Causal Inference II

Mixtape Session

Roadmap

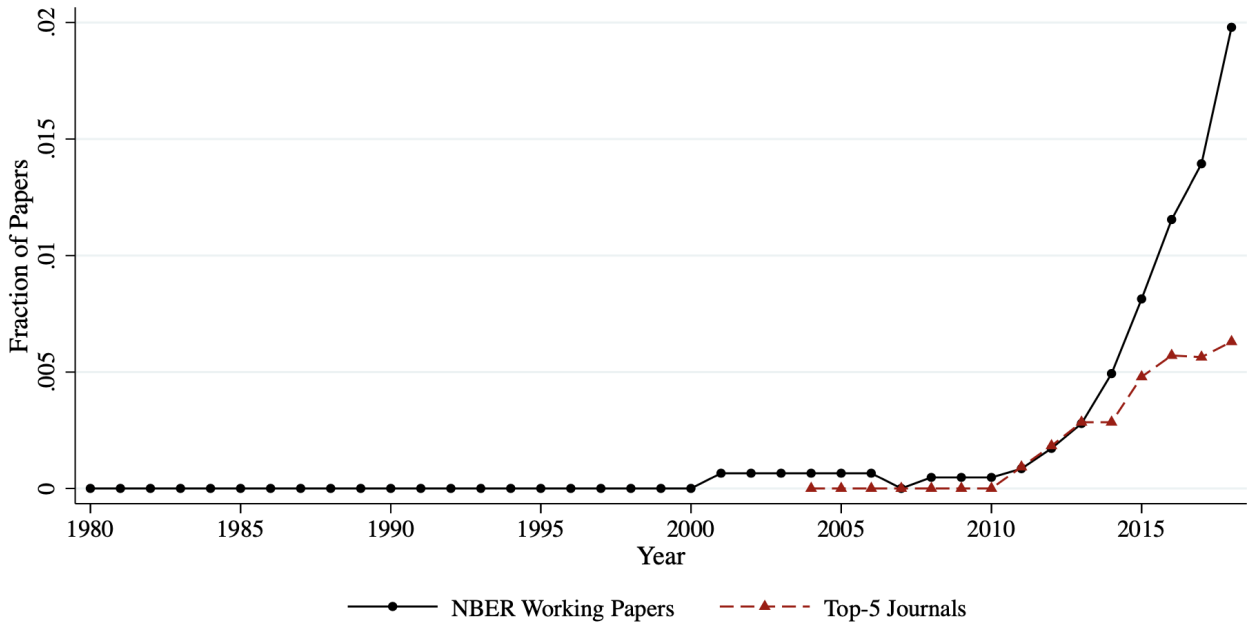
- Synthetic control
  - Abadie, Diamond and Hainmueller
  - Matrix completion with nuclear norm
  - Synthetic difference-in-differences
  - Augmented Synthetic Control
  - Augmented Synthetic Control with Staggered Rollout
  - Event studies in finance

- Concluding remarks

Day 2: Synthetic Control

- Now we move into the realm of how to handle perfect doctor scenarios without randomization
- Largely going to be a focus on modeling outcomes with panel data today
- Key characteristics will be small number of units but long panels
- We will also look at the situation when “fit” for single treated unit is good (ADH) versus bad (augmented)

D: Synthetic Control



What is synthetic control

- Synthetic control has been called the most important innovation in causal inference of the last two decades (Athey and Imbens 2017)
- Originally designed for comparative case studies, but newer developments have extended it to multiple treated units as well as differential timing
- Continues to also be methodologically a frontier for applied econometrics, so consider this talk a starting point for you

What is a comparative case study

- Single treated unit – country, state, firm
- Social scientists tackle such situations in two ways: qualitatively and quantitatively
- In political science, probably others, you see as a stark dividing line between camps, (economics doesn't have a qualitative tradition)

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.
  - → May not answer the causal questions at all because there is rarely a counterfactual, or if so, it's ad hoc.
  - → Classic example of comparative case study approach is Alexis de Toqueville's Democracy in America (but he is regularly comparing the US to France)
- You won't find someone claiming that some event caused GDP to fall $1500 when compared against France in qualitative

Traditional quantitative comparative case studies

- Traditional quantitative comparative case studies are explicitly causal designs in that there is a treatment and control
- Usually treatment is based on a natural experiment applied to a single aggregate unit (e.g., city, school, firm, state, country)
- Method compares 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 and cons

- 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:
  - → 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

- Card (1990) uses the Mariel Boatlift of 1980 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?
- The Mariel Boatlift brought 100,000 Cubans to Miami which increased the Miami labor force by 7%
- Individual-level data on unemployment from the Current Population Survey (CPS) for Miami and comparison cities


EL DORADO


EL DORADO


SILVER SEA

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.

Card's main results used diff-in-diff

Differences-in-differences estimates of the effect of immigration on unemployment[a]

	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)

[a] Notes: Adapted from Card (1990, Tables 3 and 6). Standard errors are shown in parentheses.

Parallel trends

- His estimate is unbiased if the change in Y^0 for the comparison cities correctly approximates the unobserved ΔY^0 for the treatment group
- But Card largely focused on covariates, and in a relatively casual way ("similar growth") – this predates using event studies (also Card was not particularly impressed by the study so didn't seem to really expect it to make much of a splash)
- The Black result would have been positive, too, were it not that the comparison cities growth was smaller
- Is there anything principled we could do that doesn't give the researcher so much control over control group?

Synthetic Control

- Abadie and Gardeazabal (2003) introduces synthetic control in a study of a terrorist attack in Spain (Basque) 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)
- Paper has over 5000 cites and is growing in influence over time – very popular in tech
- A combination of comparison units often does a better job reproducing the characteristics of a treated unit than single comparison unit alone

Researcher's objectives

- 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

- Precludes extrapolation (unlike regression) because counterfactual forms a convex hull
- Does not require access to post-treatment outcomes in the “design” phase of the study - no peeking
- Makes explicit the contribution of each comparison unit to the counterfactual
- Formalizing the way comparison units are chosen has direct implications for inference

Synthetic control method: disadvantages

1. Subjective researcher bias kicked down to the model selection stage
2. Significant diversity at the moment as to how to principally select models - from machine learning to modifications - as well as estimation and software

Avoiding cherry picking synthetic controls

- Ferman, Pinto and Possbaum (2020) note that there's a ton of diversity in how the models are fit
- Opens the door for "cherry picking" and remember – part of the purpose of this procedure is to reduce subjective researcher bias
- They conducted Monte Carlo simulations and make several recommendations for specifications and reporting all of them

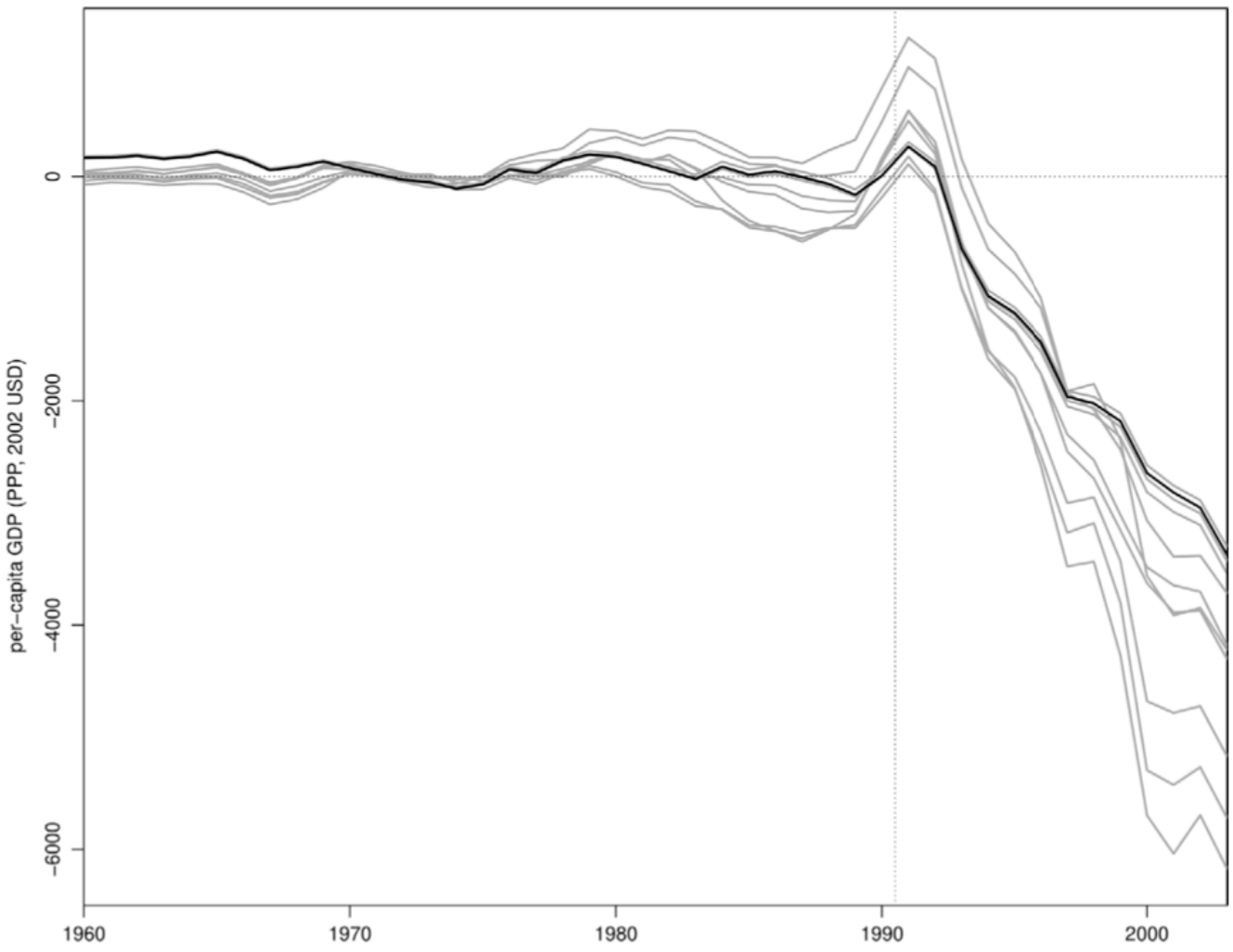
Avoiding cherry picking

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.

Avoiding cherry picking



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 period.

Synthetic control method: estimation

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

- Unit "one" is exposed to the intervention of interest (that is, "treated") during periods $T_0 + 1, \ldots, T$
- The remaining J 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.

Dynamic ATT

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

Estimating optimal weights

- Let $W = (w_2, \ldots, w_{J+1})'$ with $w_j \geq 0$ for $j = 2, \ldots, J+1$ and $w_2 + \cdots + w_{j+1} = 1$. Each value of 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 $W^* = (w_2^*, \ldots, w_{J+1}^*)'$ 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

More on the V matrix

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

- 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

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?

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

where α_t is an unknown common factor with constant factor loadings across units, and λ_t is a vector of unobserved common factors

With some manipulation

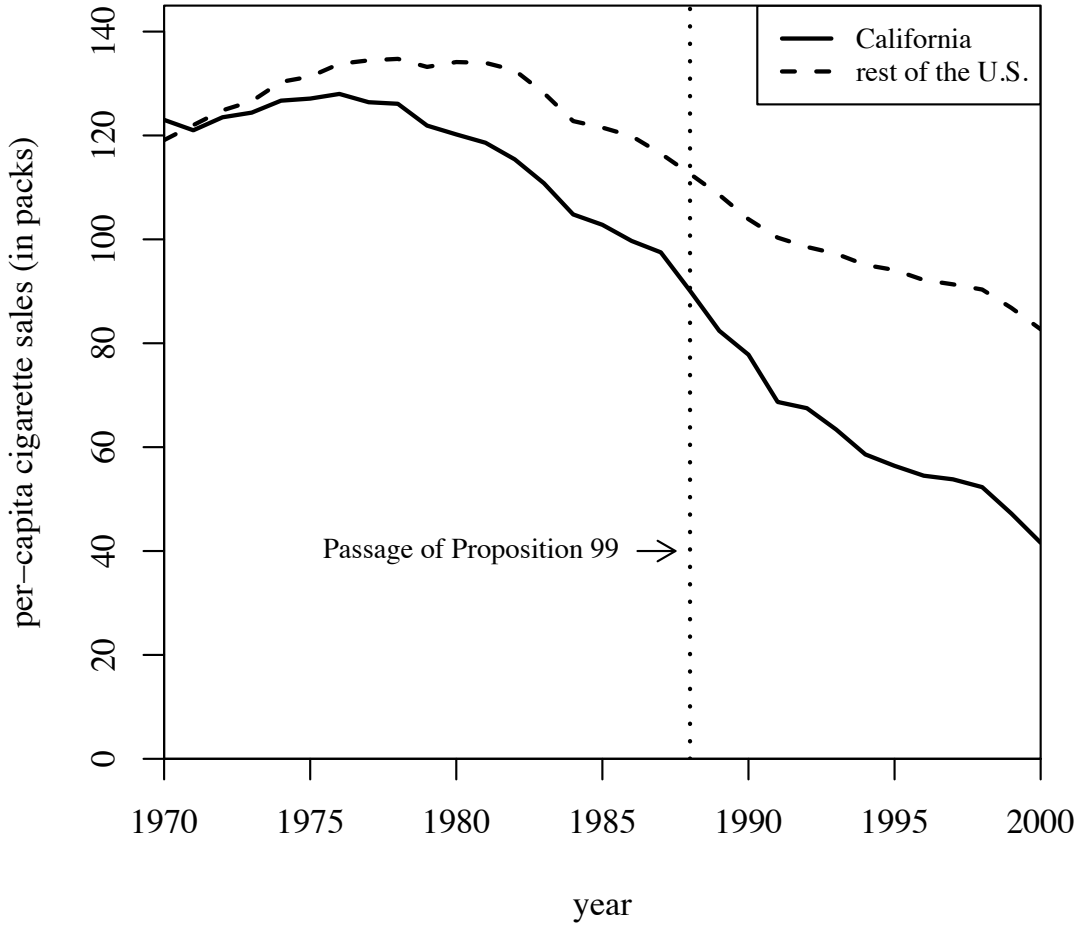
$$Y_{1t}^0 - \sum_{j=2}^{J+1} w_j^* Y_{jt} = \sum_{j=2}^{J+1} w_j^* \sum_{s=1}^{T_0} \lambda_t \left(\sum_{n=1}^{T_0} \lambda_n' \lambda_n \right)^{-1} \lambda_s' (\varepsilon_{js} - \varepsilon_{1s}) - \sum_{j=2}^{J+1} w_j^* (\varepsilon_{jt} - \varepsilon_{1t})$$

- 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
- Proof in Appendix B of ADH (2011)

