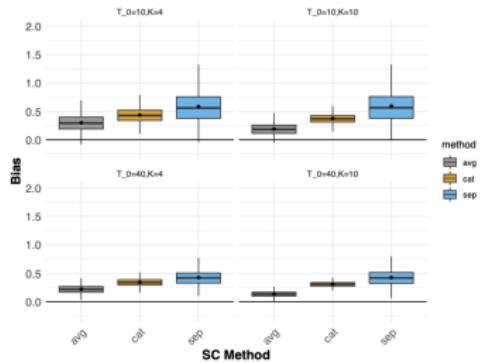
Figure 2: Point estimates for the effect of the Flint water crisis using separate weights, concatenated weights, and average weights, as well as the combined weights (setting  $\nu = 0.47$ ). The separate SCM weights yield essentially perfect pre-treatment fit for all four outcomes.

## Flint Results: Interpretation

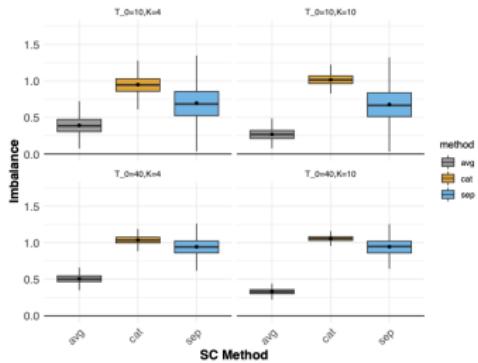
- Separate weights fit very well pre-treatment but likely overfit.
- Average and combined weights offer good fit with less overfitting.
- Estimated impacts: decline in math scores, increase in special needs.

# Simulation: Clean vs. Noisy Outcomes

- Vary strength of shared vs. idiosyncratic factors.
- When outcomes share structure: avg. & cat. reduce bias.
- When outcomes diverge: separate may be safer.



(a)

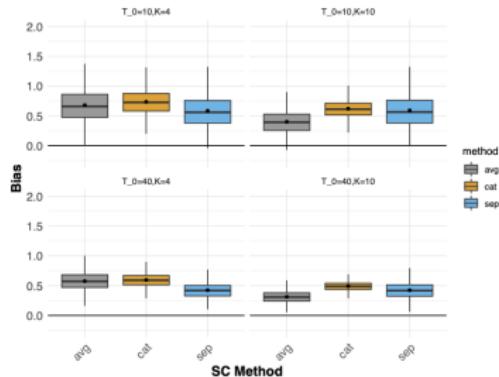


(b)

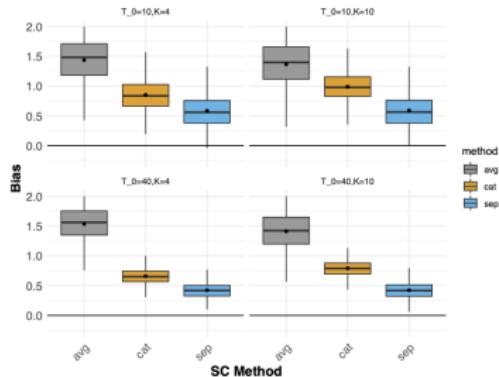
Figure D.1: Box plots of bias and imbalance using separate SCM, concatenated SCM, and average SCM over 1000 simulations.

# Simulation: Failure Case

- When  $\rho = 0$ , no shared structure  $\Rightarrow$  avg. SCM fails.
- Bias explodes as averaging collapses signal.
- Use SVD or condition number to detect weak structure.



(a) Common + idiosyncratic factors;  $\rho = 0.5$



(b) Only idiosyncratic factors;  $\rho = 0$

Figure D.2: Box plots of bias using separate SCM, concatenated SCM, and average SCM over 1000 simulations.

## Recommendations for Practice

- Standardize outcomes.
- Check SVD or hold-out diagnostics.
- Use average or combined weights when K is large and shared structure is plausible.

# Combined Objective Frontier

- Blend average and concatenated objectives.
- Tune  $\nu$  to balance fit and robustness.
- Visual: `augsynth` package frontier plot

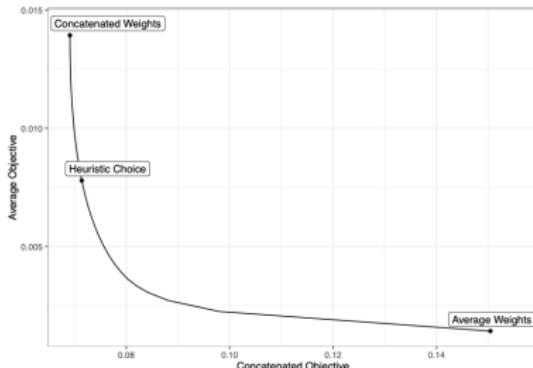
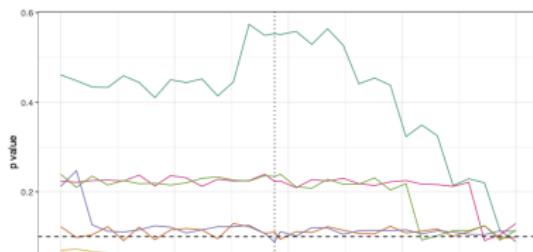


Figure E.3: Frontier plot.



## Why This Paper Matters

- Pushes SCM beyond single-outcome constraint.
- Improves interpretability and bias control.
- Leads naturally into matrix completion & machine learning SCM.

# Questions?

- Let's discuss diagnostics, estimation details, or extensions.
- Also happy to talk about **augsynth** implementation.

# Roadmap

Introduction to course

- What is Mixtape Sessions?

- Potential outcomes review

Original synthetic control method

Imperfect fit

- Bounding the bias

- Demeaned Synthetic Control

- Augmented Synthetic Control

- Applying Machine Learning to Event studies from Finance

- Concluding remarks

Multiple Outcomes

Multiple Treated Units and Staggered

- Matrix completion with nuclear norm

- Synthetic difference-in-differences

- Synthetic Control with Staggered Adoption

## Big idea

*"The main part of the article is about the statistical problem of imputing the missing values of  $Y$ . Once these are imputed, we can estimate the causal effect of interest,  $\delta$ ."*

*"To estimate average causal effect of the treatment on the treated units, we impute the missing potential control outcomes" – Athey, et al. (2021)*

All of causal inference is imputation – but some methods are more explicit and do so in a way that layers on stronger assumptions than others – and matrix completion with nuclear norm regularization is one such example.

## Overview

- Matrix completion with nuclear norm regularization is a synthetic control estimator developed by Athey, et al. (2010, JASA) that can accommodate a variety of setups
- Specifically it unites single period treatments (unconfoundedness assumptions) with single unit treatments (synthetic control)
- Uses imputation based on a low rank matrix to impute missing counterfactuals of treated units to construct estimates of the ATT
- Nuclear norm regularization is used for the imputation (as opposed to lasso or elastic net)
- Can accommodate differential timing setups where parallel trends may not hold (see Cunningham, Tripp and DeAngelo 2023, forthcoming JHR)

# What is matrix completion

- Completing a matrix means imputing the correct values for variables with missing values
- In causal inference, we are missing potential outcomes (e.g.,  $Y^0$ ), and the missingness is caused by treatment assignment
- Our target parameter is the ATT, so we are missing  $Y^0$  for all treated units

# History of matrix completion

- Open competition by Netflix in 2006 – winner would get \$1m if they could improve predictive model by ten points on RMSE
- Invited a ton of competition – from MIT teams to regular everyday joes working out of their home office
- Everyone was given a database which was then tested by Netflix on a holdout dataset
- Quick progress was made followed by very slow advances
- Winner was announced in 2009

# Netflix prize

- Gigantic sparsely populated matrix of movies ranked by users
  - I like Napoleon Dynamite and The Matrix and you like The Matrix
  - Should Netflix recommend you watch Napoleon Dynamite?
- It was implicitly causal and predictive though – “if you are shown Napoleon Dynamite, will you like it?” – is a causal question

Here's a matrix of potential outcomes,  $Y^0$ , representing units at time  $t$  that had not been treated.

$$Y_{it}^0 = \begin{pmatrix} Y_{11}^0 & Y_{12}^0 & Y_{13}^0 & \dots & Y_{1t}^0 \\ Y_{21}^0 & Y_{22}^0 & Y_{23}^0 & \dots & Y_{2t}^0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_{i1}^0 & Y_{i2}^0 & Y_{i3}^0 & \dots & Y_{it}^0 \end{pmatrix}$$

Now imagine a treatment assignment that assigns treatment in the last period  $t$  using the switching equation  $Y = DY^1 + (1 - D)Y^0$

Since this is a matrix now of potential outcomes,  $Y^0$ , we are missing anyone's potential outcome who was treated.

$$Y_{it}^0 = \begin{pmatrix} Y_{11}^0 & Y_{12}^0 & Y_{13}^0 & \dots & ? \\ Y_{21}^0 & Y_{22}^0 & Y_{23}^0 & \dots & ? \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_{i1}^0 & Y_{i2}^0 & Y_{i3}^0 & \dots & ? \end{pmatrix}$$

Matrix completion with nuclear norm will impute the last column using regularized regression:

$$Y_{it}^0 = \begin{pmatrix} Y_{11}^0 & Y_{12}^0 & Y_{13}^0 & \dots & \widehat{Y_{1t}^0} \\ Y_{21}^0 & Y_{22}^0 & Y_{23}^0 & \dots & \widehat{Y_{2t}^0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_{i1}^0 & Y_{i2}^0 & Y_{i3}^0 & \dots & \widehat{Y_{it}^0} \end{pmatrix}$$

And once you have those, you can calculate individual level treatment effects that can be used to aggregate to the ATT

## Commentary about synth imputation vs did imputation

- Important point – the assumptions in diff-in-diff use parallel trends to do imputation (e.g., Borusyak, et al. (2023), Heckman, Ichimura and Todd (1997)) – but synthetic control does not
- So you can't compare MCNN with Borusyak, et al. (2023) even though both of them are going to be plausible candidates in the same situations and may "feel" the same
- Navigating between synthetic control estimators and diff-in-diff estimators when the gaps between their use cases decline to zero will necessarily require understanding the identifying assumptions at a deep level, and it's unclear if there are available tests beyond the event studies to justify their use