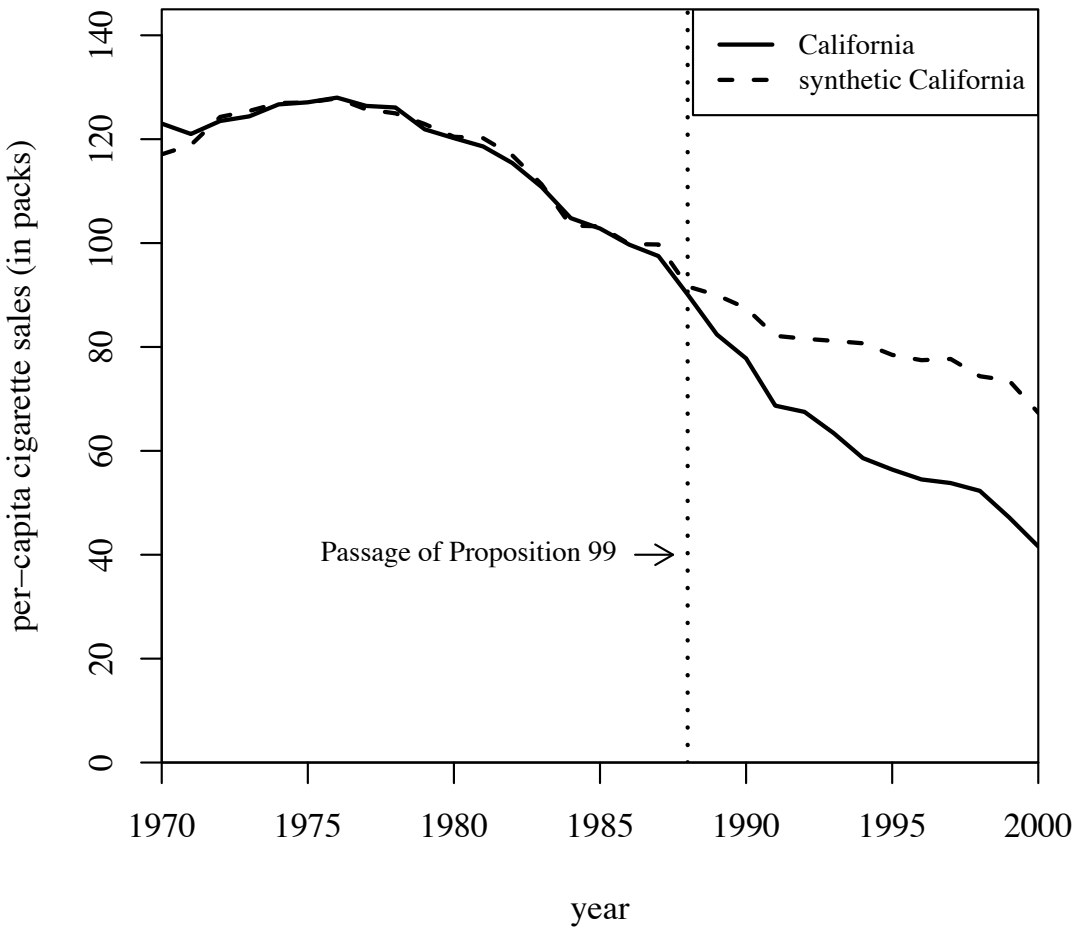
Example: California's Proposition 99

- 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

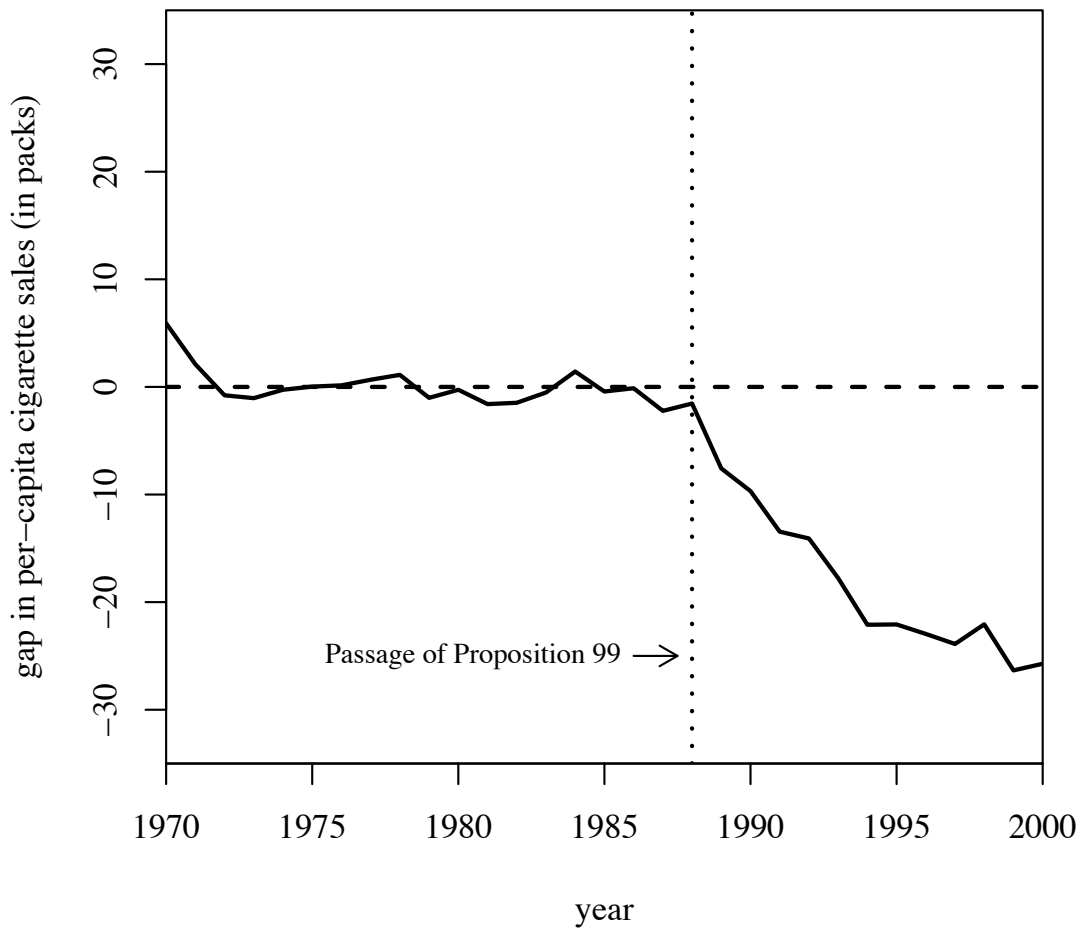


Predictor Means: Actual vs. Synthetic California

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

Note: 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



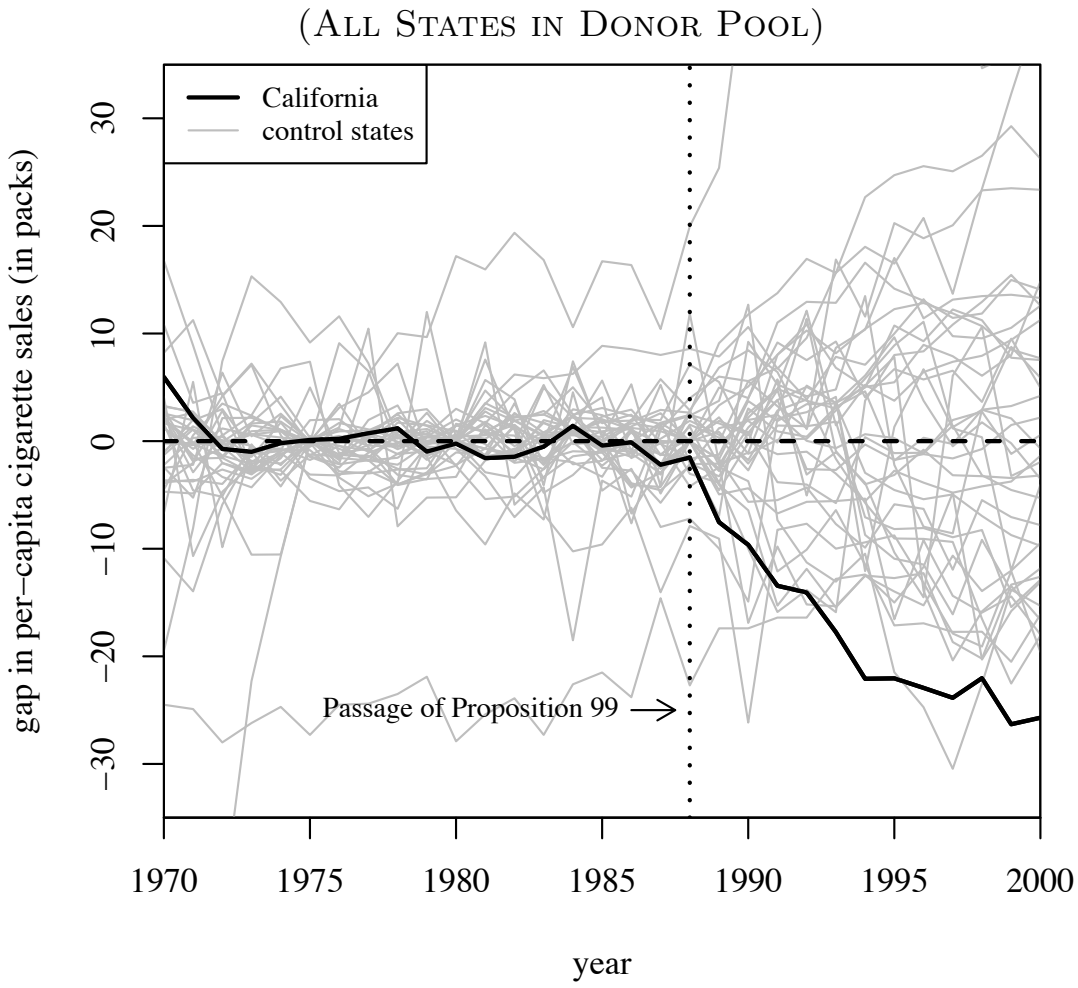
Inference

- 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:

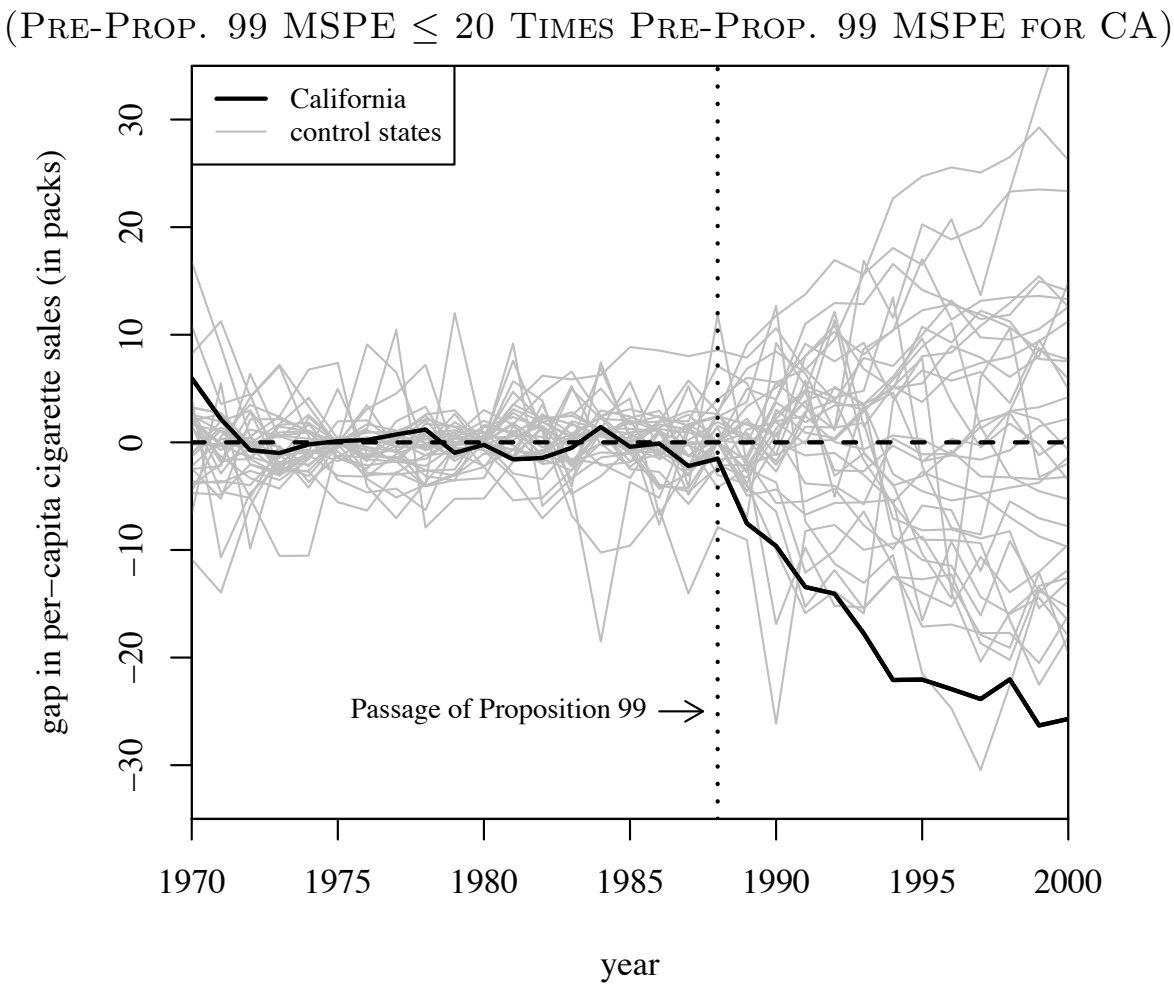
$$RMSPE = \left(\frac{1}{T - T_0} \sum_{t=T_0+1}^{T} \left(Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt} \right)^2 \right)^{\frac{1}{2}}$$

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

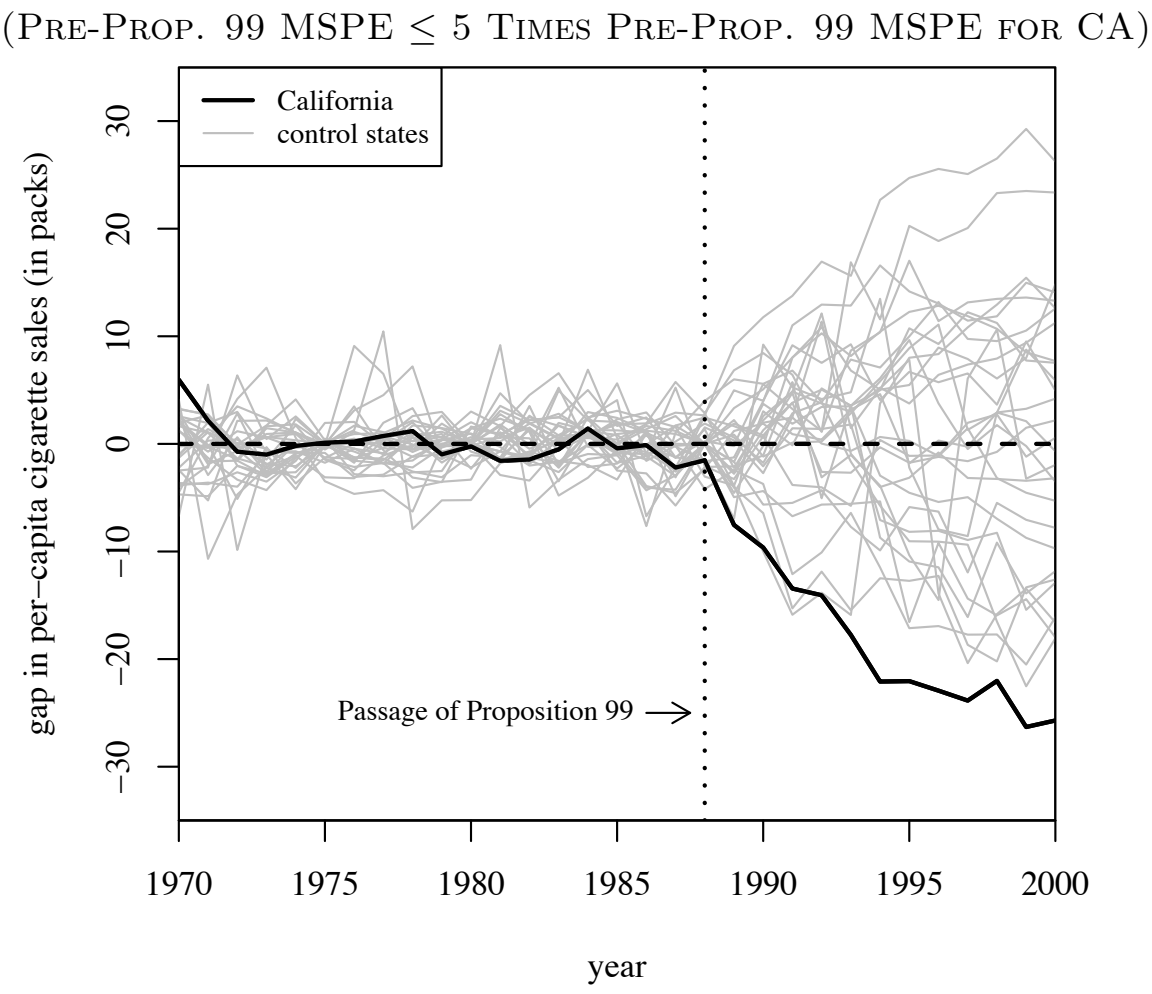
Smoking Gap for CA and 38 control states



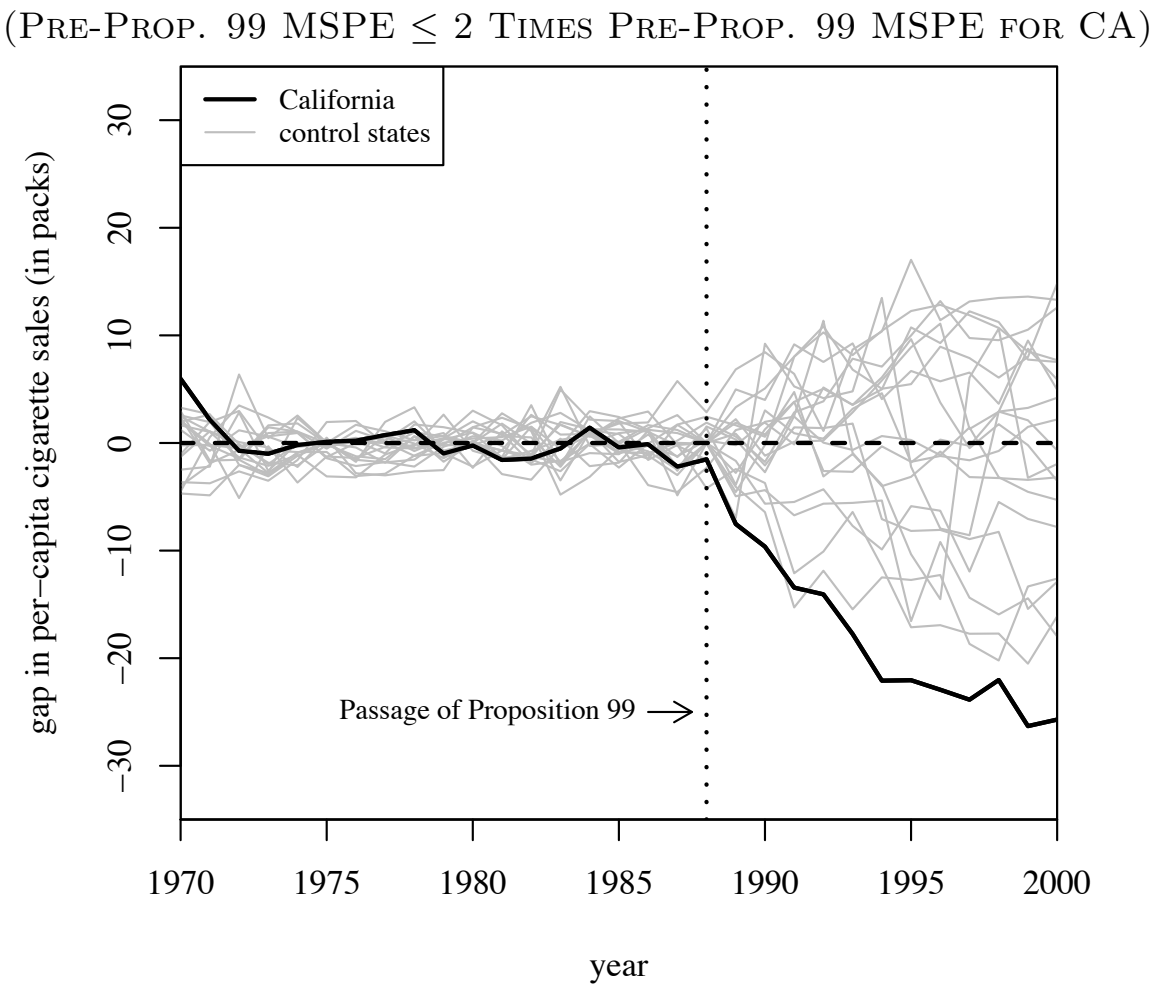
Smoking Gap for CA and 34 control states



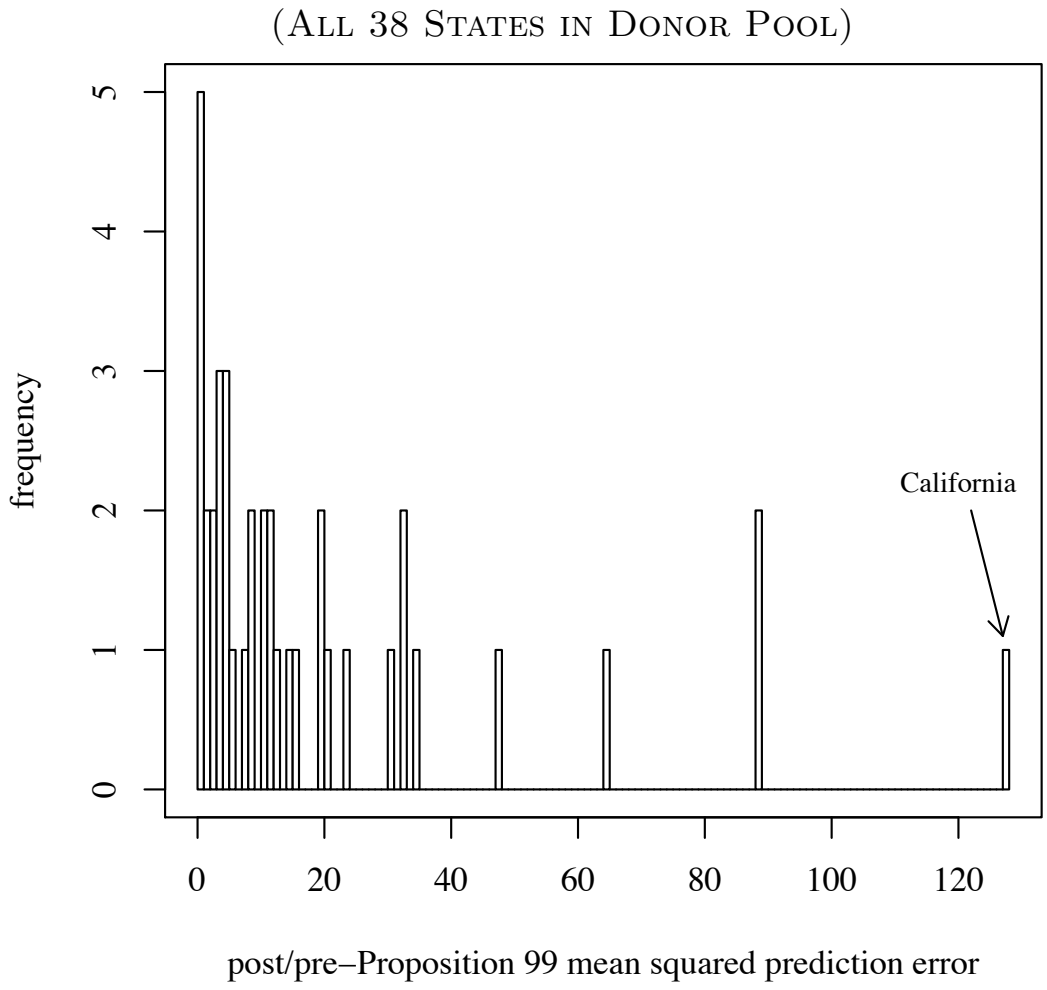
Smoking Gap for CA and 29 control states