## Two literatures

- There's two literatures they want you to have in your mind:
  1. Unconfoundedness –  $(Y^0, Y^1) \perp\!\!\!\perp D|X$  – sometimes explicitly imputes (nearest neighbor), sometimes more implicit (inverse probability weighting)
  2. Synthetic control – literally calculating a counterfactual as a weighted average over all donor pool units
- Their MCNN method will show that both are “nested” within the general framework they’ve developed making them actually special cases

# Differences between synth and unconfoundedness

- Conceptually different in the way they exploit patterns for causal inference
  - Unconfoundedness assumes that **patterns over time** are stable across units
  - Synth assumes **patterns across units** are stable over time
- Regularization nests them both and a nuclear norm ensures a low rank matrix needed for sensible imputations

## Factor models again

- Factor models and interactive effects model the observed outcome as the sum of a linear function of covariates and a unobserved component that is a low rank matrix plus noise
- Estimates are typically based on minimizing the sum of squared errors given the rank of the matrix of unobserved components with the rank itself estimated
- Nuclear norm regularization will be used for imputing the potential outcomes,  $Y^0$ , for all treated units
- Estimate plots and overall ATT using the estimated treatment effects

## Three contributions

1. Formal results for non-random missingness when block structure allows for correlation over time.
2. Shows unconfoundedness and synth are in fact matrix completion methods
  - Same objective function, but
  - Different sets of restrictions on the factors in the matrix factorization
  - MCNN doesn't impose any restrictions – just regularization to characterize the estimator – whereas synthetic control imposes convex weights
3. Applies the method to two datasets, but I'll show you a recent study of mine in what we chose to do

# Block structure

- Lots of new and old terms – unconfoundedness, vertical and horizontal regression, fat and thin matrices.
- We define the matrix first in terms of its block structure which is describing where and when the missingness is occurring in the matrix
- Then we will discuss these concepts related to the direction of the regressions themselves

# Unconfoundedness

- Much of the program evaluation literature estimates an average treatment effect (e.g., ATT) under unconfoundedness (see Lalonde 1986; Dehejia 2002; Smith and Todd 2005)
- But the focus is on a simple setup where the missingness is the last period (i.e., the post treatment period)
- In LaLonde (1986), NSW treats the workers, and then you don't observe  $Y^0$  for the treated group in the *last period*, so you use the comparison group for imputation of missing potential outcomes under presumed unconfoundedness (i.e., randomized within dimensions of  $X$ )
- Athey et al call this the “single-treated-period block structure” because only one period is missing

## Single-treated-period block structure

$$Y_{it}^0 = \begin{pmatrix} Y_{11}^0 & Y_{12}^0 & Y_{13}^0 & \dots & Y_{1T}^0 \\ Y_{21}^0 & Y_{22}^0 & Y_{23}^0 & \dots & ? \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_{i1}^0 & Y_{i2}^0 & Y_{i3}^0 & \dots & ? \end{pmatrix}$$

## Single-treated-unit block structure

$$Y_{it}^0 = \begin{pmatrix} Y_{11}^0 & Y_{12}^0 & Y_{13}^0 & \dots & Y_{1t}^0 \\ Y_{21}^0 & Y_{22}^0 & Y_{23}^0 & \dots & Y_{2t}^0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_{i1}^0 & Y_{i2}^0 & ? & \dots & ? \end{pmatrix}$$

Notice, this is the synthetic control design because a single unit (unit  $i$ ) is missing  $Y^0$  for the 3rd and  $t$ th periods.

Differential timing has a block structure too

$$Y_{it}^0 = \begin{pmatrix} Y_{11}^0 & ? & ? & \dots & ? \\ Y_{21}^0 & Y_{22}^0 & Y_{23}^0 & \dots & ? \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_{i1}^0 & Y_{i2}^0 & ? & \dots & ? \end{pmatrix}$$

So all of these so-called designs can be expressed in terms of missingness in the block structure, and our job therefore is to find an estimator that is general enough to manage all of them. Matrix completion with nuclear norm regularization is one such example and so has nice generalizations

## Thin and Fat matrices

- Next we must consider the relative number of panel units  $N$  and time periods  $T$  because this also shapes which regression style will be used for imputation
- Thin matrices are where  $N \gg T$  (relatively large numbers of units), and fat matrices are ones where  $T \gg N$  (relatively large numbers of time periods)
- Approximately square ones are where  $T$  is approximately equal to  $N$

## Vertical and horizontal regression

- Unconfoundedness has that single-treated period block structure with a thin matrix ( $N \gg T$ ).
- Uses a large number of  $N$  units to impute missing potential outcomes in the last period using controls with similar lagged outcomes (horizontal regression)
- Horizontal regression holds under unconfoundedness

## Vertical regression

- Doudchenko and Imbens (2016) and Pinto and Furman (2021) show that synthetic control can be interpreted as regressing the outcomes for the treated prior to treatment on the outcomes for controls in the same period
- It is as though you are regressing the treatment group outcomes onto the donor pool outcomes (up and down) but with restrictions on what the coefficients can be – they cannot be negative and they must sum to one
- Once synthetic control is framed as a vertical regression, it opens up the possibilities of different model specifications (like including an intercept which will shift the convex hull around which is what Doudchenko and Imbens (2016, unpublished) proposed)

# Fixed effects and factor models

- Both horizontal and vertical regressions exploit unique patterns in the data – one goes left to right, one goes up and down
- An alternative to each of them though is to consider an approach that allows for the exploitation of both stable patterns over time and stable patterns across units
- Matrix completion with nuclear norm does both (and we will see this again in their synthetic difference-in-differences)

## Matrix completion with nuclear norm

- Model the  $N \times T$  matrix of complete outcomes data matrix  $Y$  as:

$$Y = L^* + e$$

where  $E[e|L^*] = 0$

- The error term can be thought of as measurement error if you need a frame to think about it
- So you have this complete matrix,  $L^*$ , and zero mean conditional independence holds

# Assumption regarding error and matrix

## Matrix assumptions

$e$  is independent of  $L*$  and the elements of  $e$  are  $\sigma$ -sub-Gaussian and independent of each other

## Comment about regularization

- You could minimize the sum of squared differences but if the objective function doesn't depend on  $L*$ , the estimator would just spit back  $Y$  and  $\delta = 0$ .
- Authors add a penalty term  $||\lambda||$  to the objective function, but even then, not all of them do well.
- It matters whether you regularize the fixed effects or not

# Estimator

$$L* = \widehat{L} + \widehat{\Gamma} \mathbf{1}_T^T + I_N \widehat{\Delta}^T$$

where the objective function is:

$$= \arg \min_{L, \Gamma, \Delta} \left\{ \frac{1}{O} \| P_0(Y - L - \Gamma \mathbf{1}_T^T - \mathbf{1}_N \Delta^T) \|_F^2 + \Lambda \| L \| \right\}$$

## Fixed effects and regularization

- The penalty will likely be the nuclear norm but notice that the fixed effects are outside the penalty term. You could subsume them into  $L$ , they say, but they recommend you not doing this.
- Fraction of observations is relatively high and so the fixed effects can actually be estimated separately (apparently that is one difference between MCNN and the rest of the MC literature)
- The penalty will be chosen using cross-validation
- One advantage of NN is its fast and convex optimization programs will do it, whereas some others won't because of the large  $N$  or  $T$  issues

## Parting comments

- Though this model can be used for differential timing but at the moment, we haven't seen much in terms of contrasting it with the robust diff-in-diff estimators like Borusyak, et al. (2021), Callaway and Sant'Anna (2020) or any of the others
- This is going to come down to parallel trends versus the factor models, and communicating the improvements gained with synthetic control as you move away from the simplest comparative case study is likely to be challenging
- You choose the estimator based on the problem you're studying and the assumptions – you must justify it, no one else can, but you do so by appealing to assumptions

# Code

R: <https://github.com/xuyiqing/gsynth>

Stata: fect

[https://yiqingxu.org/packages/fect/stata/fect\\_md.html](https://yiqingxu.org/packages/fect/stata/fect_md.html)

# The Basic Idea of SDID

- Combines ideas from synthetic control and difference-in-differences.
  - Matches control units to treated units with pre-treatment weights.
  - Addresses biases from imperfect matching.
- Introduces time weights to prioritize key pre-treatment periods.
- Balances precision from synthetic control with broader trends from DID.

# How SDID Works

- Uses unit and time fixed effects for bias reduction.
  - Accounts for unit-specific traits (e.g., state-level differences).
  - Captures common period effects (e.g., economic booms).
- Derives time weights directly from the data.
- Flexible tool for causal inference in complex panel data.

# Intuition Behind SDID

- Approximates parallel trends by reweighting data.
  - Unit weights align control and treated units pre-treatment.
  - Time weights balance trends across periods.
- Reduces reliance on strong assumptions about pre-existing trends.
- Enhances robustness when pre-existing trends don't align perfectly.

# Assumptions in Synthetic Difference-in-Differences (SDID)

In both synthetic control and difference-in-differences, assumptions about pre-treatment trends are central. SDID combines these methods to leverage their strengths:

- Uses pre-treatment trends to reweight treated and control groups.
- Constructs a comparison group that mimics the treated group's trends.
- Relies on four key assumptions for identification.

## Assumption 1: Error Properties

- Errors are independent, identically distributed Gaussian vectors.
- Covariance matrix has bounded eigenvalues.
- Assumption is stricter than traditional DiD or synthetic control, which only assume error independence.

## Assumption 2: Sample Sizes

- Large sample sizes are critical for SDID:
  - Number of control units ( $N_{co}$ ) and pre-treatment periods ( $T_{pre}$ ) must be large.
  - Product of treated units ( $N_{tr}$ ) and post-treatment periods ( $T_{post}$ ) should also be large.
- Balance between  $T_{pre}$  and  $N_{co}$  ensures effective reweighting.

## Assumptions 3 and 4

### **Assumption 3: Systematic Component Properties**

- The systematic component ( $L$ ) has a limited number of large singular values.
- This low-rank structure captures broad patterns while reducing noise.
- Goes beyond traditional DiD and synthetic control.

### **Assumption 4: Weighting Properties and $L$**

- "Oracle" weights derived from  $L$  reduce systematic bias.
- Ensures parallel trends after reweighting without requiring perfect balance.

## Steps in SDID Estimation

- Identify unit weights for control units to match treated pre-treatment trends.
- Select time weights to balance pre- and post-treatment trends.
- Combine weights in regression to estimate the ATT.

## SDID Estimation Process

- Calculates unit and time weights to optimize comparison groups.
- Uses regularization to prevent overfitting.
  - Aligns with one-period outcome changes for untreated units.
- Solves constrained least-squares for unit weights ( $\hat{w}$ ).

## Estimation of SDiD

Synthetic DiD combines elements of DiD and synthetic control:

- Compute the regularization parameter to align with typical one-period outcome changes:

$$\Delta_{it} = Y_{i(t+1)} - Y_{it} \quad (\text{unexposed units}).$$

# Unit and Time Weights

- **Unit Weights ( $\hat{w}$ ):**
  - Match treated and control units pre-treatment.
  - Allows for a constant gap, relaxing traditional SC constraints.
- **Time Weights ( $\hat{\lambda}$ ):**
  - Balance temporal dynamics.
  - Prioritize informative pre-treatment periods.

# Estimating Weights

- Unit Weights:
  - Estimated via constrained least squares on pre-treatment data.
  - Weights are non-negative, sum to one, and allow for a level shift with regularization.
- Time Weights:
  - Estimated via constrained least squares on control data.
  - Prioritizes periods just before treatment without spreading weights broadly.

## Estimation of SDID: Unit Weights

- Estimate unit weights  $\hat{w}$  to define a synthetic control unit.
- Equation:

$$\hat{w}_1 + \hat{w}^T Y_{j,\text{pre}} \approx Y_{1,\text{pre}}$$

- Unlike traditional synthetic control, allows for an intercept term, relaxing perfect matching requirements.

## Estimation of SDID: Time Weights

- Estimate time weights  $\hat{\lambda}$  to align synthetic pre-treatment periods:

$$\hat{\lambda}_1 + Y_{1,\text{pre}} \hat{\lambda} \approx Y_{1,\text{post}}$$

- Focuses on identifying pre-treatment periods most informative for post-treatment behavior.

## Combined Estimation

- SDID estimator combines unit and time weights.
- Weighted regression minimizes residuals across units and time.
- Ensures counterfactual outcomes align with treated outcomes pre-treatment.

# Outcome Model in SDID

- Observed outcomes ( $Y$ ) consist of:
  - Systematic component ( $L$ ).
  - Treatment effect ( $D \circ \delta$ ).
  - Idiosyncratic error ( $E$ ).
- $D \circ \delta$  isolates ATT by selecting treated units and periods.
- $L$  captures trends not explained by treatment or noise.

# Oracle Weights

- Theoretical "ideal" weights for unit ( $\tilde{\omega}$ ) and time ( $\tilde{\lambda}$ ).
- Minimize bias by balancing systematic component ( $L$ ).
- Data-driven weights ( $\hat{\omega}, \hat{\lambda}$ ) approximate oracle weights.
  - Ensure pre-treatment balance.
  - Converge to oracle weights as sample size increases.

# Outcome Model

- Outcome is a combination of systematic trends, treatment effects, and noise:

$$Y = L + D \circ \delta + E$$

- Components:
  - $L$ : systematic trends across units and time ( $L = \Gamma \Upsilon^\top$ )
  - $D \circ \delta$ : treatment effects applied selectively
  - $E$ : idiosyncratic noise
- $L$  represents shared patterns unaffected by treatment or randomness.

## Role of Systematic Component $L$

- Captures trends unrelated to treatment or random noise.
- Potential source of confounding if not properly controlled.
- Treated and control groups must balance  $L$  pre-treatment for valid comparison.

## Oracle Weights

- Theoretical weights ( $\tilde{\omega}, \tilde{\lambda}$ ) minimize bias in estimating treatment effects.
- Balance systematic trends ( $L$ ) across treated and control groups.
- Provide a benchmark for robust causal inference.

## Data-Driven Weights

- Empirical weights ( $\hat{\omega}, \hat{\lambda}$ ) approximate oracle weights.
- Constructed to balance treated and control trends pre-treatment.
- Approximation improves with larger sample sizes.

# Weighted Regression for SDID

- Final estimation uses weighted DID regression:

$$\operatorname{argmin}_{\tau, \mu, \alpha, \beta} \sum_{i=1}^N \sum_{t=1}^T (Y_{it} - \mu - \alpha_i - \beta_t - W_{it}\tau)^2 \hat{w}_i^{\text{SDID}} \hat{\lambda}_t^{\text{SDID}}$$

# Regression Comparison: SC, DiD, SDiD

- Synthetic Control (SC):

$$\tau^{\text{SC}} = \operatorname{argmin}_{\tau, \lambda} \sum_{i,t} (Y_{it} - \lambda_t - \tau D_{it})^2 w_i^{\text{SC}}$$

- Difference-in-Differences (DiD):

$$\operatorname{argmin}_{\tau, \lambda, \alpha} \sum_{i,t} (Y_{it} - \lambda_t - \alpha_i - \tau D_{it})^2$$

- Synthetic Difference-in-Differences (SDiD):

$$\operatorname{argmin}_{\tau, \lambda, \alpha} \sum_{i,t} (Y_{it} - \lambda_t - \alpha_i - \tau D_{it})^2 w_i \lambda_t$$

## Key Features of SDiD

- Combines synthetic control's matching with DID's parallel trends adjustment.
- Introduces unit and time weights for improved balance.
- Does not require perfect pre-treatment matching or strictly parallel trends.

# Decomposing the Bias of SDID

$$\hat{\tau}^{sdid} - \tau = \varepsilon(\tilde{w}, \tilde{\lambda}) + B(\tilde{w}, \tilde{\lambda}) + \hat{\tau}(\hat{w}, \hat{\lambda}) - \hat{\tau}(\tilde{w}, \tilde{\lambda})$$

- **Oracle noise:** Variance introduced by weights and sample limitations.
- **Oracle confounding bias:** Systematic differences not removed by weights.
- **Deviation from oracle:** Approximation errors in empirical weights.

## Oracle Noise

- Noise arises from random variation in data.
- Small when weights  $(\hat{w}, \hat{\lambda})$  are small and sample sizes are sufficient.
- Ensures noise does not dominate the estimator.

# Oracle Confounding Bias: Units and Time

- **Units (Rows):**

$$\widetilde{w_1} + \widetilde{w_j}^T L_{j,pre} \approx \widetilde{w_1}^T L_{1,pre}$$

$$\widetilde{w_1} + \widetilde{w_j}^T L_{j,post} \approx \widetilde{w_1}^T L_{1,post}$$

Ensures weights remove bias from systematic differences across units.

- **Time (Columns):**

$$\widetilde{\lambda_1} + \widetilde{\lambda_j}^T L_{j,pre} \approx \widetilde{\lambda_1}^T L_{1,pre}$$

$$\widetilde{\lambda_1} + \widetilde{\lambda_j}^T L_{j,post} \approx \widetilde{\lambda_1}^T L_{1,post}$$

Captures temporal dynamics to align treated and control groups.

## Doubly Robust Property

- Sufficient for one set of weights (unit or time) to generalize well.
- Combination of oracle unit and time weights can compensate for failures in one dimension.
- Provides resilience against systematic confounding.

## Deviation from Oracle

- SDID approximates oracle weights when:
  - Oracle weights generalize well on training sets.
  - Regularization is appropriately tuned.
- Ensures SDID estimator is close to theoretical benchmark.

## Practical Implications

- Combines theoretical rigor with empirical flexibility.
- Balances systematic trends in pre-treatment data.
- Achieves robust causal inference for ATT estimation.

## Practical Considerations: Pre-treatment Trends

- Visually inspect pre-treatment trends to check for alignment between treated and control groups.
- Use plots to ensure parallel trends are approximately valid before treatment.
- Address any unusual patterns or discrepancies early.

# Practical Considerations: Weights and Assumptions

- Balanced weights:
  - Ensure  $\hat{\omega}$  and  $\hat{\lambda}$  are not overly concentrated.
  - Adjust regularization parameters if needed.
- Assess parallel trends:
  - Contextual knowledge remains crucial.
  - Account for potential confounders.

# Practical Considerations: SEs and Heterogeneous Effects

- Standard errors:
  - Select bootstrap or other approaches suited to your data structure.
  - Be cautious with small treated samples.
- Heterogeneous effects:
  - Consider if ATT is the correct focus.
  - Explore alternative approaches for widely varying effects.

## Key Takeaway

- Synthetic DiD offers a practical, flexible, and robust approach for causal inference in complex panel data.
- Balances strengths of synthetic control and difference-in-differences while mitigating their weaknesses.
- Poised to become a valuable tool in causal panel analysis.

## Concluding Remarks: Synthesis of Methods

- Combines synthetic control (matching precision) with difference-in-differences (parallel trends).
- Addresses challenges in synthetic control's convex hull constraint.
- Regularization allows for approximate matches with intercept terms.

# Concluding Remarks: Robustness and Applications

- Doubly robust:
  - Performs well as long as unit or time weights succeed.
- Adaptable to cases where:
  - Diff-in-diff assumptions are weak.
  - Synthetic control fit is imperfect.
- Links to augmented synthetic control for two-way bias reduction.

# Practical Problems

- **Underfitting:** Cannot achieve parallel pre-treatment trends.
  - Solution: Explore more or better controls or alternative methods.
- **Omitted Variable Bias:** External factors coincide with treatment, leading to identification failure.
- **Overfitting:** Synthetic control perfectly matches pre-treatment but fails post-treatment.
  - Analogous to RDD functional form issues.

## How to Rule Out Overfitting: Oracle Weights

- Estimator matches an "oracle" that avoids overfitting by design.
- Oracle weights minimize expected squared error, not just in-sample error.
- Weights are robust to noise in the data.

## Properties of SDID

- Approximately unbiased and normally distributed under large samples.
- Optimal variance, estimable via clustered bootstrap.
- Robust to noise and systematic confounding.

## R code: synthdid

Let's look at the code together

Code: <https://github.com/synth-inference/synthdid>

Vignettes: <https://synth-inference.github.io/synthdid/articles/more-plotting.html>

# Application: Melo, Neilson and Kemboi 2023

"Indoor Vaccine Mandates in US Cities, Vaccination Behavior and COVID-19 Outcomes" by Vitor Melo, Elijah Neilson and Dorothy Kemboi, 2023 working paper

Study investigates the effect of city-level vaccine mandates (implemented in US cities) on COVID-19 cases, deaths or vaccine uptake in the cities

Authors use Arkhangelsky, et al. (2021) "synthetic difference-in-differences", as well as conventional synthetic control and difference-in-differences and finds no effect of either the announcement or implementation of the mandate had any significant effect on the outcomes

# Motivation

- Many policies and strategies were taken to incentivize citizens to get vaccinated and reduce COVID-19 spread
- Indoor vaccine mandates, one of the more restrictive, prevented people from entering public places (e.g., theaters, restaurants) without proof of vaccination
- Many large cities (NYC, San Francisco, LA, Seattle, Boston, Philadelphia) implemented with the stated goal to raise vaccination rates and slow spread and mortality from COVID-19

# Motivation

- Vaccine viewed as crucial step toward controlling the virus and return life to normal
- Substantial number of Americans were unwilling to be immunized
- February 2021, 30% of adults say they would probably or definite not be vaccinated
- Low vaccination rates led to measures to increase uptake like mandated vaccination and weekly testing, lotteries, etc.