Smoking Gap for CA and 19 control states



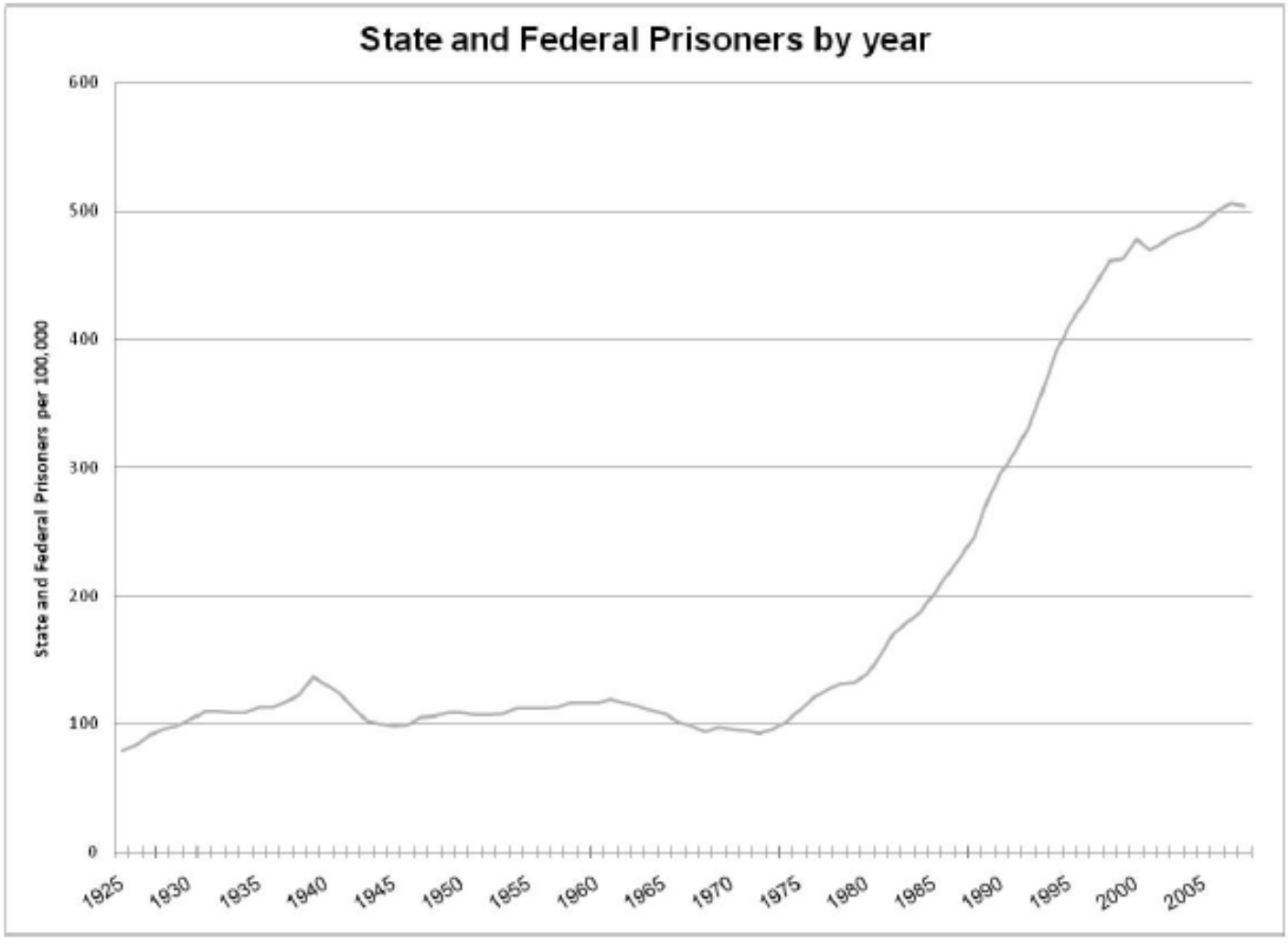
Ratio Post-Prop. 99 RMSPE to Pre-Prop. 99 RMSPE



Coding exercise

- The US has the highest prison population of any OECD country in the world
- 2.3 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



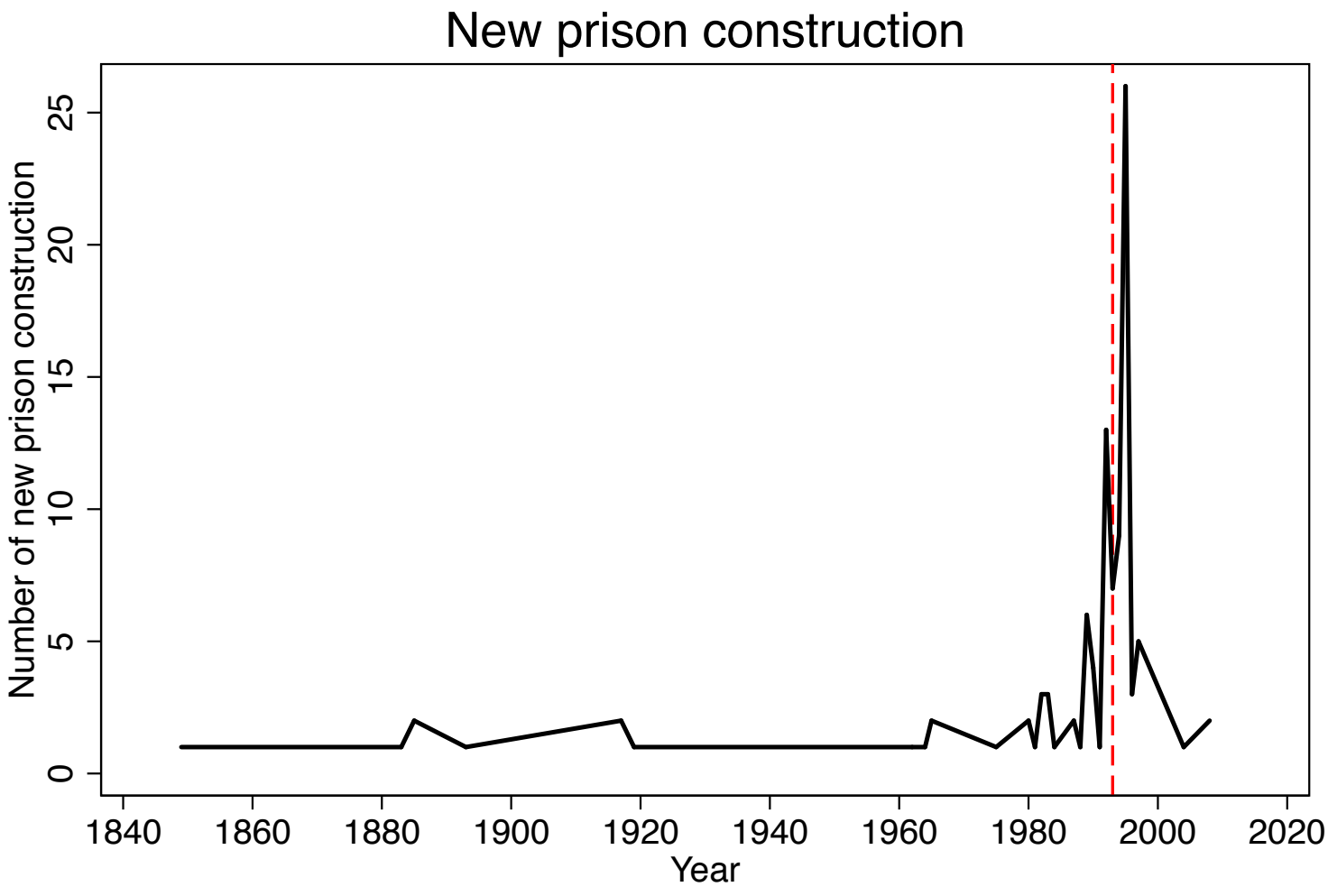
Source: www.albany.edu/sourcebook/tost_6/html

Prison constraints

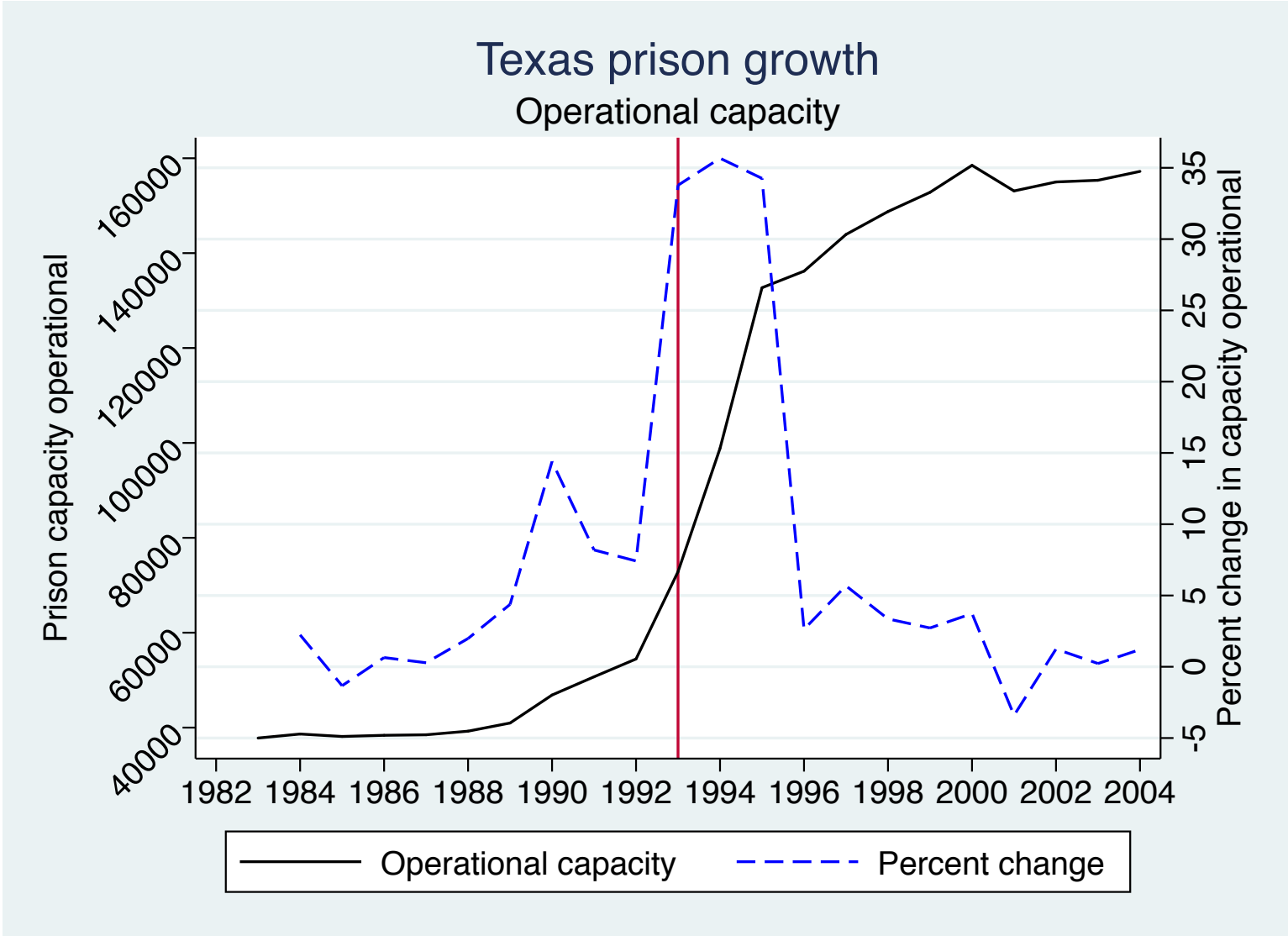
- Prisons are and have been at capacity for a long time.
- Requires managing flows through
  - → Prison construction
  - → Overcrowding
  - → Paroles

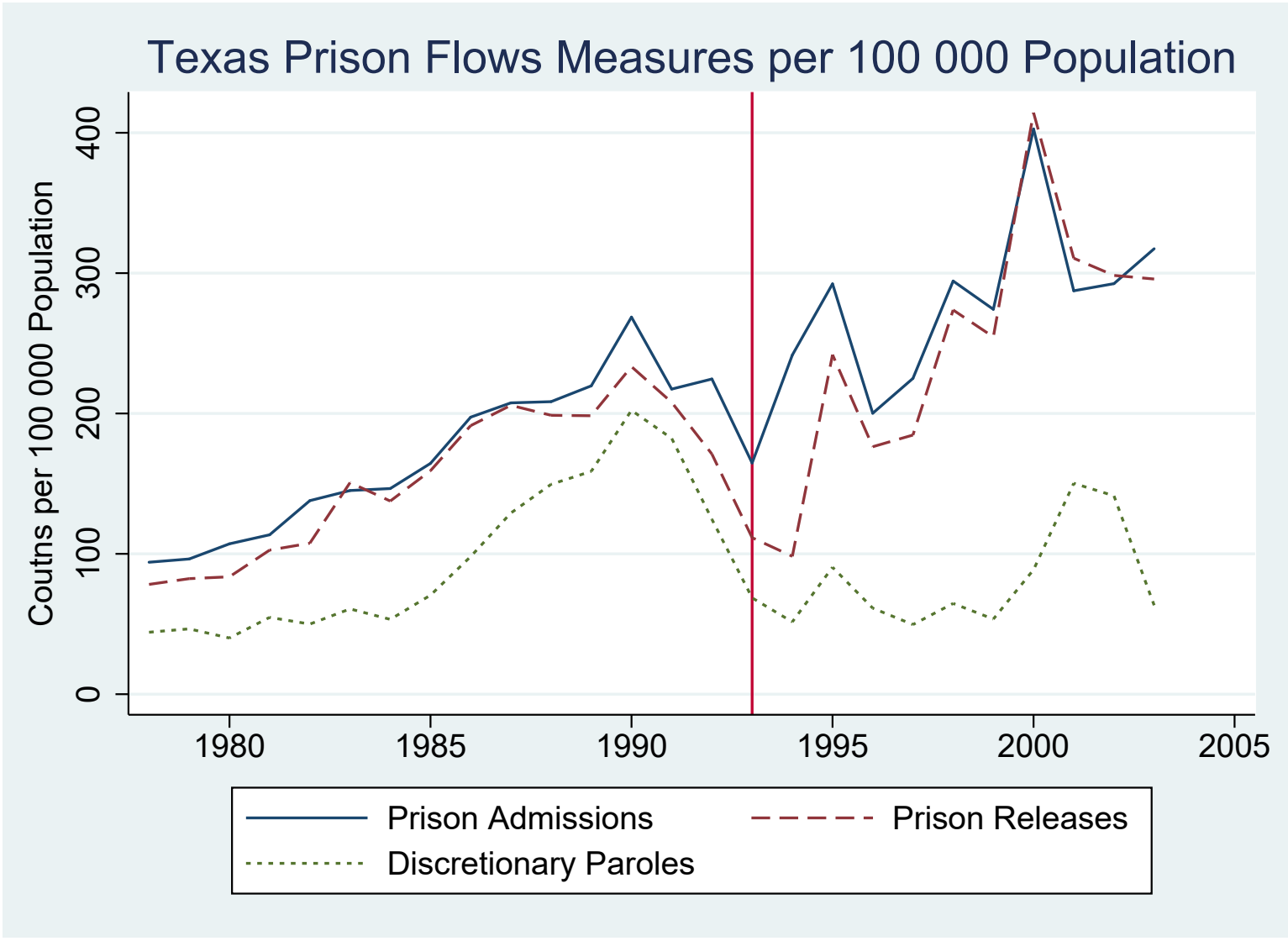
Texas prison boom

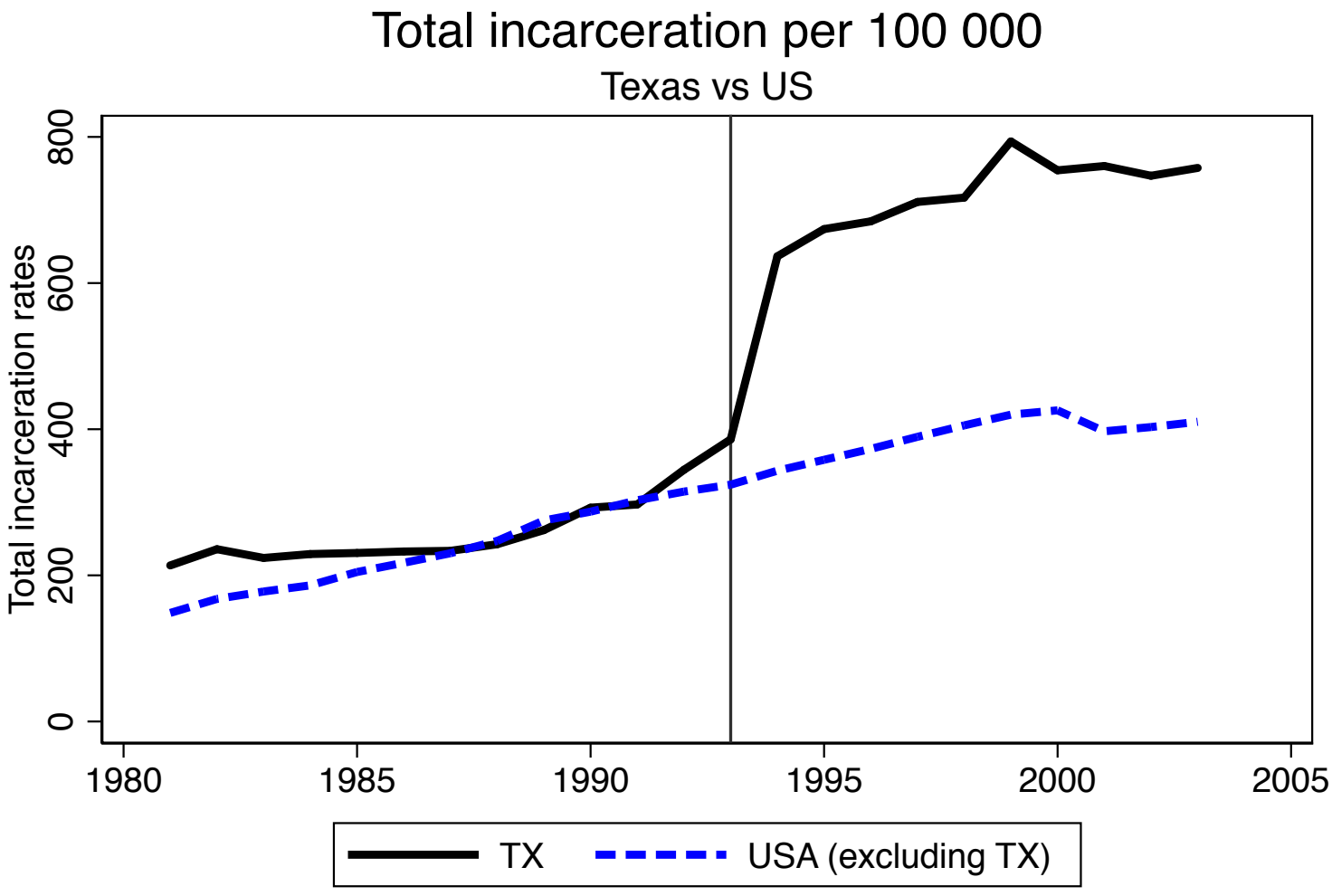
- Ruiz v. Estelle 1980
  - → Class action lawsuit against TX Dept of Corrections (Estelle, warden).
  - → TDC lost. Lengthy period of appeals and legal decrees.
  - → Lengthy period of time relying on paroles to manage flows
- Governor Ann Richards (D) 1991-1995
  - → Operation prison capacity increased 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)