# Mandates

- August 3, 2021, due to the Delta variant, NYC passed mandate requiring proof of vaccination to enter restaurants, concerts, stadiums and gyms
- Similar policies were adopted by other major cities soon after (see next table)
- I'll skip the prior literature for now

# Timing

Table: Timing of Indoor Vaccine Mandates

City	Announced	Implemented	Repealed
NYC	8/3/21	8/16/21	3/7/22
San Francisco	8/12/21	8/20/21	3/11/22
New Orleans	8/12/21	8/16/21	3/21/22
Seattle	9/6/21	10/25/21	3/1/22
Los Angeles	11/8/21	11/29/21	3/30/22
Philadelphia	12/13/21	1/3/22	2/16/22
Boston	12/20/21	1/15/22	2/18/22
Chicago	12/21/21	1/3/22	2/28/22
DC	12/22/21	1/15/22	2/15/22

## Research question

- Estimate an ATT for these cities' mandates on vaccination, cases and deaths
- Data will come from daily county level COVID-19 vaccinations, cases and deaths from the CDC aggregated to MSA by week scaled by US population estimates
- Main outcomes: Weekly measures of administered first doses of COVID-19 vaccines, cases, and deaths per 100,000 residents
- Weekly panel from December 21, 2020 to April 18, 2022 for 821 MSAs (they note various issues with data quality required dropping just under 100 MSAs) with 57,470 observations

# Descriptive Statistics

*Table:* Descriptive Statistics

Variable	All MSAs			Treated MSAs			Untreated MSAs		
	Mean	SD	Median	Mean	SD	Median	Mean	SD	Median
First Doses per 100,000	817.47	1,344.30	458.98	1,253.50	1,237.18	827.71	812.66	1,344.65	455.01
Cases per 100,000	273.75	373.61	147.73	247.47	394.75	121.95	274.04	373.37	148.16
Deaths per 100,000	3.56	5.87	1.90	2.03	2.31	1.17	3.58	5.90	1.91
Number of observations		57,470			630			56,840	

Notes: The unit of observation is MSA week. Our sample consists of 821 MSAs, 9 of which are treated, and the period spans 70 weeks from December 21, 2020, to April 18, 2022.

## Great discussion of synth DiD

*"The basic idea is that the unit weights are chosen to find a convex combination of potential control states whose treatment trend in the outcome variable of interest is most parallel to that of the treated state. The inclusion of the intercept term  $\omega_0$  (made possible because of the inclusion of the unit fixed effects) is one way in which the SDID unit weights differ from those of the synthetic control weights. Instead of the weights needing to make the pre-trend control unit perfectly match that of the treated unit, as is the case with the synthetic control estimator, allowing for this intercept makes it sufficient for the weights to just make the trends parallel."*

Table 3: Announcement of Indoor COVID-19 Vaccine Mandates and First-Dose Vaccine Uptake

	<i>Dependent Variable : Weekly First Doses per 100,000</i>		
	Difference-in-Differences (1)	Synthetic Control (2)	SDID (3)
<b>Panel A. Boston</b>			
Average Effect ( $\hat{\tau}$ )	319.04	-140.04	72.69
95% Confidence Interval	(-1047.25, 1685.33)	(-1211.06, 930.99)	(-1160.96, 1306.33)
<b>Panel B. Chicago</b>			
Average Effect ( $\hat{\tau}$ )	-39.95	-28.4	-172.34
95% Confidence Interval	(-1197.23, 1117.34)	(-894.70, 837.91)	(-1188.18, 843.50)
<b>Panel C. Los Angeles</b>			
Average Effect ( $\hat{\tau}$ )	-143.09	-242.61	-185.31
95% Confidence Interval	(-1142.48, 856.31)	(-740.82, 255.59)	(-966.80, 596.19)
<b>Panel D. New Orleans</b>			
Average Effect ( $\hat{\tau}$ )	-341.38	-219.38	-209.07
95% Confidence Interval	(-1642.73, 959.97)	(-724.08, 285.32)	(-721.84, 303.70)
<b>Panel E. New York</b>			
Average Effect ( $\hat{\tau}$ )	-575.97	123.77	-82.59
95% Confidence Interval	(-1907.72, 755.79)	(-398.14, 645.68)	(-605.48, 440.30)
<b>Panel F. Philadelphia</b>			
Average Effect ( $\hat{\tau}$ )	104.16	-295.41	-303.02
95% Confidence Interval	(-1148.25, 1356.57)	(-1252.58, 661.76)	(-1401.35, 795.31)
<b>Panel G. San Francisco</b>			
Average Effect ( $\hat{\tau}$ )	-1197.67*	-42.89	-195.37
95% Confidence Interval	(-2504.92, 109.58)	(-566.19, 480.41)	(-726.44, 335.71)
<b>Panel H. Seattle</b>			
Average Effect ( $\hat{\tau}$ )	-736.58	-97.14	-207.02
95% Confidence Interval	(-1978.53, 505.38)	(-688.32, 494.03)	(-840.35, 426.32)
<b>Panel I. Washington DC</b>			
Average Effect ( $\hat{\tau}$ )	-253.99	18.77	-76.53
95% Confidence Interval	(-1620.28, 1112.31)	(-1059.12, 1096.67)	(-1309.86, 1156.80)

*Notes:* This table reports the average estimated effects of announcing an indoor COVID-19 vaccine mandate on first-dose vaccine uptake as measured by weekly first doses per 100,000 residents using the difference-in-differences, the synthetic control, and the SDID estimators ( $\hat{\tau}$  from equations (2), (3), and (1)). Also reported are 95% confidence intervals using the placebo variance estimation approach outlined in section 4.2. Significance levels are reported as \*\*\* p<0.01, \*\* p<0.05, and \* p<0.1.

Table 4: Announcement of Indoor COVID-19 Vaccine Mandates and COVID-19 Cases

	<i>Dependent Variable : Weekly COVID-19 Cases per 100,000</i>		
	Difference-in-Differences (1)	Synthetic Control (2)	SDID (3)
<b>Panel A. Boston</b>			
Average Effect ( $\hat{\tau}$ )	274.32	240.05	224.57
95% Confidence Interval	(-252.03, 800.67)	(-272.99, 753.09)	(-267.13, 716.27)
<b>Panel B. Chicago</b>			
Average Effect ( $\hat{\tau}$ )	139.6	184.48	121.14
95% Confidence Interval	(-299.65, 578.84)	(-245.53, 614.49)	(-289.63, 531.91)
<b>Panel C. Los Angeles</b>			
Average Effect ( $\hat{\tau}$ )	202.06	340.28***	176.49
95% Confidence Interval	(-74.98, 479.09)	(97.34, 583.22)	(-58.08, 411.05)
<b>Panel D. New Orleans</b>			
Average Effect ( $\hat{\tau}$ )	-22.81	6.15	-27.28
95% Confidence Interval	(-216.80, 171.19)	(-182.99, 195.28)	(-217.93, 163.36)
<b>Panel E. New York</b>			
Average Effect ( $\hat{\tau}$ )	-53.04	7.02	4.62
95% Confidence Interval	(-251.54, 145.46)	(-186.94, 200.98)	(-190.87, 200.12)
<b>Panel F. Philadelphia</b>			
Average Effect ( $\hat{\tau}$ )	110.41	290.62	114.41
95% Confidence Interval	(-368.76, 589.58)	(-180.84, 762.08)	(-329.83, 558.66)
<b>Panel G. San Francisco</b>			
Average Effect ( $\hat{\tau}$ )	-107.37	65.71	-95.48
95% Confidence Interval	(-311.98, 97.24)	(-135.80, 267.22)	(-297.46, 106.50)
<b>Panel H. Seattle</b>			
Average Effect ( $\hat{\tau}$ )	19.85	20.72	-16.99
95% Confidence Interval	(-239.41, 279.12)	(-202.52, 243.95)	(-247.34, 213.36)
<b>Panel I. Washington DC</b>			
Average Effect ( $\hat{\tau}$ )	149.7	600.71	190.26
95% Confidence Interval	(-376.66, 676.05)	(86.71, 1114.72)	(-301.70, 682.22)

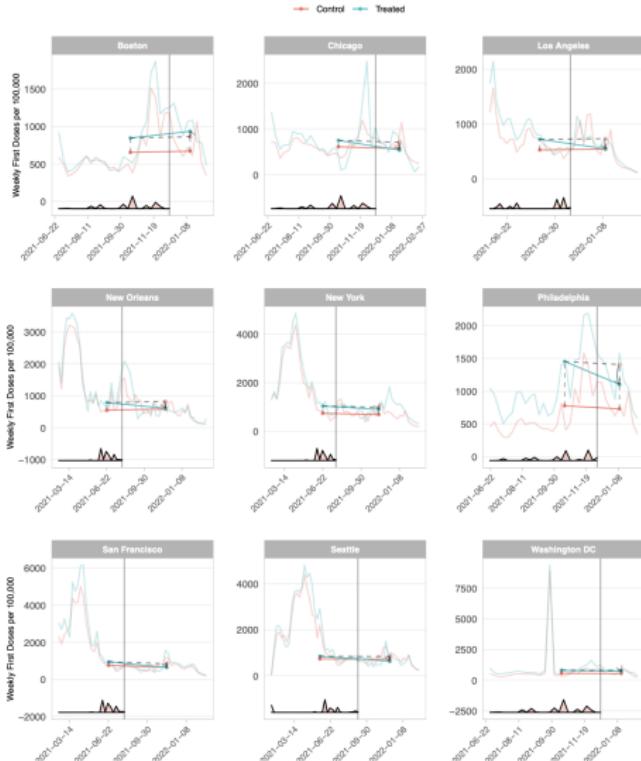
*Notes:* This table reports the average estimated effects of announcing an indoor COVID-19 vaccine mandate on the number of weekly COVID-19 cases per 100,000 residents using the difference-in-differences, the synthetic control, and the SDID estimators ( $\hat{\tau}$  from equations (2), (3), and (1)). Also reported are 95% confidence intervals using the placebo variance estimation approach outlined in section 4.2. Significance levels are reported as \*\*\* p<0.01, \*\* p<0.05, and \* p<0.1.