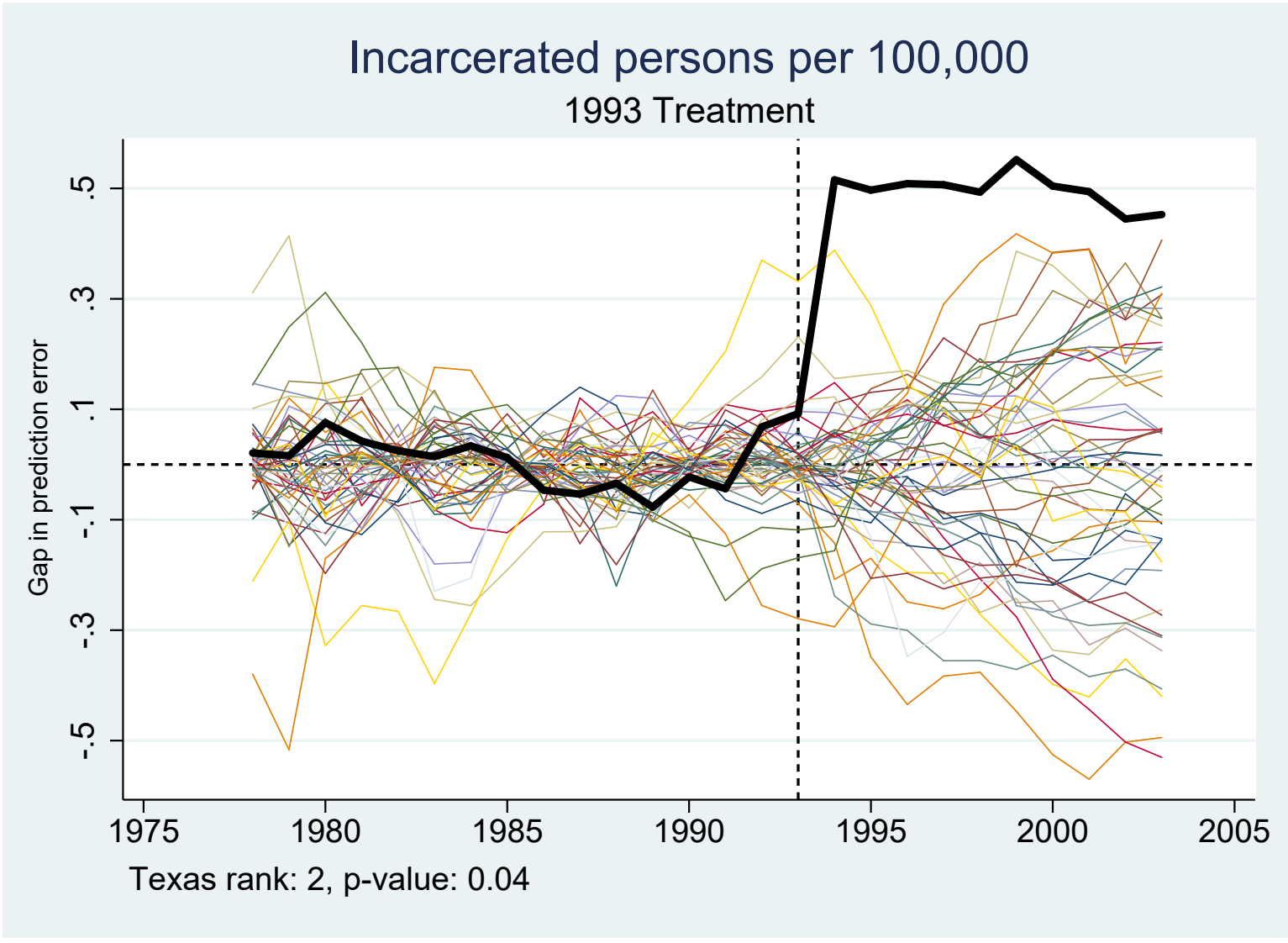


Red dashed line is 1993


Texas prison growth
Operational capacity
Prison capacity operational
Percent change in capacity operational
40000
60000
80000
100000
120000
140000
160000
-5
0
5
10
15
20
25
30
35
1982
1984
1986
1988
1990
1992
1994
1996
1998
2000
2002
2004
Operational capacity
Percent change


Texas Prison Flows Measures per 100 000 Population
Coutns per 100 000 Population
0
100
200
300
400
1980
1985
1990
1995
2000
2005
Prison Admissions
Prison Releases
Discretionary Paroles


Total incarceration per 100 000
Texas vs US
Total incarceration rates
0
200
400
600
800
1980
1985
1990
1995
2000
2005
TX
USA (excluding TX)
1993 starts the prison expansion


Incarcerated persons per 100,000
1993 Treatment
Gap in prediction error
.5
.3
.1
-.1
-.3
-.5
1975
1980
1985
1990
1995
2000
2005
Texas rank: 2, p-value: 0.04

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

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, δ."

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

Overview

- Athey, et al. (2021) unites two literatures – unconfoundedness and synthetic control
- Combines computer science with statistics to create the matrix completion with nuclear norm (MCNN) estimator
- Nuclear norm regularization is used for the imputation

What is matrix completion

- Completing a matrix means guessing at the correct values that are missing
- Hence the “completion” is just another name for “filling in” the matrix
- In causal inference, if the matrix is a matrix of potential outcomes (e.g., Y^0), then missingness is caused by treatment assignment

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, SUTVA, that flips treatment from 0 to 1 in the last period t:

$$Y = DY^0 + (1 - D)Y^1$$

Ask yourself: why are there question marks in the last column?

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

Matrix completion seeks to do the following:

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

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 (100m users ranking 100k movies)
- I like Silver Linings Playbook and Lars and the Real Girl and you like Silver Linings Playbook
- Probably you'll also like Lars and the Real Girl
- So we are using correlations in the columns to "complete" missing values
- When you think about it, while it seems predictive (and it is), isn't it really a causal design?
- "If I watch Lars and the Real Girl, will I like it?"

Types of imputation

- I didn't always think of causal inference in terms of imputation because often the method was just taking existing values and manipulating them, rather than filling in missing values
- But the fundamental problem of causal inference states that causal inference is a missing data problem, so it makes sense you'd be imputing
- I tend to think therefore in terms of implicit and explicit imputation methods
- Borusyak, et al. (2021) and Athey, et al. (2021) both seem more like "explicit" imputation methods
- Callaway and Sant'Anna (2020) on the other hand is an implicit method, as is did methods more generally

Two literatures

- Lots of moving parts in this interesting paper, so my goal here is purely explainer and mostly high level at that.
- I want you to be competent and conversant in it so we also have some R code
- 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

- 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
- Nuclear norm ensures a low rank matrix needed for sensible imputations

The Gist

- 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. Nuclear norm is important here
2. Shows unconfoundedness and synth are in fact matrix completion methods
   - → they all have the same objective function based on the Frobenius norm for the difference between the latent matrix and the observed matrix
   - → Each approach imposes different sets of restrictions on the factors in the matrix factorization
   - → MCNN by contrast doesn't impose any restrictions – just regularization to characterize the estimator
3. Applies the method to two datasets, but I'm going to skip it though for now

Block structure

- Lots of jargon in this article – unconfoundedness, vertical and horizontal regression, fat and thin matrices.
- Unfortunately, you need to learn it all so let me try and organize it
- We define the matrix first in terms of its block structure which is describing where and when the missingness is occurring in the matrix

Unconfoundedness

- Much of the unconfoundedness literature estimates an ATE under unconfoundedness
- But it tends to focus only on a simple setup where the missingness is the last period
- Think about LaLonde (1986) – NSW treats the workers, and then you don't observe Y^0 for the treated group in the *last period*
- This is the "single-treated-period block structure" because only one *period* is missing

Single-treated-period block structure

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

Single-treated-unit block structure

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

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

Staggered adoption

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

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. Their MCNN will be that.

Thin and Fat matrices

- We also have to 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 basically where $N >> T$, but fat matrices are ones where $T >> N$
- Approximately square ones are where T is approximately equal to N

Vertical and horizontal regression

- Two special combinations of missing data patterns and matrix shape need special attention because they are the focus of large but separate literatures
- Unconfoundedness has that single-treated period block structure with a thin matrix ($N >> T$).
- You use a large number of units and impute missing potential outcomes in the last period using controls with similar lagged outcomes
- This is the horizontal regression – imagine just running OLS on the lags and taking predicted values
- The horizontal regression holds under unconfoundedness

Vertical regression

Doudchenko and Imbens (2016) and Pinto and Furman (2019) show that Abadie, Diamond and Hainmueller (2011) can be interpreted as regressing the outcomes for the treated prior to treatment on the outcomes for controls in the same period

Fixed effects and factor models

- Both horizontal and vertical regressions exploit other patterns
- 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
- This is where their matrix completion with nearest neighbor model comes in – it does that very thing

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 1

Apart from the unconfoundedness assumption, we have this weird assumption!

Assumption 1

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

Lots of matrix forms can be defined this way. But let's not get lost in the weeds – we are still just trying to estimate $L*$! That's the main storyline, not the side quest, to use Red Dead Redemption words I understand

All imputations are wrong but some are useful

- You can impute something a million different ways.
- $1 + 1 + 1 + 1 = 4$ is an imputation of the fifth unknown element and frankly just looking at it, seems wrong.
- 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 $\delta = 0$.
- They add a penalty term $||\lambda||$ to the objective function, but even then, not all of them do well.
- Turns out, it actually matters whether you regularize the fixed effects or not (just like it matters whether you regularize the constant in LASSO apparently – I decided to take their word for it)

Estimator

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

where the objective function is:

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

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

Other norms

- One thing I thought was interesting was that the nuclear norm allowed for the construction of a low rank $L*$ matrix, but other norms actually would have weird properties
- I remember once me asking Imbens (like I had even a clue what I was talking about), “Why not use elastic net? Why are you using the nuclear norm?” He said elastic net would spit out all zeroes. I remember thinking “Why did I think I would understand what he told me?”
- 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
- There’s almost like a cross walk, too, between this and Borusyak, et al. (2021) but I don’t quite see it except they both leverage imputation

Conclusion

- Ultimately, this is just another model though that can be used for differential timing but at the moment, no one knows how it performs in simulations alongside Borusyak, et al. (2021), Callaway and Sant'Anna (2020) or any of the others
- So I can't really answer questions about when to use it and not to – it comes down to these very narrow assumptions
- 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: ??

New developments

- We have been interested in the effect of the treatment on the treated (ATT), and have been using panel data largely to do this – first with the DiD, then the synthetic control model
- Athey and Imbens have been very active in applying machine learning methods to causal inference ("most important innovation in causal inference of the last 15 years" - Athey and Imbens)
- Their work on synthetic control has been in this spirit
- The synthetic DiD bears some similarities to their MCNN model, but focuses on estimating weights, not the L^* matrix

When to use this

- Matrix completion with nuclear norm regularization allows for staggered adoption
- Use this setting for 2x2 situations with more than one treatment group
- So the “block” is a binary treatment with units treated in some late period
- It will dominate the Abadie, Diamond and Hainmueller (2010) as they will show and addresses overfitting and other things through estimating oracle weights (which I’ll explain towards the latter half)
- Very technical paper – this is not your mother’s econometrics. We are shifting towards more structural estimation; perhaps a trend
- Although parallel trends was in some ways already structural because it placed restrictions on the parallel trends (as opposed to relying on randomization)

Model selection

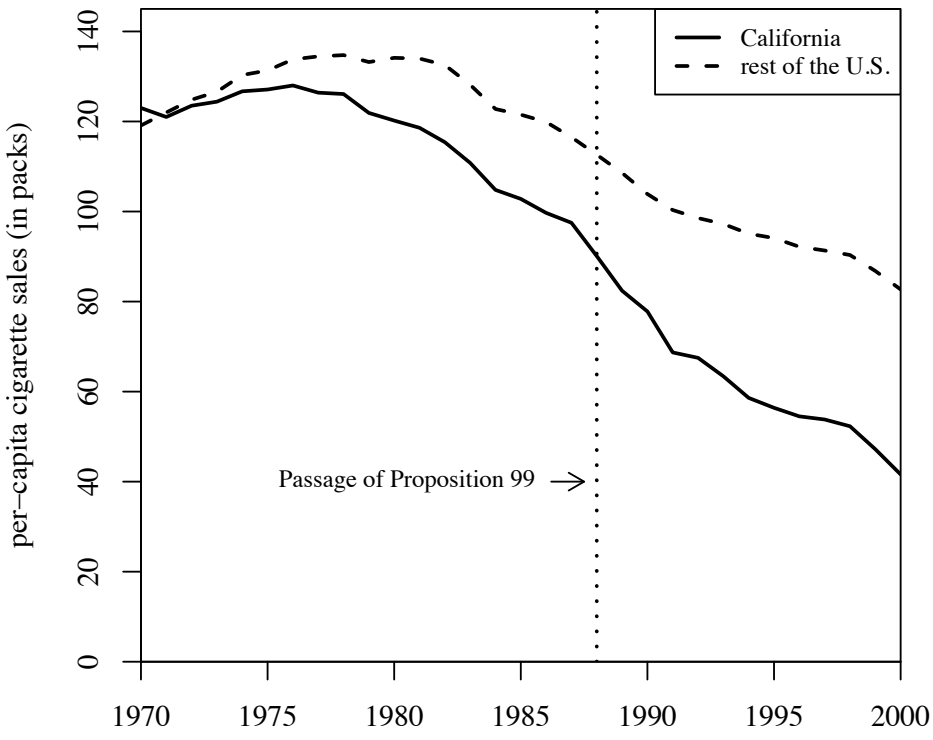
- Another thing is that ADH tends to rely on an “eyeball test” for the pre-trend fitting
- “Researcher degrees of freedom” vs “reducing subjective researcher bias”
- This will allow for a more principled approach

Imperfect fits

- Recall that ADH needs to fit a pre-treatment convex hull to model the heterogeneity
- Often, though, the fit is imperfect for various reason because weights are constrained to be non-negative and sum to one
- But this can be problematic if the treatment group can't be approximated by a weighted average of other units since the weights are fractions
- So they're going to allow for a constant level shift to "get there"

Diff-in-diff, parallel trends and pre-trends

- Recall the identifying assumption in DiD – parallel trends
- Untestable, but we often use pre-trends for an indirect test
- But in the smoking example, parallel trends didn't hold for many states
- Choice of control units matter – the average trends for many control states are roughly parallel, but not all



ADH Synth

- ADH sought a weighted average over the control units to recreate the pre-trend through a fitting exercise
- Synthetic control becomes the weighted average of controls, and then the focus is just on estimating weights
- All we ask is that the weighted average follow the same dynamic path as treatment group (a fit for each period)
- Doudchenko and Imbens (2015) note that this is just a "vertical regression" which yields coefficients on the control units (as opposed to the lags in T)

$$Y_{1,t}^{0} = \sum_{j=2}^{J+1} \widehat{\omega_j} \times Y_{j+1,t}$$

- To the degree the fit is good pre-treatment, then the gaps post-treatment measure ATT at a point in time

Weighting across controls

Assume that the synthetic control at any period is $Y_{1,t} \approx \sum_{j=2}^{J+2} w_i \times Y_j$

- Synthetic control – weights, $\widehat{w}$, control units to get weighted average controls
  1. Use the pre-treatment data to find the optimal weights that when aggregated over control units predict treatment group outcomes ("fit")
  2. Assumes that there's a stable relationship over time, though, because this is going to be our estimated counterfactual post-treatment
- This is shown to be equivalent to a "vertical regression" where you regress units against the higher column units to get those weights
- May require regularization in the regression (if there are more units than time periods)

Weighting across time dimensions

- Forecasting – time weights, $\widehat{\lambda}$, periods to get weighted average periods
    1. Use the controls to learn an average of periods that forecast what we see post-treatment
    2. Imagine a regression, in other words, that yields coefficients on covariates, not on units, to predict future counterfactual
    3. Assumes that this relationship remains valid for the treated and we use the same average of periods to impute the Y^0 for our treatment group
- This is equivalent to a "horizontal regression" where you regress outcomes against the leads (i.e., Y_{it} against $Y_{i,t-1}$) – this is what was meant by unconfoundedness from the MCNN lecture
- Again may need regularization if there are more time periods than units

Difference-in-differences model

- They tend to equate DiD with a TWFE model

$$Y(0)_{it} = \mu + \alpha_i + \gamma_t + \varepsilon_{it}$$

  and solve for the unknown parameters
- More generally, these are the factor models

Reconciling these things

- Vertical regression (i.e., the ADH synth approach) assumes there is a stable relationship between units over time (hence why the weights accurately estimate counterfactuals post-treatment)
- Horizontal regression (i.e., the unconfoundedness approach) is similar, but assumes a stable relationship between outcomes in the treatment period and pre-treatment periods that is the same for all units
- DiD regression (TWFE): assumes an additive outcome model that captures differences between time and units

So the focus becomes about choosing between these methods

Synthetic DiD

Synthetic DID takes synth and forecasting to create a *synthetic DiD* version

- Combine these two – weighting controls using pre-treatment and weighting time using controls, then applying a type of DiD differencing – to create the synthetic DiD model
- There is a focus, just like ADH, on estimating appropriate weights
- It's doubly robust – only one has to remain valid
- Constant effects will get differenced out and the synthetic control can be *parallel* to treatment, as opposed to *identical* in pre-treatment period

Estimation of SDiD

Synthetic DiD is DiD with a synthetic control and a pre-treatment period (on the baseline, just like CS).

1. Compute the regularization parameter to match the size of a typical one-period outcome change, $\Delta_{it} = Y_{i(t+1)} - Y_{it}$, for unexposed

Estimation of SDiD

2. Estimate unit weights $\widehat{w}$ defining a synthetic control unit (just like Abadie, Diamond and Hainmueller 2010) using the pre-treatment data

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

but they allow for an intercept term so that now the weights no longer need to make the unexposed pre-trends *perfectly* match the treatment group (hence convex hull can fail to hold)

Estimation of SDiD

3. Estimate the time weights $\widehat{\lambda}$ defining a synthetic pre-treatment period using control data

$$\widehat{\lambda}_{j=1} + Y_{1,pre}\widehat{\lambda} \approx Y_{1,post}$$

Estimation

4. Computer the SDID estimator via the weighted DID regression

$$\arg\min_{\tau,\mu,\alpha,\beta} = \left\{ \sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{it} - \mu - \alpha_i - \beta_t - W_{it}\tau)^2 \widehat{w}_i^{sdid} \widehat{\lambda}_t^{sdid} \right\}$$

Estimating the weights

Our focus then becomes about estimating $\widehat{w}$ and $\widehat{\lambda}$

5. Estimate the control weights, $\widehat{w}$, defining the control group unit via constrained least squares on the pre-treatment data. This requires weights to be non-negative and sum to one and allows for a level shift with regularization. Synthetic control is a weighted average like in ADH

Estimating the weights

6. We then estimate the time weights. $\widehat{\lambda}$, defining the synthetic pre-treatment period via constrained least squares on the control data with analogous time constraints

More formalization

Assumed data generating process – outcome is “low rank matrix” (MCNN) plus noise

$$Y = L + \tau D + E$$

where L is the systematic component and the conditional expectation of the error matrix E given the assignment matrix D and the systematic component of L is zero.

We won’t estimate L^* though, unlike MCNN

Data generating process – noise and signal

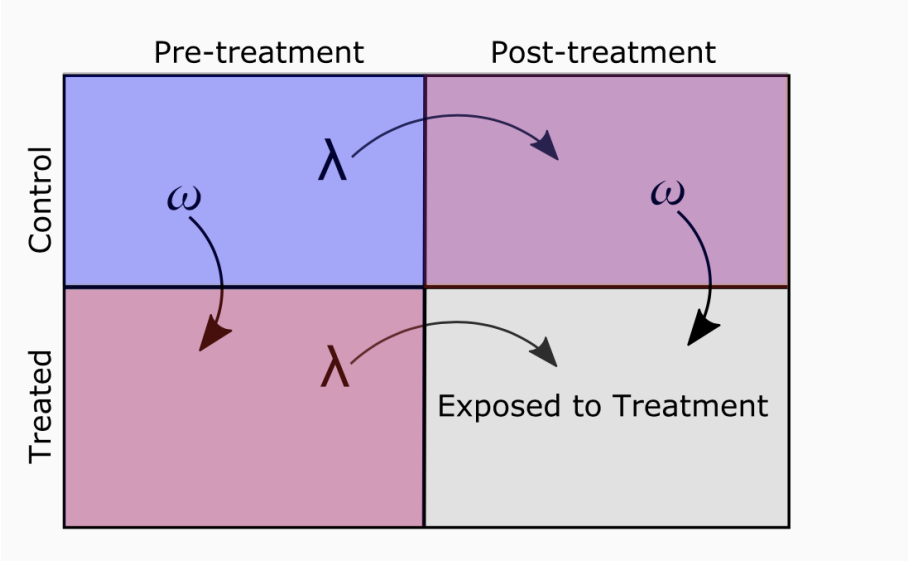
$$Y = L + \tau D + E$$

The treatment cannot depend on the error term, but may depend on the systematic elements of L (i.e., D is not randomized). Think of L as the signal, τ a matrix of treatment effects and E the noise with no autocorrelation over time or between units. The only thing random is E, our noise matrix.

Estimating the weights – high level

- Modify synthetic control weights – use penalized least squares to get a weighted average of control units with pre-trends "parallel" to the treated unit average
- But they'll allow for a constant, unlike ADH synth
- And then they'll do the same thing for the time weights, but this time they won't regularize because they want to weight more intensively the periods "just before" – ridge, they note, would "spread out the weights" over multiple time periods and they don't want that
- I'll get more into this with the oracle weights, but for now I'll just note it conceptually

Picture



(credit: David Hirshberg January 2020 slides because I can't make this picture to save my life)

Regression

- SC is weighted linear regression with no unit FEs:

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

- DiD is unweighted regression with unit FEs and time FEs:

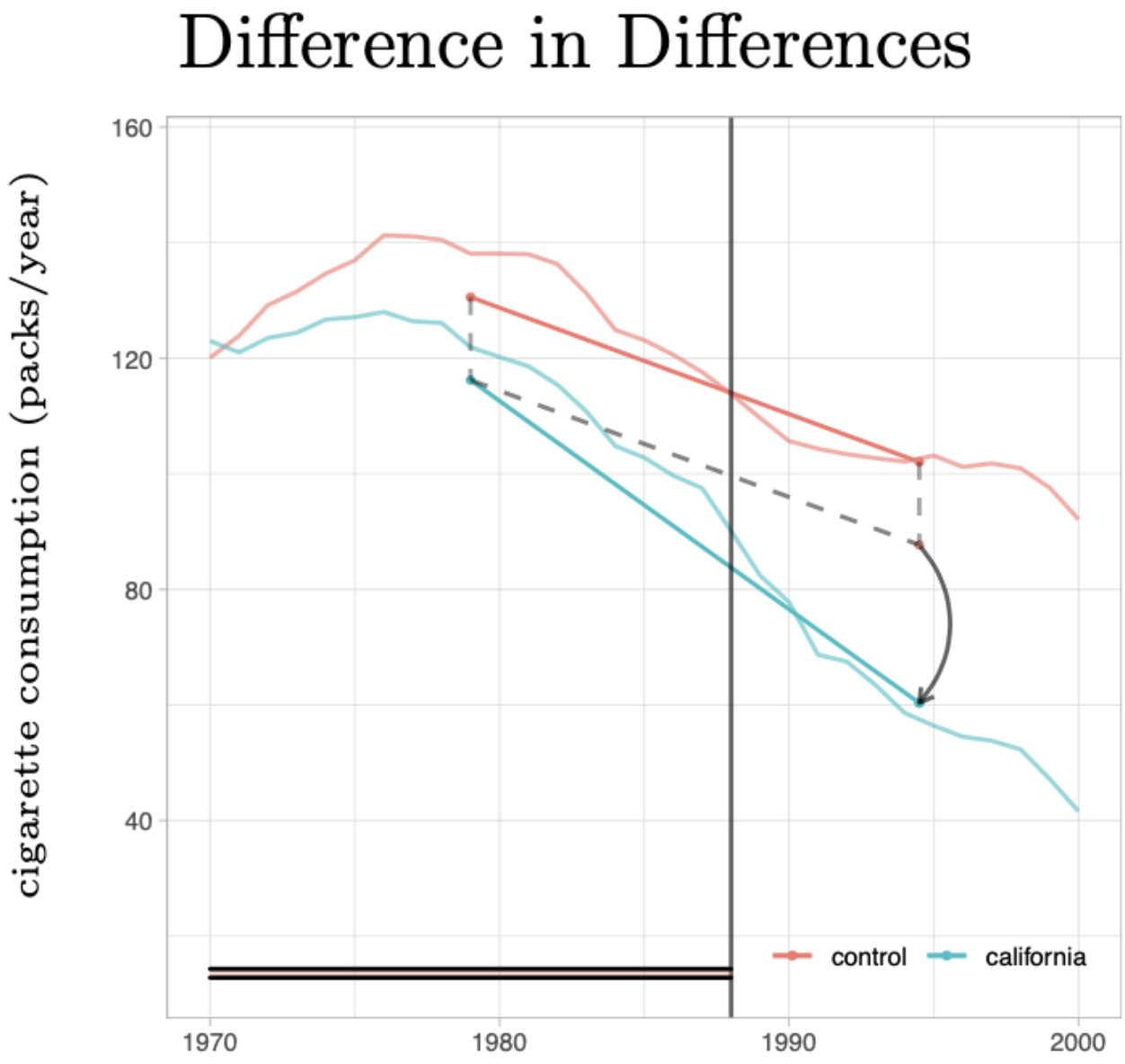
$$\text{argmin}_{\tau,\lambda,\alpha} \sum_{i,t} (Y_{it} - \lambda_t - \alpha_i - \tau D_{it})^2$$

- SDiD is weighted regression with unit FEs and time FEs:

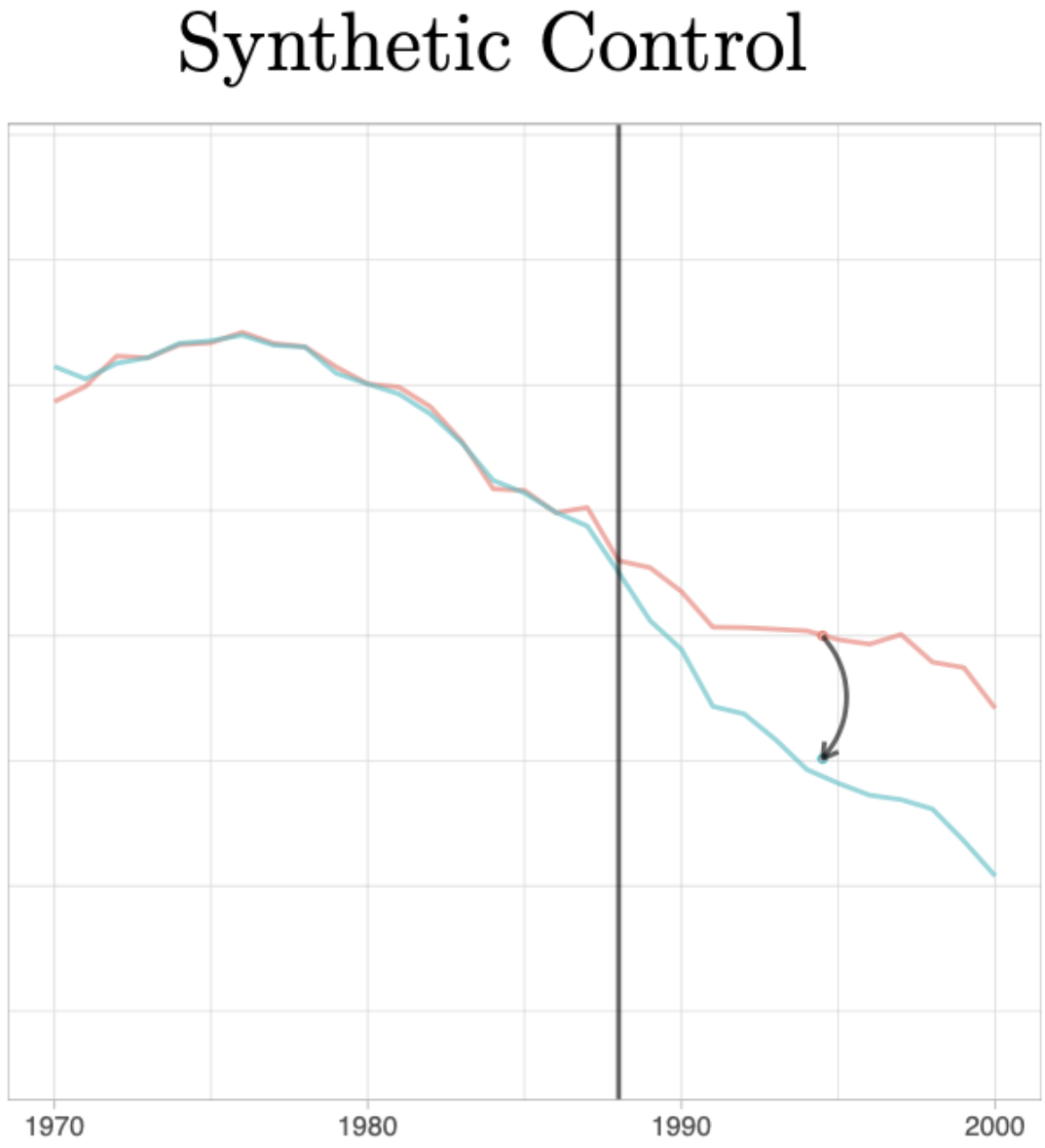
$$\text{argmin}_{\tau,\lambda,\alpha} \sum_{i,t} (Y_{it} - \lambda_t - \alpha_i - \tau D_{it})^2 \times w_i \times \lambda_t$$

Formal results overview

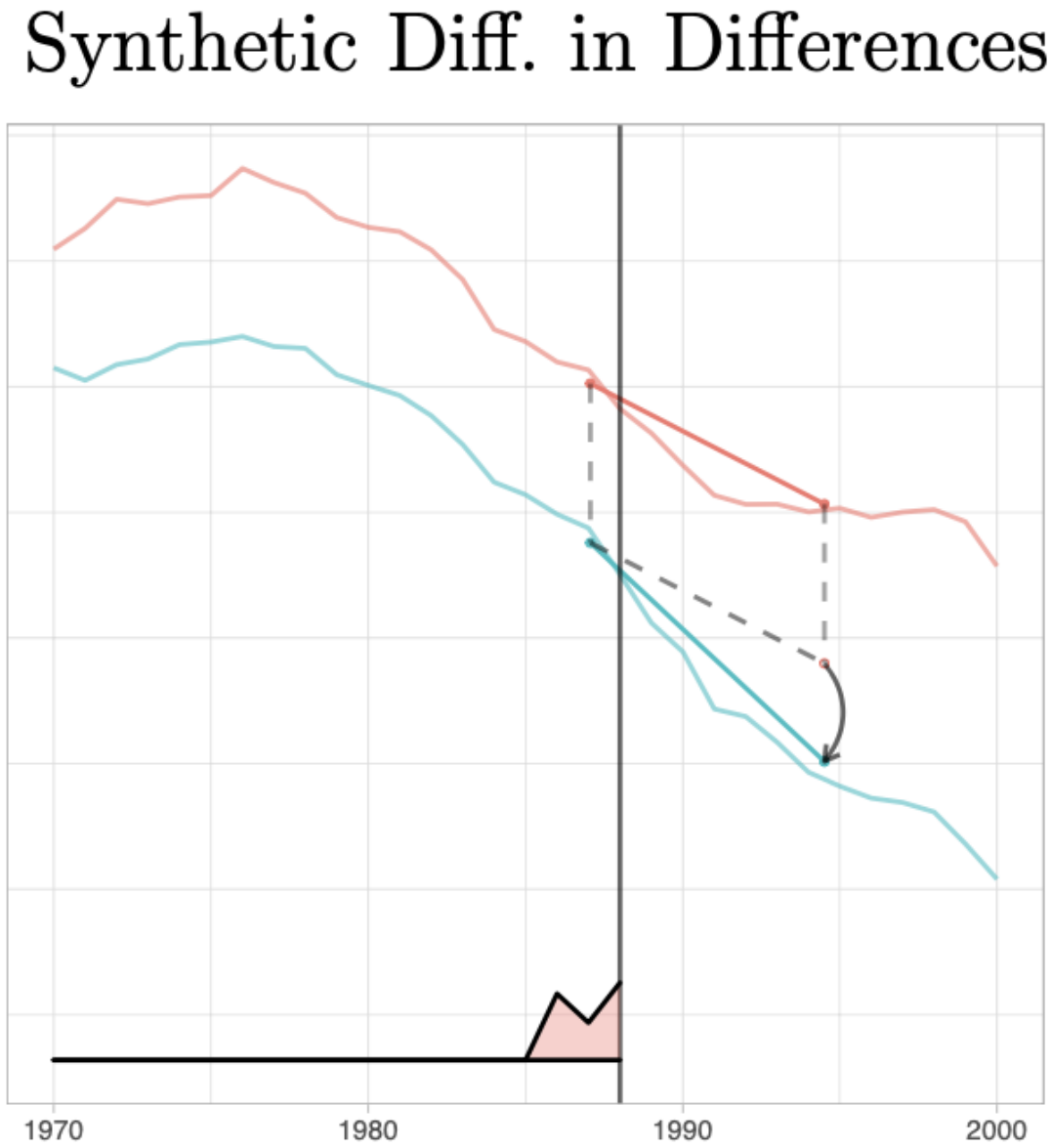
- Formal results will show SDiD is “doubly robust” (recall Sant’Anna and Zhao 2020)
- Factor model on the outcome can be a latent factor model but true model is that signal model and it’ll still be consistent
- Asymptotic normality of $\widehat{\tau}^{SDiD}$
- With oracle weights, SDiD will have “good weights”
- You can do inference through resampling like jackknife, bootstrap and randomization inference



Estimated decrease: -27.3 (17.7)



Estimated decrease: -19.6 (9.9); bad fit just prior bc weights are fitting everywhere



Estimated decrease: -15.4 (8.4). Jagged line left of 1988 is the weighting of those years

Practical problems

- Underfitting. What if I can't get a parallel synthetic control? I know because it's visible. This is an underfitting problem. We need more controls, better controls, or another method.
- Omitted variable bias. Something else happens exactly when the treatment occurs. Sorry – there isn't a solution, because you're not identified.
- Overfitting. We get a synthetic control, but it's because the plot over fit the data. This means that you've not approximated the counterfactual post-treatment. No different than in RDD when you're unable to identify the counterfactual due to functional form problems.

How to rule out overfitting: oracle weights

- Their estimator is equivalent to an “oracle estimator” which cannot overfit
- Oracle uses unit and time weights that don’t depend on the noise
- Weights minimize MSE; oracle weights minimize **expected** SE

Decomposing the bias of SDID

$$
\begin{aligned}
\widehat{\tau}^{sdid} - \tau &= \varepsilon(\widetilde{w}, \widetilde{\lambda}) + B(\widetilde{w}, \widetilde{\lambda}) + \widehat{\tau}(\widehat{w}, \widehat{\lambda}) - \widehat{\tau}(\widetilde{w}, \widetilde{\lambda}) \\
&= \text{oracle noise} + \\
&\quad \text{oracle confounding bias} + \\
&\quad \text{deviation from oracle}
\end{aligned}
$$

So they characterize these terms

Oracle noise

First term: the oracle noise

$$\varepsilon(\widetilde{w}, \widetilde{\lambda})$$

Tends to be small when the weights are small and there are a sufficient number of exposed units and time periods.

Oracle confounding bias (rows / units)

$$B(\widetilde{w}, \widetilde{\lambda})$$

Will be small when the pre-exposure oracle row (units) regression fits well and generalizes to the exposed rows :

$$\widetilde{w_1} + \widetilde{w_j}^T L_{j,pre} \approx \widetilde{w}_1^T L_{1,pre}$$

and

$$\widetilde{w_1} + \widetilde{w_j}^T L_{j,post} \approx \widetilde{w}_1^T L_{1,post}$$

Oracle confounding bias (columns / time)

$$B(\widetilde{w}, \widetilde{\lambda})$$

Will be small when the pre-exposure oracle column (time) regression fits well and generalizes to the exposed columns :

$$\widetilde{\lambda_1} + \widetilde{\lambda_j}^T L_{j,pre} \approx \widetilde{\lambda}_1^T L_{1,pre}$$

, and

$$\widetilde{\lambda_1} + \widetilde{\lambda_j}^T L_{j,post} \approx \widetilde{\lambda}_1^T L_{1,post}$$

Oracle confounding bias – neither do well

What if neither model generalizes well on its own, then there is a doubly robust property

It is sufficient for one model to predict the generalization error of the other

“The upshot is even if one of the sets of weights fails to remove the bias from the presence of L, the combination of oracle unit and time weights can compensate for such failures”

Deviation from Oracle

Core theoretical claim (All formalized in their asymptotic analysis): SDID estimator will be close to the oracle when

- The oracle time and unit weights look promising on their respective training sets

$$\widetilde{w_1} + \widetilde{w_j}^T L_{j,pre} \approx \widetilde{w}_1^T L_{1,pre}$$
$$\widetilde{\lambda_1} + \widetilde{\lambda_j}^T L_{j,pre} \approx \widetilde{\lambda_1}^T L_{1,pre}$$

- and regularization is not too large for either weight

Properties

Under some assumptions, they provide then that SID:

1. SDiD is approximately unbiased and normal
2. SDiD has a variance that is optimal and estimable via clustered bootstrap

Placebo Simulation

- Big picture still – they do a simulation to evaluate bias, RMSE of estimates compared to the observed outcome, but they don't want to use randomization because that may not catch the distinct time trend
- They want the simulation to be "realistic" not "ideal" (i.e., design based identification using randomized treatment dates)
- Bertrand, et al. (2004) randomly assigned a set of states in the CPS to a placebo treatment and the rest the control and examine how well different approaches to inference for DiD covered the true effect of zero
- Only methods that were robust to serial correlation of repeated observations for a given unit (e.g., clustering by level of treatment) attained valid coverage

Treatment assignment process

- Policy: abortion laws, gun laws, minimum wages with outcome hours and unemployment rate
- Logistic regression to predict presence of regulation on four state factors from simulation outcome model M
- Goodness of fit shows that treatment assignment responds strongly to unobserved latent factors
- Assign treatment to states with probabilities from the logistic model

Some details of this placebo simulation

- They calculate average earnings over 40 years and 50 states by subtracting the overall mean and dividing by the standard deviation to get a matrix Y with $||Y||_2^2 = 1$
- They fit a rank 4 factor model M
- They then extract TWFE from there based on unit and time fixed effects F
- Extract low rank matrix as $L = M - F$
- Calculate residuals $E = Y - M$ on an AR(2) model
- Compared SDID, DiD, synthetic control and matrix completion under different baseline scenarios and SDID tends to better

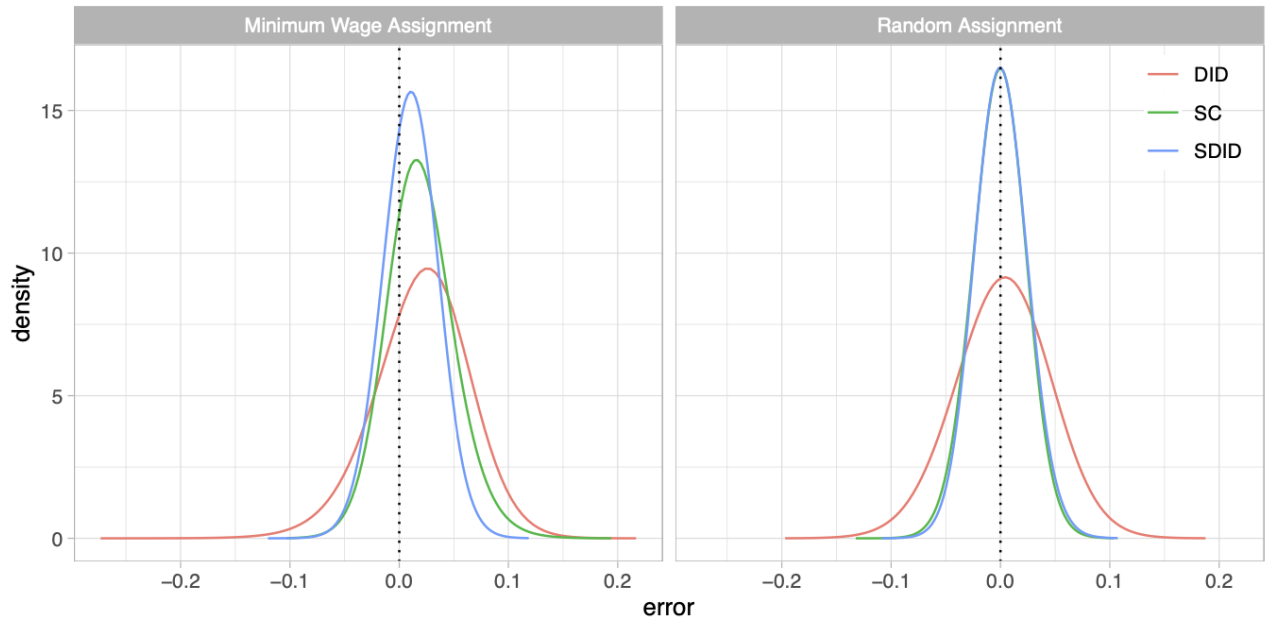


Figure 2: Distribution of the errors of SDID, SC and DID in the setting of the "baseline" (i.e., with minimum wage) and random assignment rows of Table 2.