Table 5: Announcement of Indoor COVID-19 Vaccine Mandates and COVID-19 Deaths

	<i>Dependent Variable : Weekly COVID-19 Deaths per 100,000</i>		
	Difference-in-Differences (1)	Synthetic Control (2)	SDID (3)
<b><i>Panel A. Boston</i></b>			
Average Effect ( $\hat{\tau}$ )	2.32	1.65	1.38
95% Confidence Interval	(-4.75, 9.39)	(-5.97, 9.28)	(-4.76, 7.53)
<b><i>Panel B. Chicago</i></b>			
Average Effect ( $\hat{\tau}$ )	1.94	1.46	1.39
95% Confidence Interval	(-4.21, 8.09)	(-5.10, 8.03)	(-4.06, 6.84)
<b><i>Panel C. Los Angeles</i></b>			
Average Effect ( $\hat{\tau}$ )	-0.2	0.67	-0.3
95% Confidence Interval	(-5.26, 4.86)	(-3.85, 5.19)	(-4.52, 3.92)
<b><i>Panel D. New Orleans</i></b>			
Average Effect ( $\hat{\tau}$ )	-0.65	-2.5	-1.37
95% Confidence Interval	(-4.48, 3.18)	(-6.07, 1.07)	(-4.96, 2.22)
<b><i>Panel E. New York</i></b>			
Average Effect ( $\hat{\tau}$ )	-2.37	-2.66	-1.91
95% Confidence Interval	(-6.16, 1.43)	(-6.09, 0.76)	(-5.42, 1.60)
<b><i>Panel F. Philadelphia</i></b>			
Average Effect ( $\hat{\tau}$ )	2.76	-2.16	2.21
95% Confidence Interval	(-4.11, 9.63)	(-9.35, 5.02)	(-3.69, 8.11)
<b><i>Panel G. San Francisco</i></b>			
Average Effect ( $\hat{\tau}$ )	-2.19	-4.72**	-2.66
95% Confidence Interval	(-6.14, 1.76)	(-8.51, -0.93)	(-6.36, 1.04)
<b><i>Panel H. Seattle</i></b>			
Average Effect ( $\hat{\tau}$ )	-1.07	-1.08	-1.43
95% Confidence Interval	(-5.04, 2.91)	(-4.63, 2.47)	(-5.09, 2.22)
<b><i>Panel I. Washington DC</i></b>			
Average Effect ( $\hat{\tau}$ )	0.46	-0.92	0.2
95% Confidence Interval	(-6.61, 7.53)	(-8.55, 6.70)	(-5.95, 6.35)

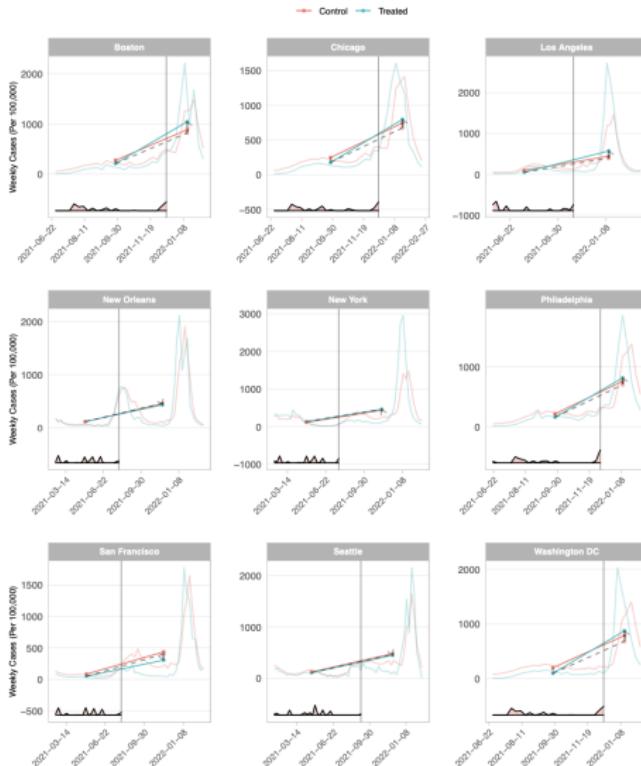
*Notes:* This table reports the average estimated effects of announcing an indoor COVID-19 vaccine mandate on the number of weekly COVID-19 deaths per 100,000 residents using the difference-in-differences, the synthetic control, and the SDID estimators ( $\hat{\tau}$  from equations (2), (3), and (1)). Also reported are 95% confidence intervals using the placebo variance estimation approach outlined in section 4.2. Significance levels are reported as \*\*\* p<0.01, \*\* p<0.05, and \* p<0.1.

Figure 1: Trends in Weekly First Doses per 100,000 in Treated MSAs and Their Respective Synthetic Controls



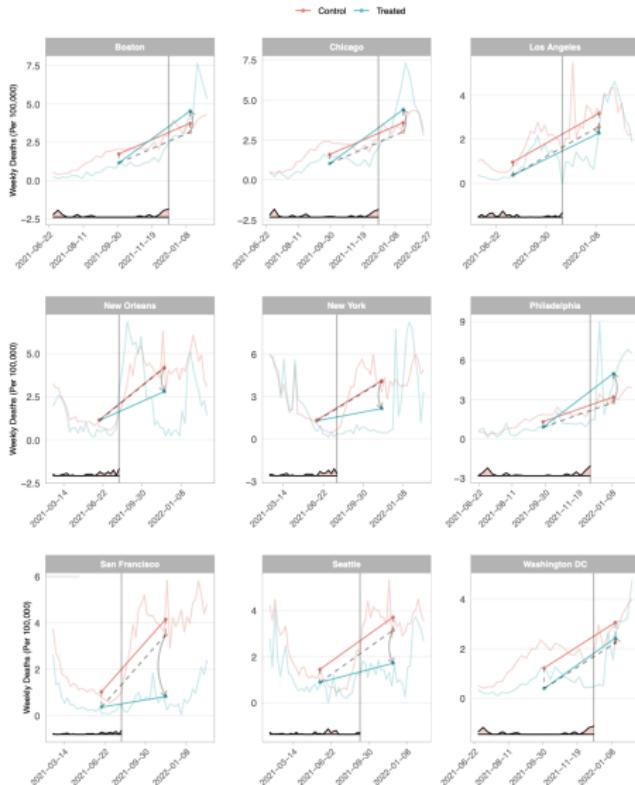
*Notes:* Each plot shows trends in weekly first doses of COVID-19 vaccinations per 100,000 residents for each MSA that adopted an indoor vaccine mandate and for their corresponding synthetic control. The weights used to average pre-treatment time periods are shown at the bottom of the plots. The curved arrows indicate the estimated average treatment effect ( $\hat{\tau}$  from equation (1)) and the vertical lines represent the week each MSA announced their vaccine mandate.

Figure 2: Trends in Weekly COVID-19 Cases per 100,000 in Treated MSAs and Their Respective Synthetic Controls



*Notes:* Each plot shows trends in weekly COVID-19 cases per 100,000 residents for each MSA that adopted an indoor vaccine mandate and for their corresponding synthetic control. The weights used to average pre-treatment time periods are shown at the bottom of the plots. The curved arrows indicate the estimated average treatment effect ( $\hat{\tau}$  from equation (1)) and the vertical lines represent the week each MSA announced their vaccine mandate.

Figure 3: Trends in Weekly COVID-19 Deaths per 100,000 in Treated MSAs and Their Respective Synthetic Controls



*Notes:* Each plot shows trends in weekly COVID-19 deaths per 100,000 residents for each MSA that adopted an indoor vaccine mandate and for their corresponding synthetic control. The weights used to average pre-treatment time periods are shown at the bottom of the plots. The curved arrows indicate the estimated average treatment effect ( $\hat{f}$  from equation (1)) and the vertical lines represent the week each MSA announced their vaccine mandate.

# Conclusion

- They also report synth and DiD analysis as robustness – something to keep in mind is the presentation of results are subjective
- Rather than showing regression results with more controls, we tend to now see different DiD and synth estimators as the robustness
- Authors fail to find strong evidence the vaccine mandates slowed COVID-19

# Synthetic Control with Staggered Adoption

- Synth was originally designed for a single treated unit, no extrapolation, non-negative weights summed to one
- Previous Ben-Michael, Feller and Rothstein (2021a) paper addressed imperfect fit in the pre-trends using bias correction and slightly negative weighting
- This new paper (Ben-Michael, Feller and Rothstein 2021b) focuses on multiple units by allowing differential timing
- This and matrix completion with nuclear norm regularization seem to be relevant for the new differential timing papers in diff-in-diff

# Synthetic Control with Staggered Adoption

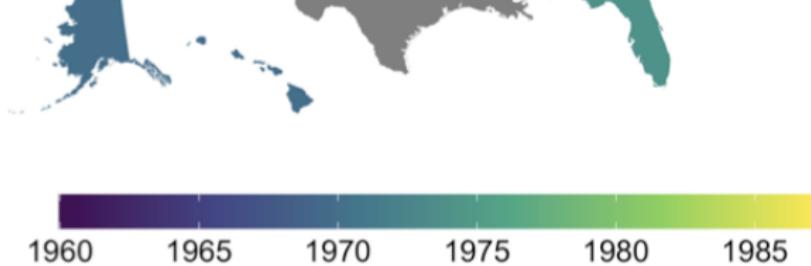
- Groups of units are treated, but they are treated at different time periods
- Standard approach is DiD and event studies
  - Tons of papers recently (e.g., Callaway and Sant'Anna 2021; Sun and Abraham 2021; de Chaisemartin D'Haultfouille 2020; Borusyak, et al. 2023)
  - Identifying assumption in all DiD papers is *parallel trends*
- When parallel trends is not viable, then ATT estimate is biased by a non-parallel trends bias term
- Synthetic control methods were updated to accommodated multiple treated groups

## Synthetic Control with Staggered Adoption

- Core idea of the paper: find a coherent way to manage multiple synthetic controls and aggregating them into a single parameter estimate
- Goal is to balance the imperfect biases in the pooled and the separate unit-level estimates
- Their working example will be a teacher union collective bargaining law and teacher salary study