	RMSE					Bias				
	SDID	SC	DID	MC	DIFP	SDID	SC	DID	MC	DIFP
Baseline	**0.28**	0.37	0.49	0.35	0.32	0.10	0.20	0.21	0.15	0.07
Outcome Model										
No Corr	**0.28**	0.38	0.49	0.35	0.32	0.10	0.20	0.21	0.15	0.07
No $\boldsymbol{M}$	0.16	0.18	**0.14**	**0.14**	0.16	0.01	0.04	0.01	0.01	0.01
No $\boldsymbol{F}$	0.28	**0.23**	0.49	0.35	0.32	0.10	0.04	0.21	0.15	0.07
Only Noise	0.16	**0.14**	**0.14**	**0.14**	0.16	0.01	0.01	0.01	0.01	0.01
No Noise	0.06	0.17	0.47	**0.04**	0.11	0.05	0.04	0.20	0.00	0.01
Assignment Process										
Gun Law	**0.26**	0.27	0.47	0.36	0.30	0.08	-0.03	0.15	0.15	0.09
Abortion	**0.23**	0.31	0.45	0.31	0.27	0.04	0.16	0.03	0.02	0.01
Random)	**0.24**	0.25	0.44	0.31	0.27	0.01	-0.01	0.02	0.01	-0.00
Outcome Variable										
Hours	1.90	2.03	2.06	**1.85**	1.97	1.12	-0.49	0.85	1.00	1.00
U-rate	**2.25**	2.31	3.91	2.96	2.30	1.77	1.73	3.60	2.63	1.69
Assignment Block Size										
$T_{\text{post}} = 1$	**0.50**	0.59	0.70	0.51	0.54	0.20	0.17	0.38	0.21	0.12
$N_{\text{tr}} = 1$	**0.63**	0.73	1.26	0.81	0.83	0.03	0.15	0.11	0.05	-0.02
$T_{\text{post}} = N_{\text{tr}} = 1$	1.12	1.24	1.52	**1.07**	1.16	0.14	0.24	0.33	0.16	0.11

Table 2: Simulation Results for CPS Data. The baseline case uses state minimum wage laws to simulate treatment assignment, and generates outcomes using the full data-generating process described in Section II.1.1, with $T_{\text{post}} = 10$ post-treatment periods and at most $N_{\text{tr}} = 10$ treatment states. In subsequent settings, we omit parts of the data-generating process (rows 2-6), consider different distributions for the treatment exposure variable D_i (rows 7-9), different distributions for the outcome variable (rows 10-11), and vary the number of treated cells (rows 12-14). The full dataset has $N = 50$, $T = 40$, and outcomes are normalized to have mean zero and unit variance. All results are based on 1000 simulation replications and are multiplied by 10 for readability.

Inference

This can be used to motivate practical methods for large-sample for inference. You can use conventional confidence intervals to conduct asymptotically valid inference, and they discuss three ways: jackknife, bootstrap, and placebo variance estimation.

	Bootstrap			Jackknife			Placebo		
	SDID	SC	DID	SDID	SC	DID	SDID	SC	DID
Baseline	0.96	0.93	0.89	0.93	—	0.92	0.95	0.88	0.96
Gun Law	0.97	0.96	0.92	0.94	—	0.93	0.94	0.95	0.93
Abortion	0.96	0.94	0.93	0.93	—	0.95	0.97	0.91	0.96
Random	0.96	0.96	0.92	0.93	—	0.94	0.96	0.96	0.94
Hours	0.92	0.96	0.94	0.89	—	0.95	0.91	0.90	0.96
Urate	0.78	0.74	0.38	0.71	—	0.42	0.74	0.77	0.41
$T_{\text{post}} = 1$	0.93	0.94	0.84	0.92	—	0.88	0.92	0.90	0.92
$N_{\text{tr}} = 1$	—	—	—	—	—	—	0.97	0.95	0.96
$T_{\text{post}} = N_{\text{tr}} = 1$	—	—	—	—	—	—	0.96	0.94	0.94
Resample, $N = 200$	0.94	0.96	0.92	0.95	—	0.93	0.96	0.95	0.94
Resample, $N = 400$	0.95	0.91	0.96	0.96	—	0.95	0.96	0.90	0.96
Democracy	0.93	0.96	0.55	0.94	—	0.59	0.98	0.97	0.79
Education	0.95	0.95	0.30	0.95	—	0.34	0.99	0.90	0.94
Random	0.93	0.95	0.89	0.96	—	0.91	0.95	0.94	0.91

Table 4: Coverage results for nominal 95% confidence intervals in the CPS and Penn World Table simulation setting from Tables 2 and 3. The first three columns show coverage of confidence intervals obtained via the clustered bootstrap. The second set of columns show coverage from the jackknife method. The last set of columns show coverage from the placebo method. Unless otherwise specified, all settings have $N = 50$ and $T = 40$ cells, of which at most $N_{\text{tr}} = 10$ units and $T_{\text{post}} = 10$ periods are treated. In rows 7-9, we reduce the number of treated cells. In rows 10 and 11, we artificially make the panel larger by adding rows, which makes the assumption that the number of treated units is small relative to the number of control units more accurate (we set N_{tr} to 10% of the total number of units). We do not report jackknife and bootstrap coverage rates for $N_{\text{tr}} = 1$ because the estimators are not well-defined. We do not report jackknife coverage rates for SC because, as discussed in the text, the variance estimator is not well justified in this case. All results are based on 400 simulation replications.

Some practical considerations

More treated units is worse – when we add treated units, the oracle standard deviation decreases faster leaving too little room for other sources of error to disappear in the noise

More practical considerations

Circumstances are ideal if the signal matrix L admits a good oracle synthetic control and synthetic pre-treatment period and it's too complex

- What is good? Oracle control weights distribute mass over enough control units
- Oracle time weights should distribute the rest of its mass over enough time periods

More practical considerations

Interestingly, this is an overlap assumption (like common support in matching and CS DiD):

- Many control units are like the treated ones
- Many pre-treatment periods are comparable to post-treatment ones

More practical considerations

What is “not too complex” signal matrix L? It’s one that looks different from the matrix of noise

- More about the rank of the matrix – it must be moderate rank
- Moderate means smaller than the square root of the number of control units
- A state’s behavior isn’t idiosyncratic, but characterized by a blend of industries, etc. of relatively few trends

More practical considerations

- Including more controls won't hurt you bc the set of weights is small and the error is insensitive to dimension
- Less than ideal circumstances can be problematic. The error gets worse:
  - → Signal is too complex
  - → Fit and dispersion of the oracle weights is poor

Some comments

- Conceptually, this is ADH synth combined with a simple 2x2 DiD where the weights are based on estimated time and control group weights
- Oracle weights will make improvements that don't suffer from some of the practical problems, like overfitting, that we said
- Synth DiD dominates synthetic control
- Still remains to be seen how we are going to go about choosing between these, but some things we may need to put down (ADH)

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

"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)*

What is augmented synthetic control?

- Recall the quote from earlier by Athey and Imbens – “synthetic control is the most important innovation in causal inference of the last 15 years”
- They haven’t been the only ones working on this – Eli Ben-Michael, Avi Feller and Jesse Rothstein have two new papers on the subject as well
- This model will “augment” the original synthetic control model by Abadie, Diamond and Hainmueller (2010) by adjusting for pre-treatment imbalance
- As with the Athey, et al. papers, they will augment using more contemporary machine learning methods – **penalized ridge regression**

The gist of their argument

1. ADH (“synth”) needs perfect fit and so is biased in practical settings due to the curse of dimensionality as it won’t be the case we get weights constrained to be “on the simplex”
2. Their augmentation will introduce an outcome model to estimate the bias caused by covariate imbalance
3. Introduces ridge regularization linear regression to estimate new weights to reweight synth
4. Think of it as “bias reduction” like Abadie and Imbens (2011) plus it will have doubly robust properties and be equivalent to inverse probability weighting
5. When synth is imbalanced, augmented synth will reduce bias reweighting and bias correction, and when synth is balanced, they are the same

Some topical observations

- Foregoes estimating *donor pool unit weights* (e.g., ADH, synth did, MCNN)
- Synth sequels are using penalization/regularization for estimation
- Relaxes some of the original ADH constraints, like non-negative weights (i.e., no extrapolation)
  - → This is used to address bias caused by imbalance
  - → Negative weights puts them back in the convex hull which recall we need
  - → 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

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

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:

$$\min_w = ||V_X^{1/2}(X_1 - X_0'w)||_2^2 + \psi \sum_{D_j=0} f(w_j)$$

$$\text{s.t. } \sum_{j=2}^{N} w_j = 1 \text{ and } w_j \geq 0$$

$Y_0'w*$ (i.e., optimally weighted donor pool) is the unit 1 “synthetic control”

Predicting counterfactuals

Synth minimizes the following norm:

$$\min_w = ||V_X^{1/2}(X_1 - X_0'w)||_2^2 + \psi \sum_{D_j=0} f(w_j)$$

$$\text{s.t. } \sum_{j=2}^{N} w_j = 1 \text{ and } w_j \geq 0$$

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:

$$\min_w = ||V_X^{1/2}(X_1 - X_0'w)||_2^2 + \psi \sum_{D_j=0} f(w_j)$$

$$\text{s.t. } \sum_{j=2}^{N} w_j = 1 \text{ and } w_j \geq 0$$

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:

$$\min_w = ||V_X^{1/2}(X_1 - X_0'w)||_2^2 + \psi \sum_{D_j=0} f(w_j)$$

$$\text{s.t.} \sum_{j=2}^{N} w_j = 1 \text{ and } w_j \geq 0$$

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 constrain.”

Convex hull

Synth minimizes the following norm:

$$\min_w = ||V_X^{1/2}(X_1 - X_0'w)||_2^2 + \psi \sum_{D_j=0} f(w_j)$$

$$\text{s.t.} \sum_{j=2}^{N} w_j = 1 \text{ and } w_j \geq 0$$

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 (“simplex”)

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

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

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

Perfect fit models heterogeneity

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

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}, follows a factor model where $m(.)$ are pre-treatment outcomes:

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

- Since $\widehat{m(.)}$ estimates the post-treatment outcome, let's view it as estimated bias, analogous to bias correction for inexact matching (Abadie and Imbens 2011)

Bias correction

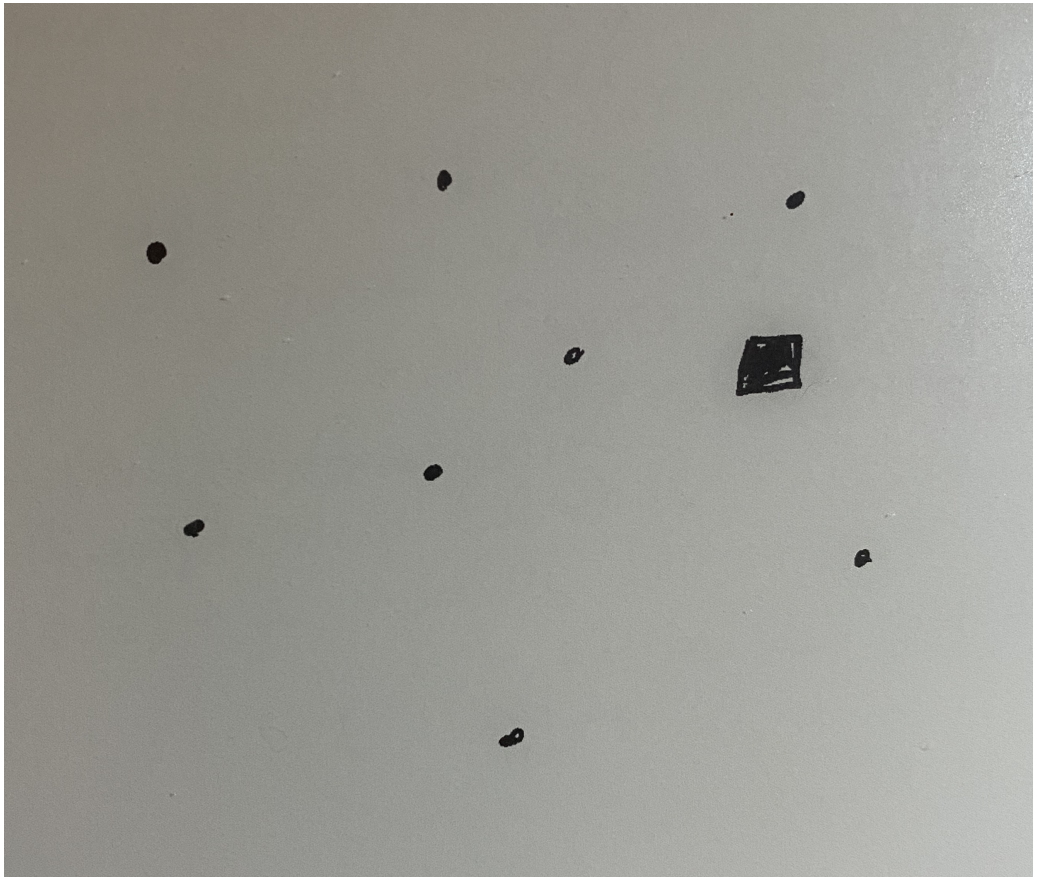
$$Y_{jt}^{0} = m_{jt} + \varepsilon_{jt}$$

- When the weights achieve exact balance, the bias of synthetic control decreases with T
- The intuition is that for a large T (T not transitory shocks), you achieve balance by balancing the latent parameter on the unobserved heterogeneity in our factor model