# Estimating effects under staggered adoption

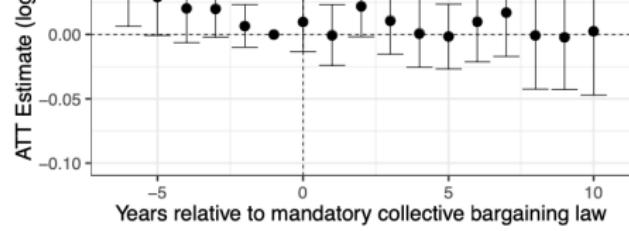
- **Staggered adoption:** Multiple units adopt treatment over time
- **Common approaches can fail:** Little guidance when this happens
  - Diff-in-diff requires parallel trends assumption
  - Synth designed for single treated units, poor fit on average
- **Partially pooled synthetic control**
  - Modify the optimization problem to target overall and state-specific fit
  - Account for the level differences with an intercept shift



# Teacher unions and teacher salaries/spending

Their application is about teacher unions

- 1964-1987: 33 states grant collective bargaining rights to teachers
- Long literature exploited the timing (Hoxby 1996; Lovenheim 2009)
- Impact on student spending, teacher salaries
  - Hoxby (1996) finds increased spending by 12%
  - Paglayan (2019) estimates precise zero in an event study model using ever-treated states
- Traditionally this was done using two-way fixed effects and event studies but these are known to have bias with heterogeneous treatment effects (Goodman-Bacon 2021)
- They're going to re-analyze using all states and synth models and we can review it too using their code



## Two approaches before now vs theirs

1. Separate Synthetic Control – Donohue, et al. 2019 estimates separate synth model for each state that passed concealed carry laws then averaged the estimates;
  - only works if you can find good for each one obviously
  - Can lead to poor fit for the average leading to bias when the average treatment effect is the target parameter
2. Pooled Synthetic Control – Minimizes the average pre-treatment imbalance across all treated units
  - Can achieve nearly perfect fit for the average treated unit
  - Can yield substantially worse unit-specific fits
3. Partially pooled (their proposal) – minimizes a weighted average of the two imbalances

# Intuition of the Partially Pooled Approach

We want to balance the average of the underlying factor loadings

- Balancing individual units may cause large imbalance in the average if errors all go in the same direction
- Balancing the average outcome may not balance factor loadings if imbalance for different treated units offset each other
- Trade these off one another

## Average separate synths (separate SCM)

- Suppose the first  $J$  units are treated at times  $T_1, \dots, T_J$
- Suppose we find a synthetic control for each, with  $w_{ij}$  the weight on donor unit  $i$  for treated unit  $j$
- Our estimate of the ATT at event time  $k$  will then be

$$\hat{\delta} = \frac{1}{J} \sum_{j=1}^{J+1} \left( Y_{j,T_j+k} - \sum_i w_{ij} Y_{i,T_j+k} \right)$$

Average of  $J$  separate synth estimates

## Average treated unit (pooled SCM)

Alternatively, we can think of it as Synth estimate for average treated unit which they call the pooled SCM

$$\hat{\delta} = \frac{1}{J} \sum_{j=2}^{J+1} Y_{j,T_j+k} - \frac{1}{J} \left( \sum_{j=2}^{J+1} \sum_i w_{ij} Y_{i,T_j+k} \right)$$

## Two definitions of ATT

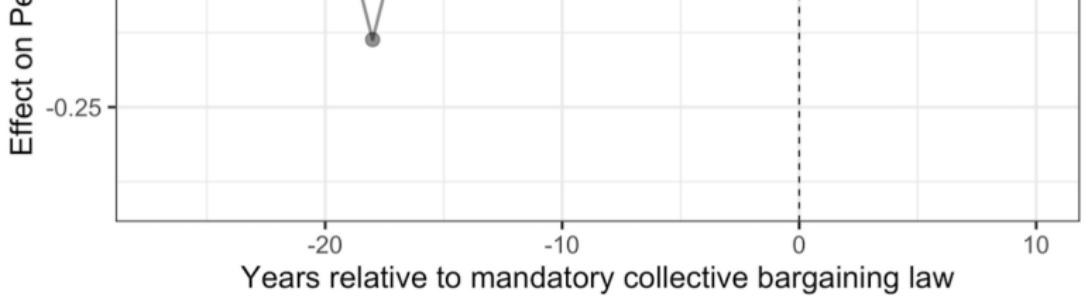
$$\begin{aligned}\hat{\delta}_{Separate\;SCM} &= \frac{1}{J} \sum_{j=1}^J \left( Y_{j,T_j+k} - \sum_i w_{ij} Y_{i,T_j+k} \right) \\ \hat{\delta}_{Pooled\;SCM} &= \frac{1}{J} \sum_{j=2}^{J+1} Y_{j,T_j+k} - \frac{1}{J} \left( \sum_{j=2}^{J+1} \sum_i w_{ij} Y_{i,T_j+k} \right)\end{aligned}$$

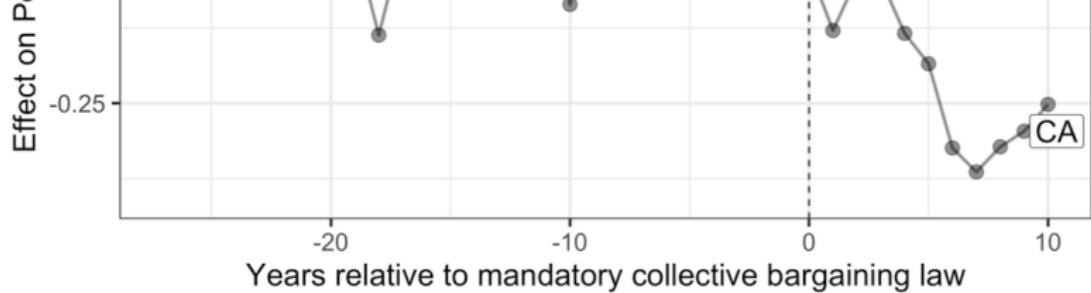
## Returning to that optimization problem

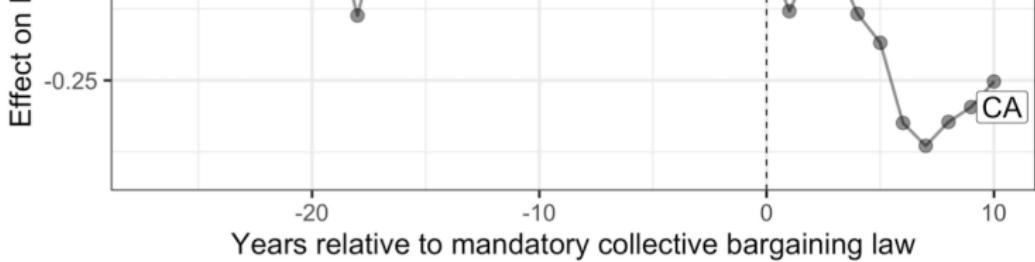
So they ask: do we want to optimize the sum of the separate imbalances from the separate SCM or the imbalance of the sum (the pooled imbalance) from the pooled SCM?

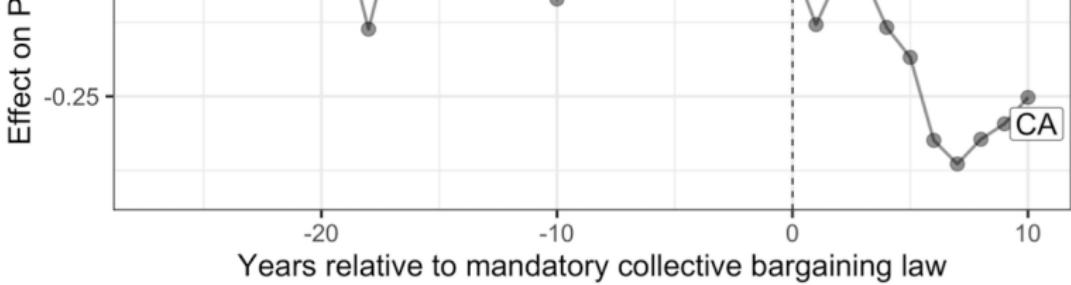
$$\sum_{j=2}^{J+1} \left\| X_j - \sum_i w_{ij} X_i \right\|^2 \text{ or } \left\| \sum_{j=2}^{J+1} X_j - \sum_i w_{ij} X_i \right\|^2$$

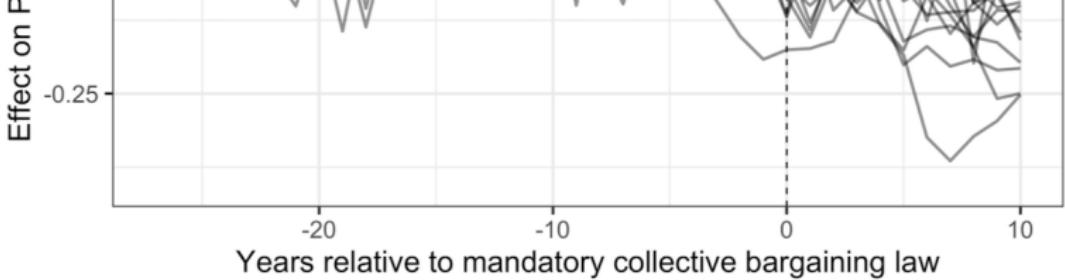
where  $j$  is treatment group and  $i$  is donor pool units. Notice summations are inside or outside the norm. You will get different solutions obviously.