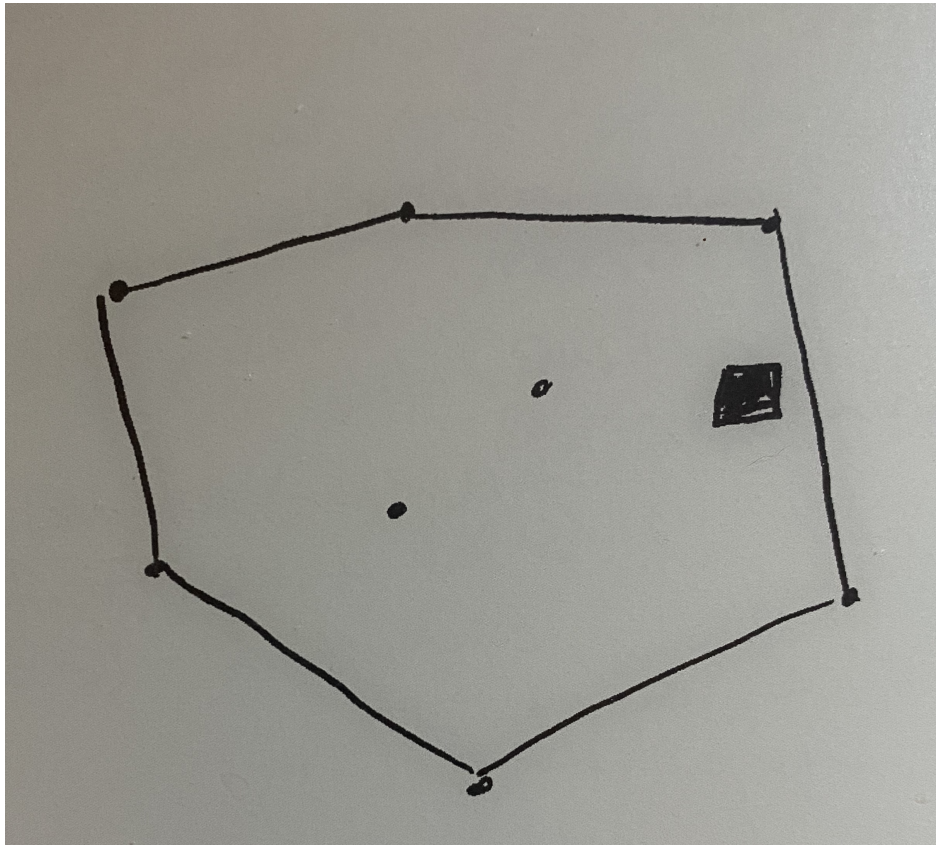
Common practice

- 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

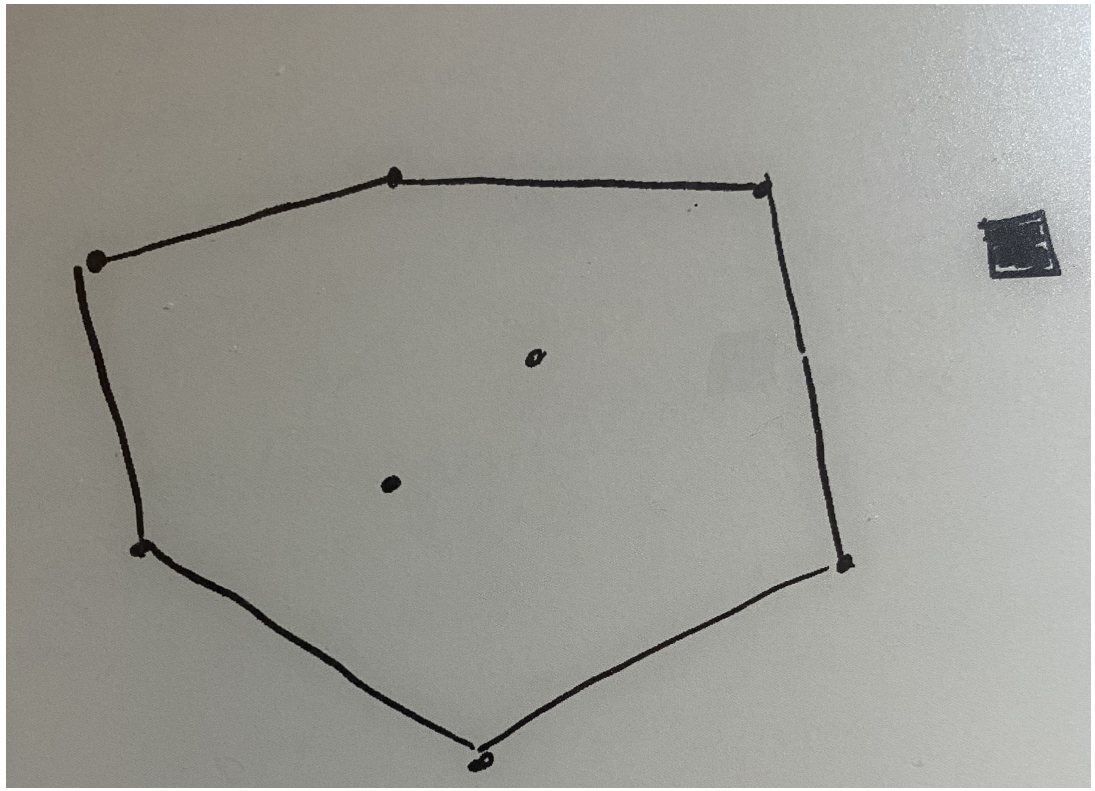
Treatment and control units



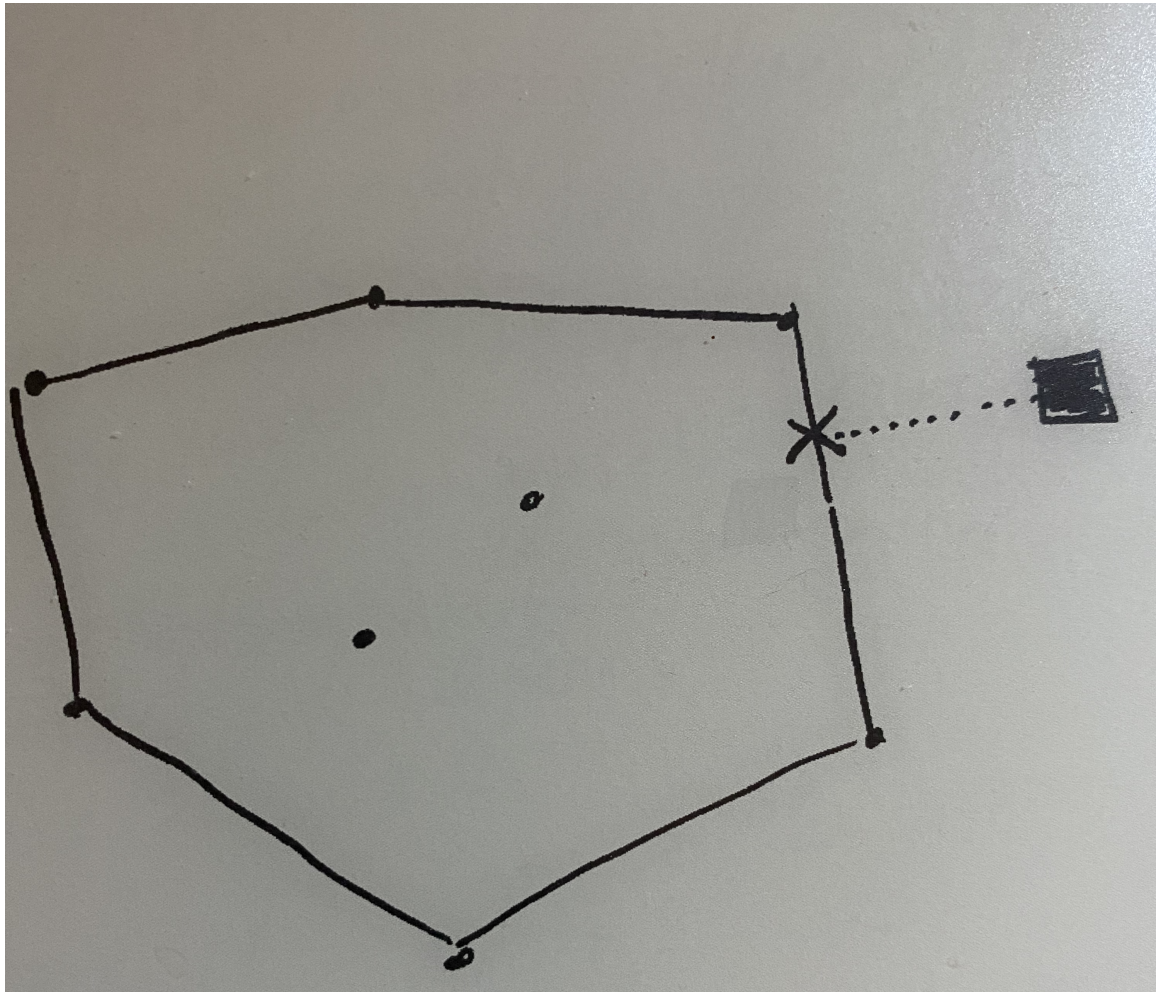
Convex hull – ideal for synth



Outside the convex hull bc of dimensionality



Outside the convex hull bc of dimensionality



Estimating the bias

- Adjust the synthetic control approach to adjust for poor fit pre-treatment.
- Recall our factor model – let $\widehat{m}_{jT}$ be an estimator for the post-treatment control potential outcome Y_{jT}^0.
- The augmented synthetic control estimator for Y_{jt}^0 is on the next slide

Setup of the estimator

Let's adjust synthetic control for this bias. First we'll apply the **bias correction**. Then we'll do the doubly robust augmented **inverse probability weighting**. Let $Y_1^{aug,0}$ be the augmented potential outcome

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

Interpreting line 1

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

(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 $\widehat{m}$.

Interpreting line 1

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

(1) Since $\widehat{m}$ 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

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

(1) I actually cover the bias correction of Abadie and Imbens 2011 in the mixtape! The subclassification chapter

Interpreting line 1

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

(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

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

(2) The second equation is equivalent to a double robust estimation which begins with an outcome model but then re-weights it to balance residuals.

Interpreting line 2

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

(2) The second equation has a connection to inverse probability weighting (they show this in an appendix)

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

Go back to that weighting but use the ridge parameters

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

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

The ridge weights are key to the augmentation

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

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

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

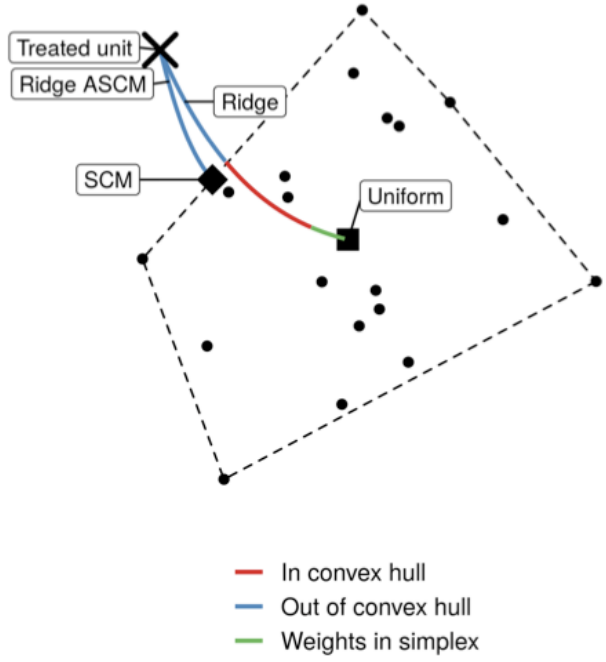
Relaxing the constraint from synth that weights be non-negative, as non-negative weights prohibit extrapolation. But we don't have synthetic control on the simplex, so we *must* extrapolate, otherwise synth will be biased.

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 $\widehat{\lambda}^{ridge}$
- When synth has good pre-treatment fit or when λ^{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!



(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 α for δ by inverting the hypothesis test (see Appendix A for the exact formula)
   - $\rightarrow$ 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

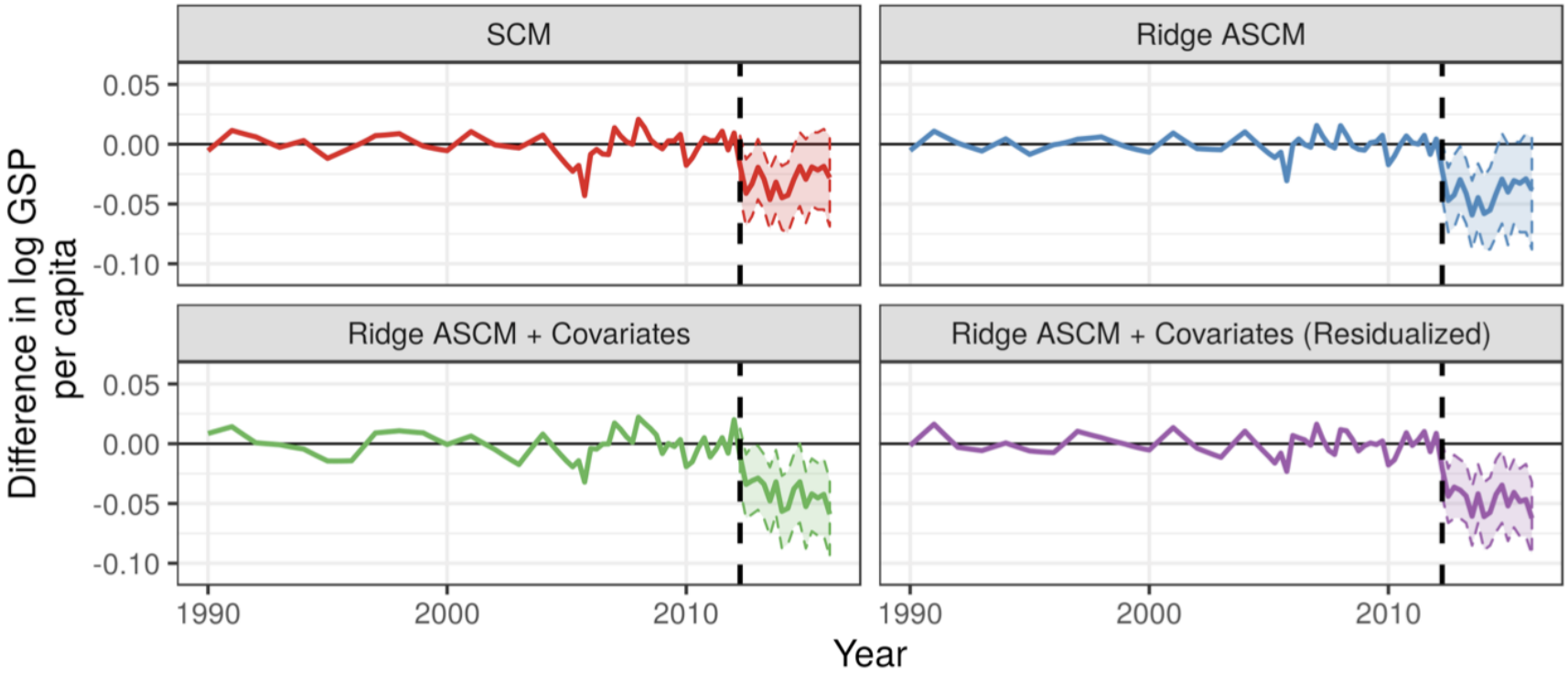
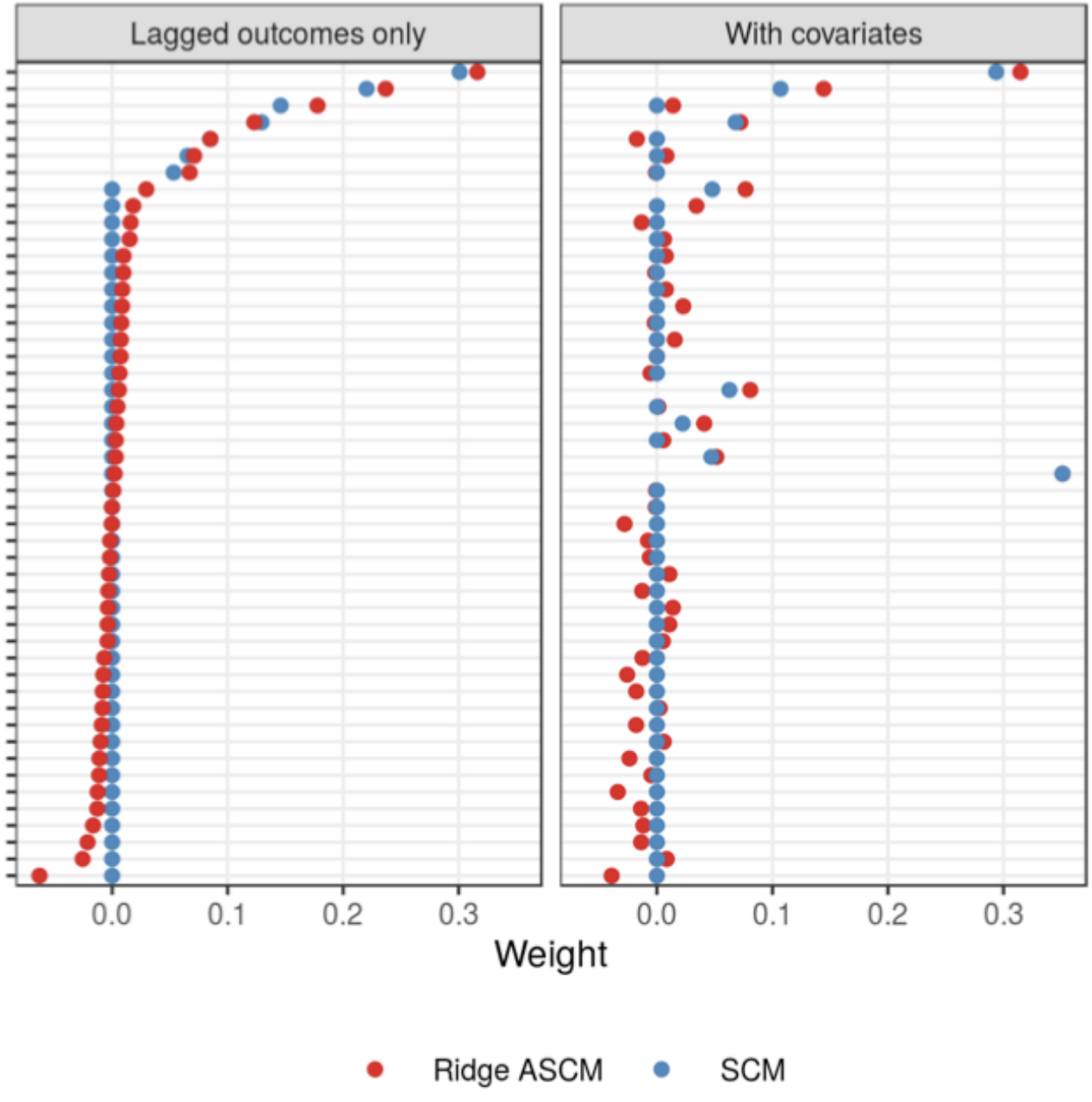


Figure 6: Point estimates along with point-wise 95% conformal confidence intervals for the effect of the tax cuts on log GSP per capita using SCM, Ridge ASCM, and Ridge ASCM with covariates.


Lagged outcomes only
With covariates
0.0
0.1
0.2
0.3
Weight
Ridge ASCM
SCM

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)

Some minor points

- We've motivated augmented synth as a kind of bias correction, but you can also think of it as correcting synth with an inverse probability weight (Appendix E)
- There's an implicit estimate of a propensity score model with ridge regularization
- Weights are odds of treatment (they're ATT weights), i.e., they're the inverse probability weighting scheme from Abadie (2005)

Augmented synth is better

- 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
- They also say synth DiD is a special case of their augmented synth method, which is interesting as synth DiD is also meant to nest all such modifications too (but they don't discuss augmented synth)

R code

R: https://github.com/ebenmichael/augsynth

- ADH was 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 regularization
- Ben-Michael, Feller and Rothstein (2021b) focus on the single unit by allowing differential timing
  - → Augmented synth is a double-robust style (or bias corrected) estimator
  - → This synth is similar to “shrunken”/empirical Bayes/random effects estimation
- It sort of fits with the newer differential timing papers, like matrix completion with nuclear norm regularization had, even though neither are technically DiD
- More machine learning regularization as we’ve been seeing

Paper's Contribution

1. Extend synth to staggered adoption (as opposed to one unit)
2. Show results using an example of unions on spending
3. Propensity score weighting with shrinkage

Motivation

Synthetic control is not a very good propensity score estimator

- Uses pre-treatment outcomes as covariates (some use other X covariates)
- Small N and large K means propensity score must be regularized with probability approaching 1 so perfect balance is not achieved
- You can borrow information to reduce bias from imbalance – for instance from an outcome model (e.g., a regression) and/or other treated units
- You can then combine to produced a weighted event study estimator

Stepping back

- A unit is treated at some period t and we want to know that event's effect on Y
- Standard approach is DiD and event studies
  - → Tons of papers recently (e.g., Goodman-Bacon 2021, Callaway and Sant'Anna 2020)
  - → But what units all DiD papers is *parallel trends*
- That assumption may be wrong, in which case the models' findings will be wrong

Synth with staggered adoption

- Synth was designed originally for the comparative case study – i.e., one treatment group
- What do we do when there's more than one treatment group?
- People in the past tried different things
  - → If they were all treated at the same time, they'd average the treated units and construct a synthetic control for the average
  - → If it was staggered, then they'd fit synth for each treated unit separately, then average those estimates
- They're going to propose optimizing a weighted average of the *global balance* (for the average treated unit) and the sum of *unit-specific balance* for each treated unit

Intuition

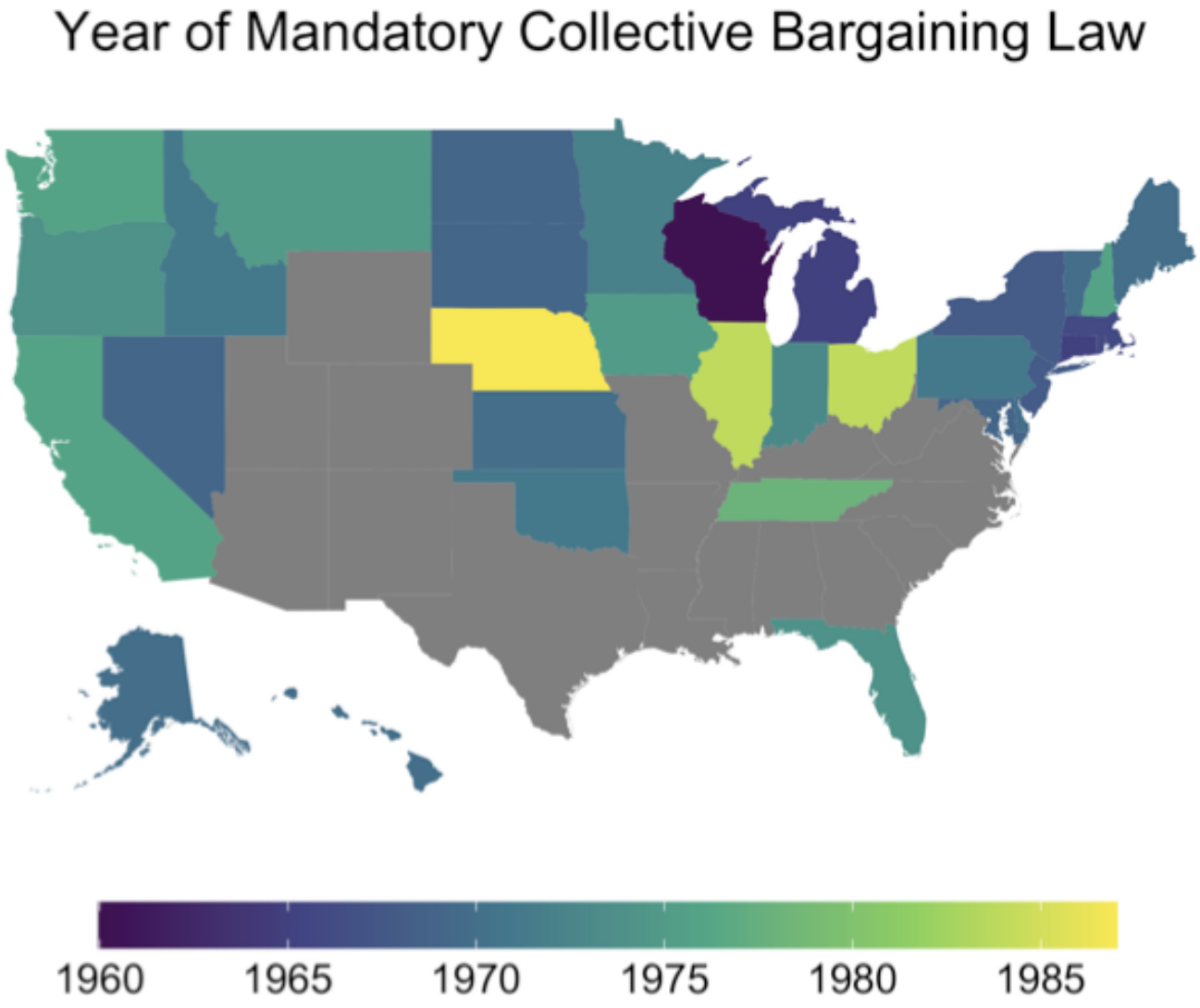
We want to balance the average of the underlying factor loadings