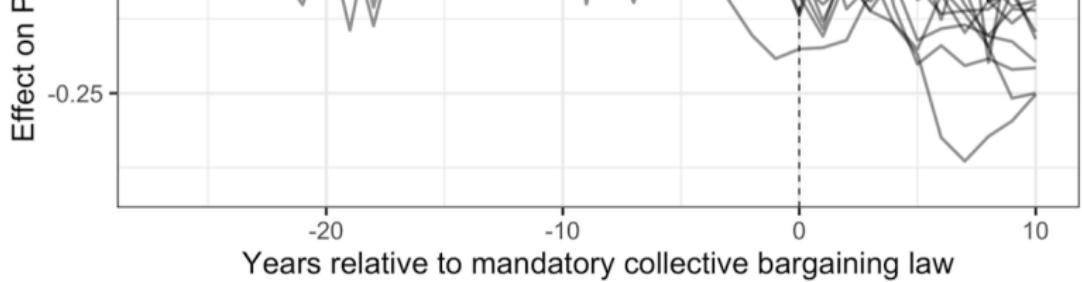


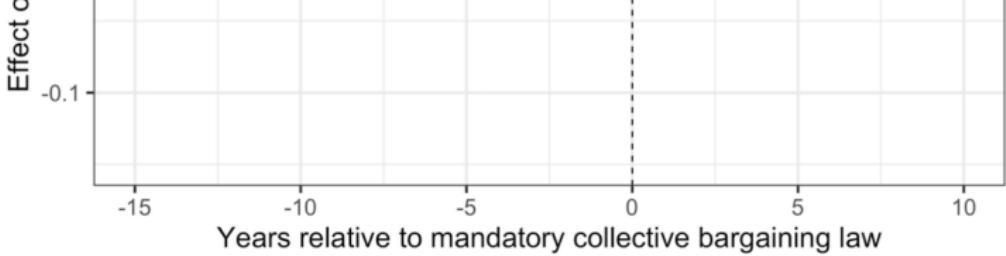


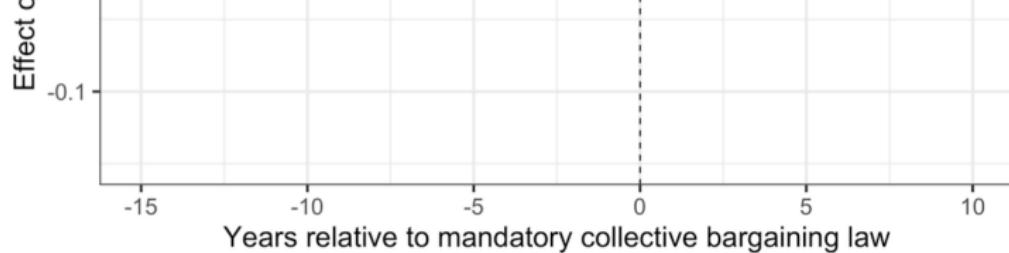


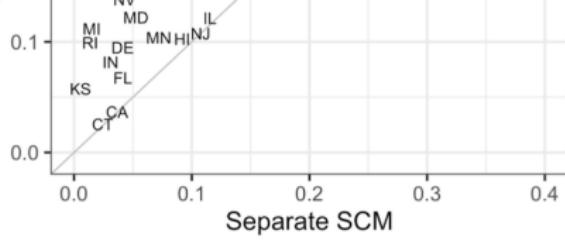


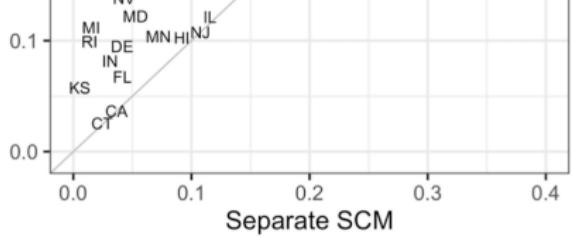












- E.g., different factors are different in different time periods.
- Once time is re-centered, matching the average need not mean matching the factors.

## Proposal: Partially pool synth

Instead of minimizing pooled imbalance or average state imbalance, minimize a *weighted average*:

$$\begin{aligned} \min_{\Gamma \in \Delta^{synth}} \quad & v \|\text{Pooled balance}\|_2^2 \\ & + (1 - v) \frac{1}{J} \sum_{j=2}^{J+1} \|\text{State balance}\|_2^2 \\ & + \text{penalty} \end{aligned}$$

“Returns” to this are highly convex: setting  $v$  just a little below 1 yields a big improvement in state-level imbalance with very little cost in pooled imbalance

## $v$ hyperparameter

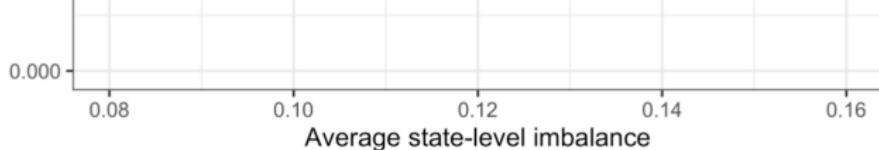
- $v$  is the hyperparameter  $v \in [0, 1]$
- It governs the relative importance of the two objectives;
- higher values of  $v$  correspond to more weight on the pooled fit relative to the separate fit
- When  $v = 0$ , it's the separate SCM and when it's  $v = 1$  it's the pooled SCM

## Intermediate choice of $v$

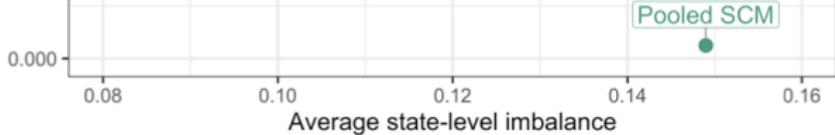
- It is important to control both the pooled fit (for the ATT) and the unit-level fits (for both the ATT and the unit-level estimates)
- The hyper-parameter  $v$  controls the relative weight of these in the objective
- In general, we want to find good estimates of both the overall ATT and the unit-level effects

## Intermediate choice of $v$

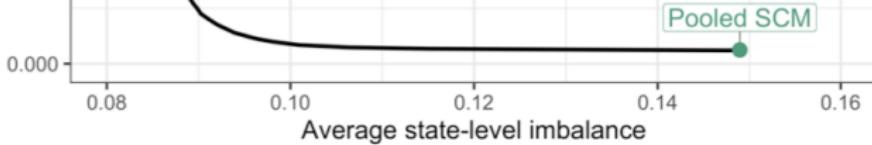
- Following figures is the balance possibility frontier: the y-axis shows the pooled imbalance and the x-axis shows the unit-level imbalance
- The curve traces how these change as we vary  $v$  from the SCM solution upper left to the pool lower right
- Strongly convex relationship which means we can accept a very small increase in pooled imbalance from the pooled solution and get large reductions in unit-level imbalance and vice versa



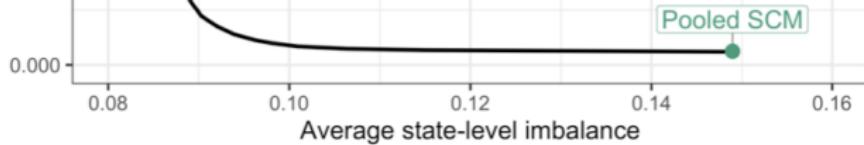
$$\min_{\Gamma \in \Delta^{\text{scm}}} \quad \nu \|\text{Pooled Balance}\|_2^2 + (1 - \nu) \frac{1}{J} \sum_{j=1}^J \|\text{State Balance}_j\|_2^2 + \text{penalty}$$



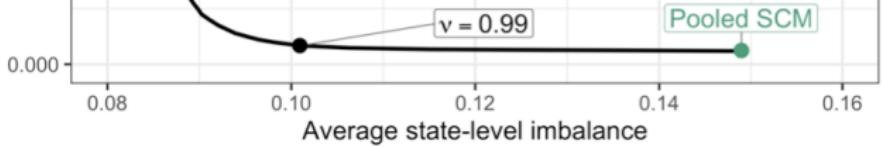
$$\min_{\Gamma \in \Delta^{\text{scm}}} \quad \nu \|\text{Pooled Balance}\|_2^2 + (1 - \nu) \frac{1}{J} \sum_{j=1}^J \|\text{State Balance}_j\|_2^2 + \text{penalty}$$



$$\min_{\Gamma \in \Delta^{\text{scm}}} \quad \nu \|\text{Pooled Balance}\|_2^2 + (1 - \nu) \frac{1}{J} \sum_{j=1}^J \|\text{State Balance}_j\|_2^2 + \text{penalty}$$



$$\min_{\Gamma \in \Delta^{\text{scm}}} \quad \nu \|\text{Pooled Balance}\|_2^2 + (1 - \nu) \frac{1}{J} \sum_{j=1}^J \|\text{State Balance}_j\|_2^2 + \text{penalty}$$



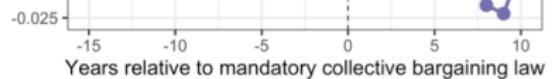
$$\min_{\Gamma \in \Delta^{\text{scm}}} \quad \nu \|\text{Pooled Balance}\|_2^2 + (1 - \nu) \frac{1}{J} \sum_{j=1}^J \|\text{State Balance}_j\|_2^2 + \text{penalty}$$

## Interpreting

- So even going from  $v = 1$  to  $v = 0.99$  cuts the unit-level imbalance by 30 percent with almost no change in the pooled fit
- In many cases, it will be possible to trade off a small increase in pooled imbalance for a large decrease in unit-level imbalance
- This would yield a better estimator of the overall ATT and the unit-level estimates at little cost
- The balance possibility frontier is a tool you use to try and trace out the trade-offs between the pooled and unit-level fit and choose which ever  $v$  they want

## Heuristic for $v$

- They use a simple heuristic for choosing  $v$
- The ratio of the pooled fit to the average unit-level fit
- The key idea is that if the separate problem with  $v = 0$  achieved good pooled fit on its own, then you want a small  $v$  which ensures good unit and pooled fit
- If the pooled fit of separate is poor, then there can be substantial gains to giving the pooled higher priority and setting  $v$  large



## Augment staggered adoption

1. Estimate an outcome model
2. Estimate the partially pooled synth model
3. Use the outcome model to adjust synth for imbalance (bias correction) or alternatively just use synth on the residuals from the outcome model (double robust)

## Special case: weighted event study

- Estimate unit fixed effects via pre-treatment average:  $\bar{Y}_{i,T_j}^{pre}$
- Estimate synth using residuals (Doudchenko and Imbens 2017; Ferman and Pinto 2018)

$$\hat{Y}_{j,T_j+k}^{aug}(0) = \bar{Y}_{j,T_j}^{pre} + \sum_{i=1}^N \hat{w}_{ij} \left( Y_{i,T_j+k} - \bar{Y}_{i,T_j}^{pre} \right)$$

where  $Y(0) = Y^0$

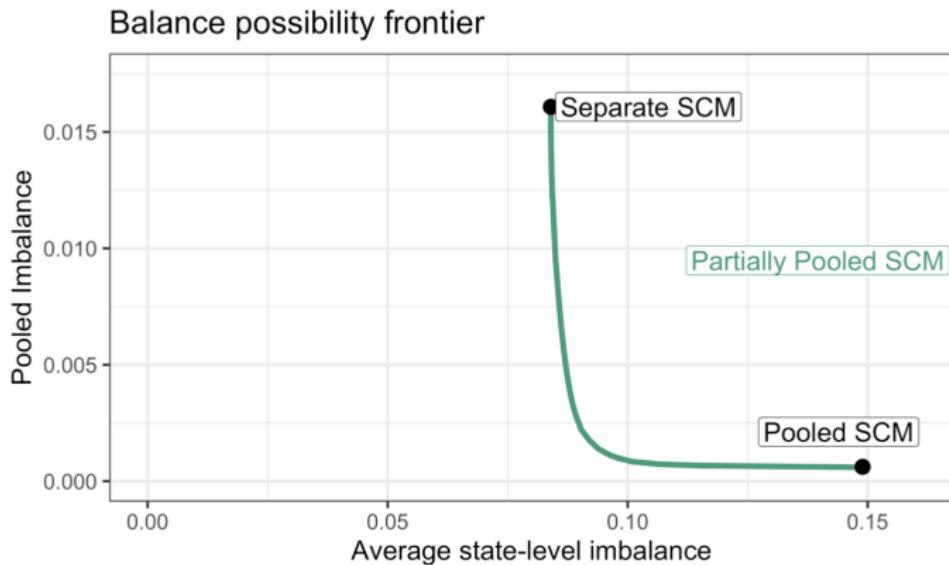
## Special case: weighted event study

Treatment effect estimate is **weighted diff-in-diff**:

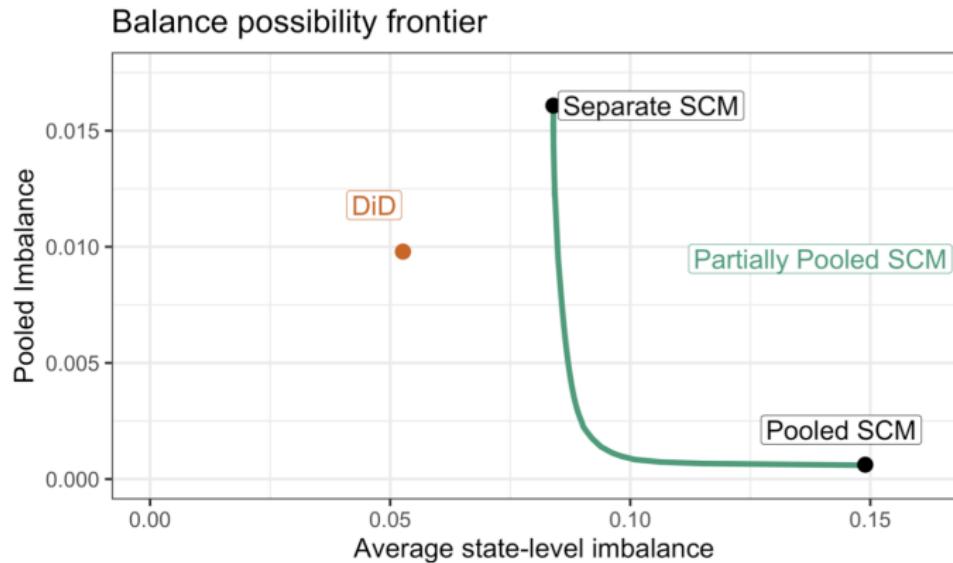
$$\hat{\delta}_{jk}^{aug} = \left( Y_{j,T_j+k} - \bar{Y}_{j,T_j}^{pre} \right) - \sum_{i=1}^N \hat{w}_{ij} \left( Y_{i,T_j+k} - \bar{Y}_{i,T_j}^{pre} \right)$$

Uniform weights correspond to "standard DiD"

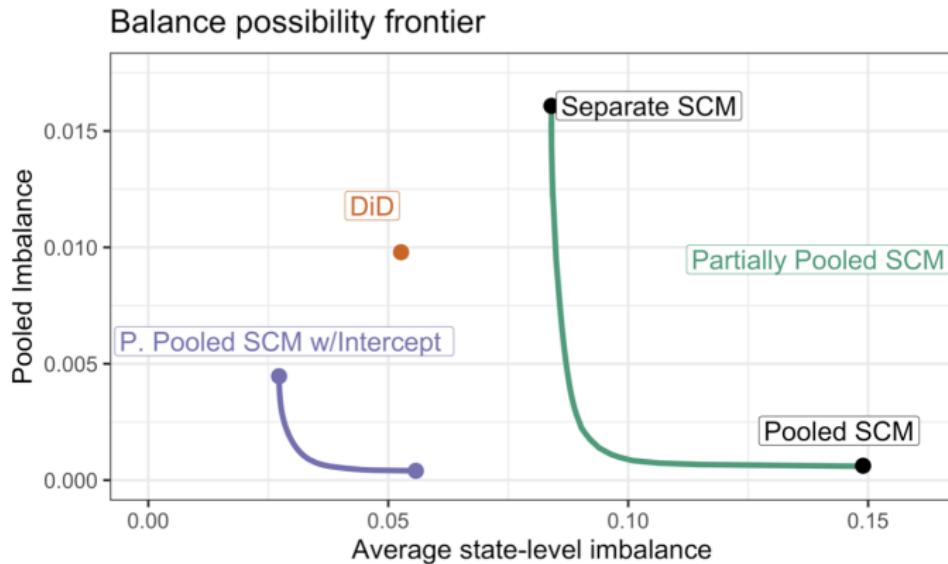
Weighted event studies shift the balance possibility frontier far inward



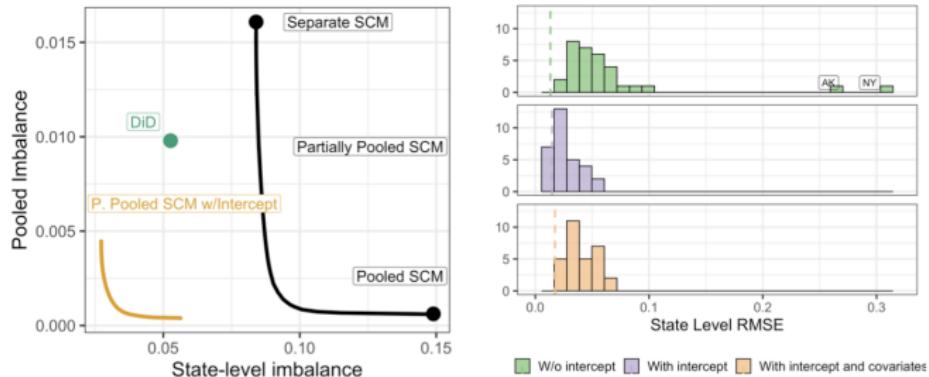
Weighted event studies shift the balance possibility frontier far inward



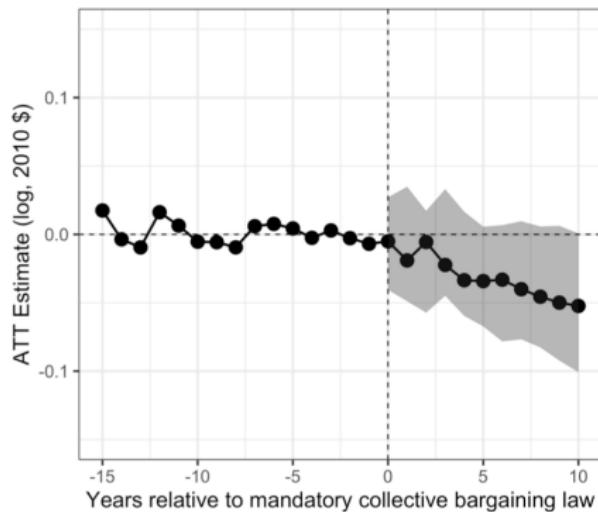
Weighted event studies shift the balance possibility frontier far inward



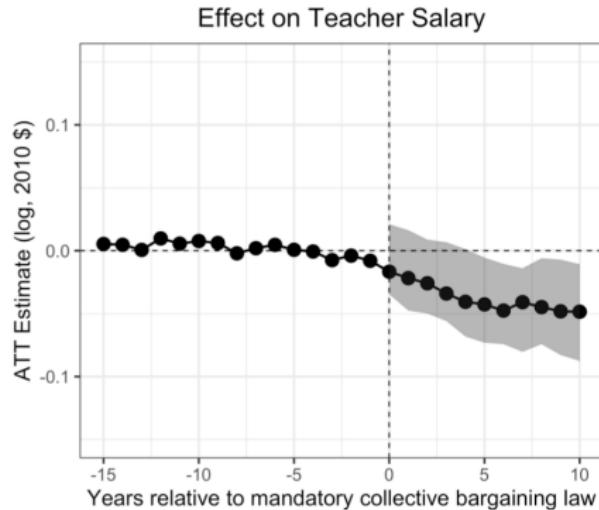
Weighted event studies shift the balance possibility frontier far inward



When we construct an adequate comparison group, the event study estimate of the effect on spending is zero or negative



When we construct an adequate comparison group, the event study estimate of the effect on teacher salaries is negative



# Conclusions

- Synth is useful for very difficult problems in which parallel trends is implausible
- With large  $T$  and perfect balance, you can use synth to get approximately unbiased treatment effect estimates under reasonable DGPs (we saw in the original ADH)
- But perfect balance is a unicorn and doesn't happen in most settings
- What do we do when it doesn't? Give up? Salvage the estimates somehow? How?

# Conclusions

- Augmented synth allows us to salvage the method, using an outcome model to remove bias from imperfect balance
- Partially pooled synth allows extension to the staggered adoption setting
- Combining the two methods gives us the best hope
  - A simple fixed effect outcome model leads to a weighted event study
  - This generalizes recent recommendations for two-way fixed effects

# Summarizing

- Synthetic control was developed for the comparative case study; it is a kind of matching estimator with an underlying factor model for its identification (not parallel trends)
- Advancements have been made along multiple dimensions – bias adjustments, demeaning, as well as exploring more general structures than just factor models
- It is now a more robust, general causal panel method but the assumptions needed to justify it need "due diligence"

## Closing remark

- Focus on the treatment assignment mechanisms carefully to help understand how unobserved time varying confounders may be threatening your results, pay close attention to issues around observable matching bias, remember the importance of the "long panel"
- Extrapolation based on the negative weighting should be done with the idea of bias reduction (augmented ridge), not simply for the purpose of fitting
- Good luck!