- 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

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
- They're going to re-analyze using all states and synth models


Year of Mandatory Collective Bargaining Law
1960
1965
1970
1975
1980
1985

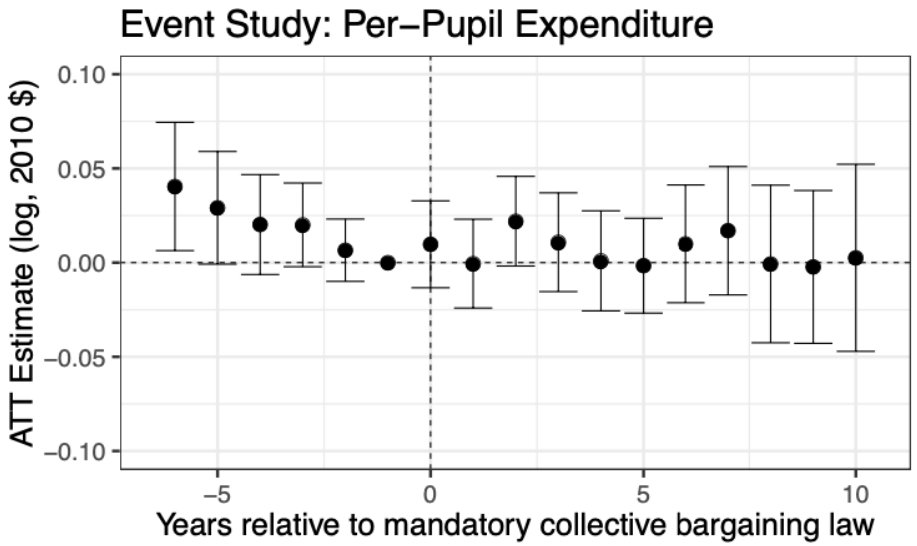
Their paper

1. **Methods**: Extend synth to staggered adoption
2. **Substance**: Application will find minimal effect of unions on spending
3. **Connection**: Propensity score weighting with shrinkage

Standard approach in economics is event study

$$Y_{it} = \text{unit}_i + \text{time}_t + \sum_{\ell=2}^{L} \delta_\ell \, \mathbb{I}\{T_i = t - \ell\} + \sum_{k=0}^{K} \tau_k \, \mathbb{I}\{T_i = t + k\} + \varepsilon_{it}$$

Key assumption: Parallel trends.



Synth treatment effect: average all the synths

- Suppose the first J units are treated at times $T_1, \ldots, T_J$
- 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

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

Average of J separate synth estimates

Synth treatment effect: average treated unit

Or, we can think of it as Synth estimate for average treated unit

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

Two definitions of ATT

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

Optimization problem

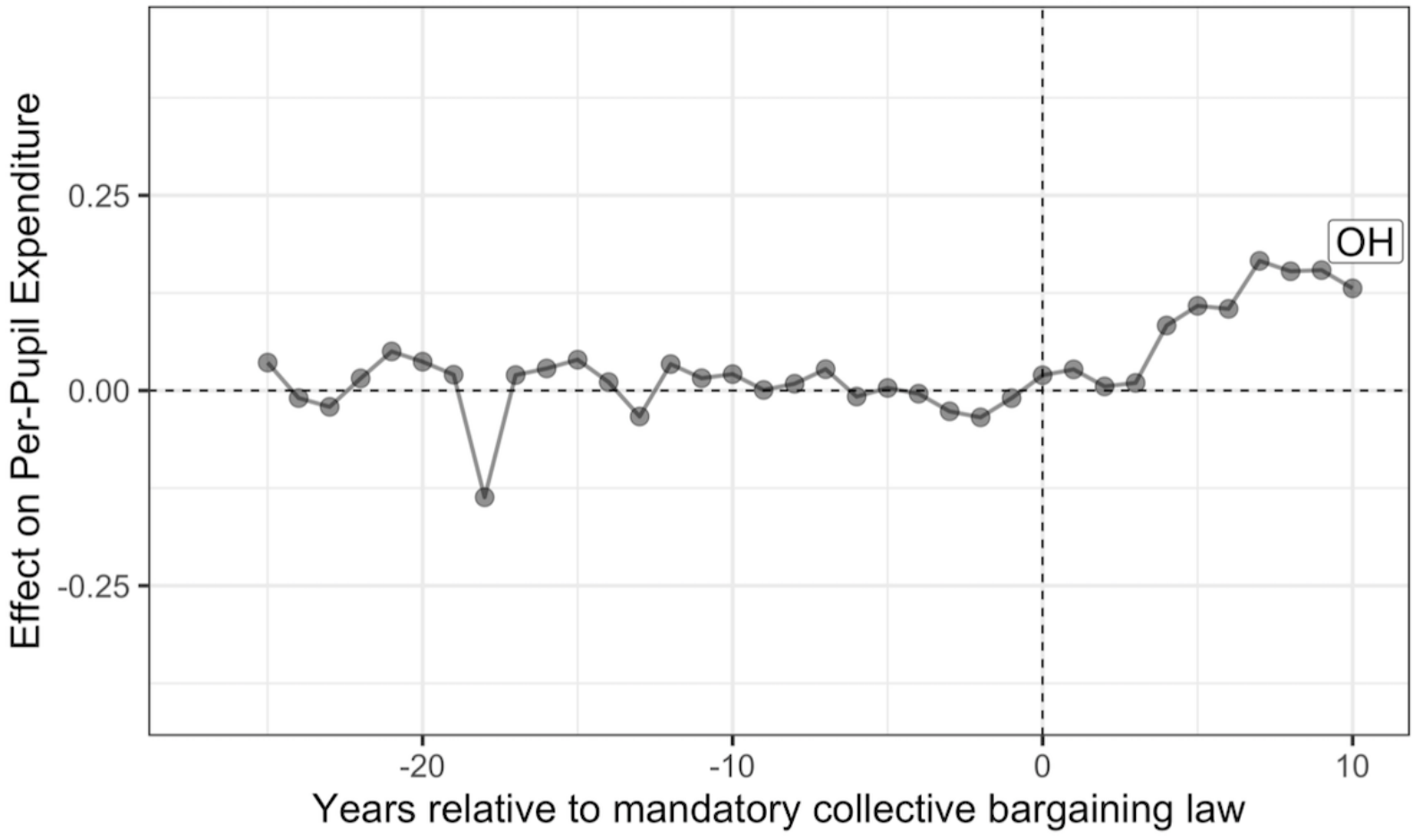
Do we want to optimize the sum of the separate imbalances or the imbalance of the sum (the pooled imbalance)?

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

where j is treatment group and i is donor pool units. Notice summations are inside or outside the norm

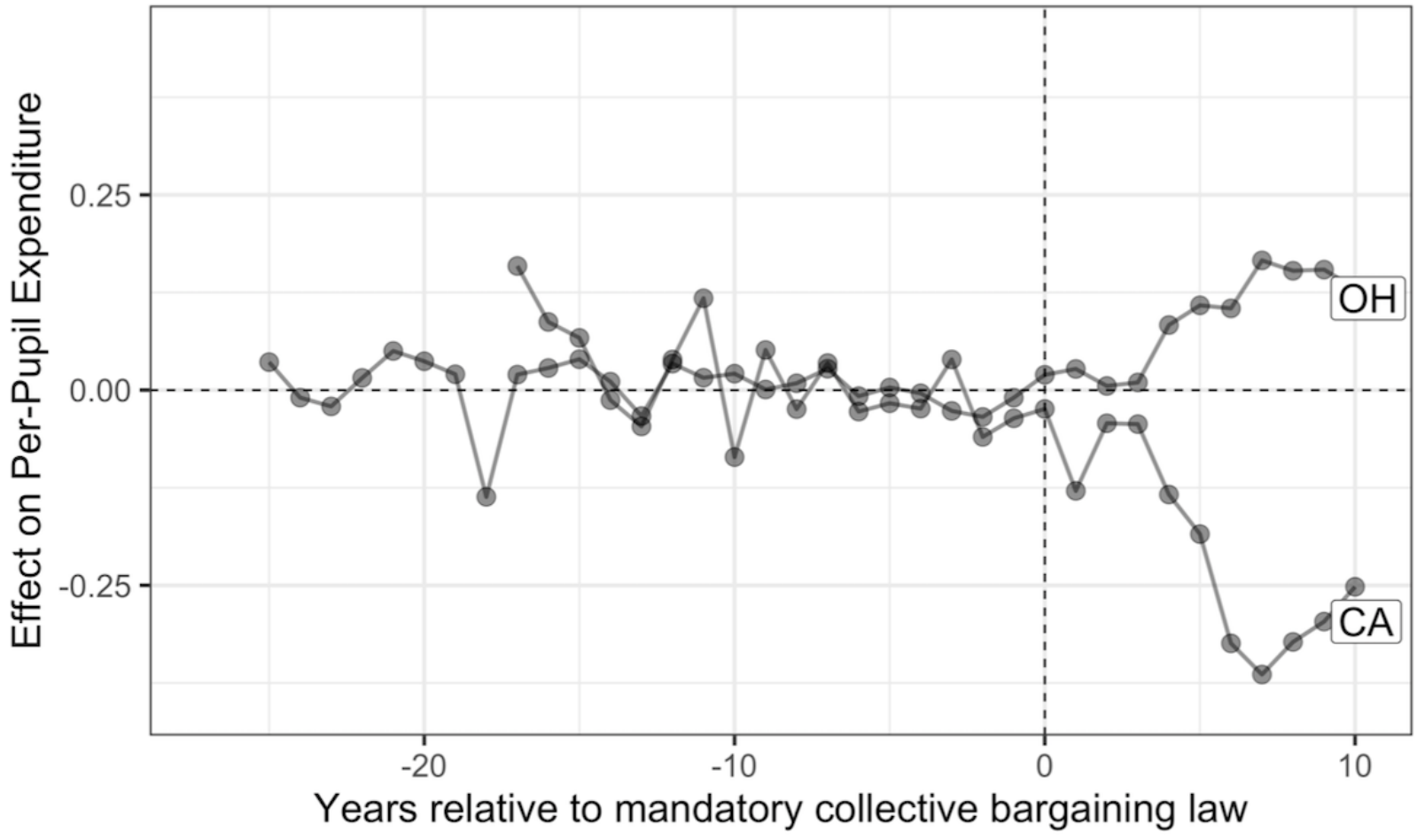
Separate SCM

The synthetic control for Ohio is pretty good.



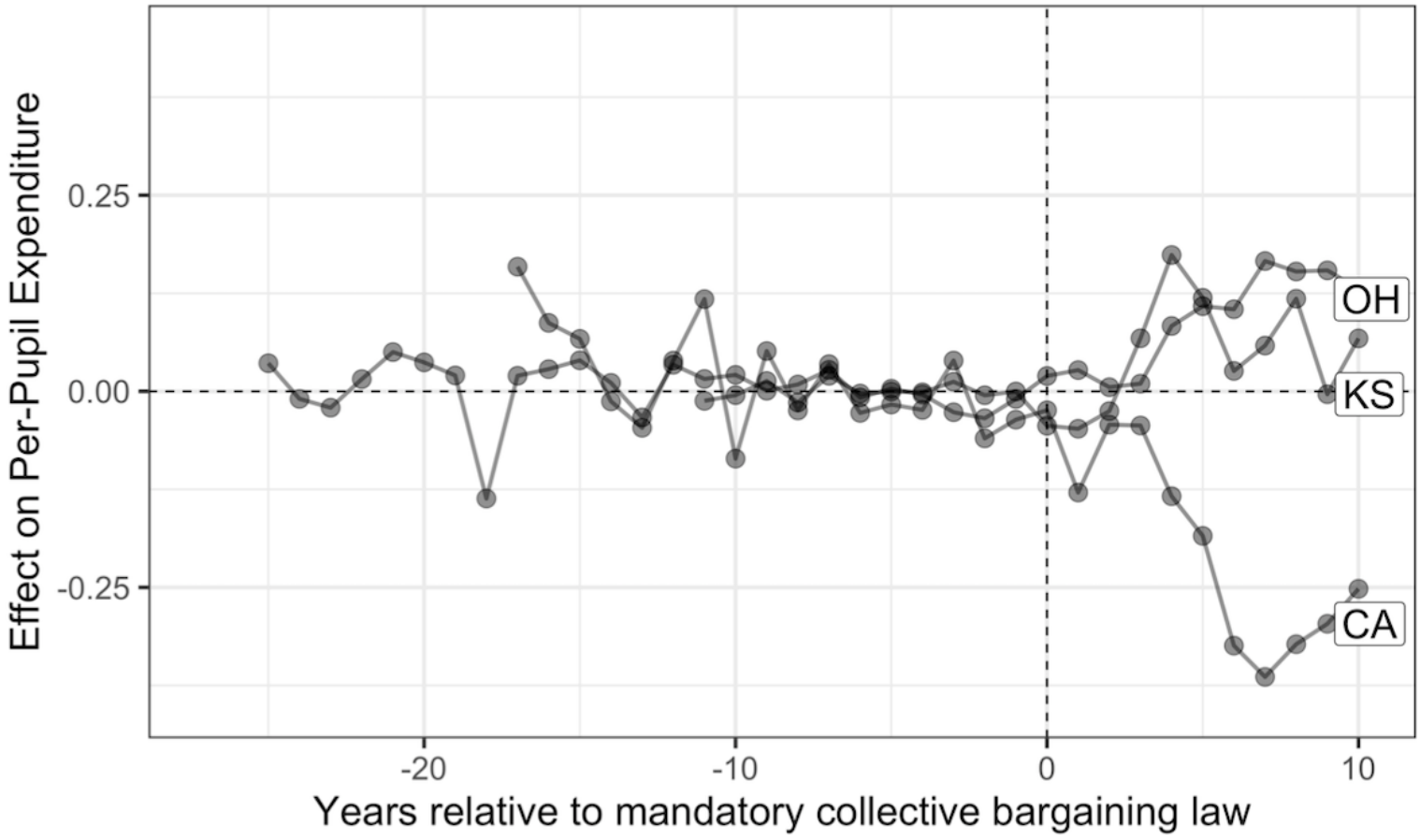
Separate SCM

We also fit California pretty well.



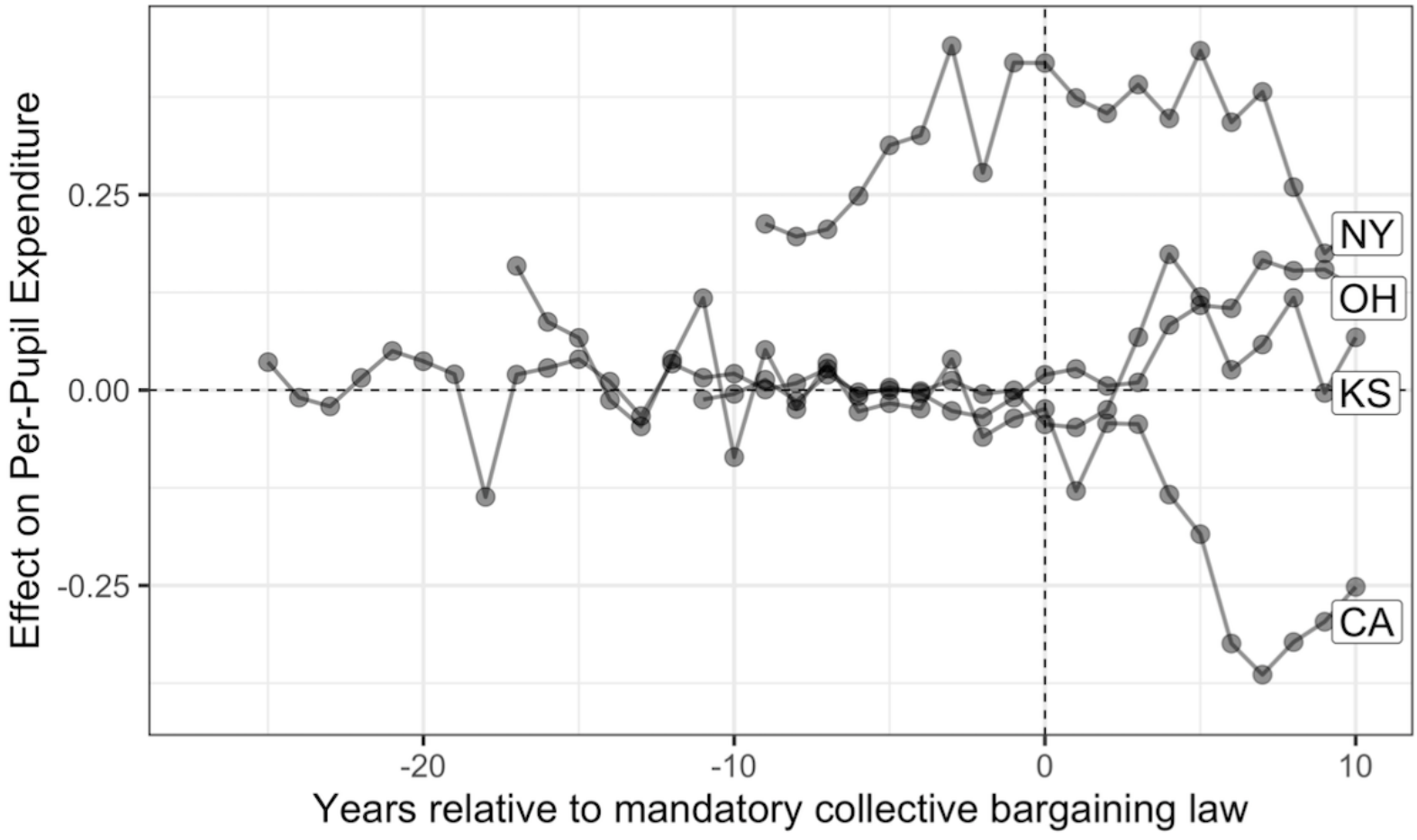
Separate SCM

And Kansas.



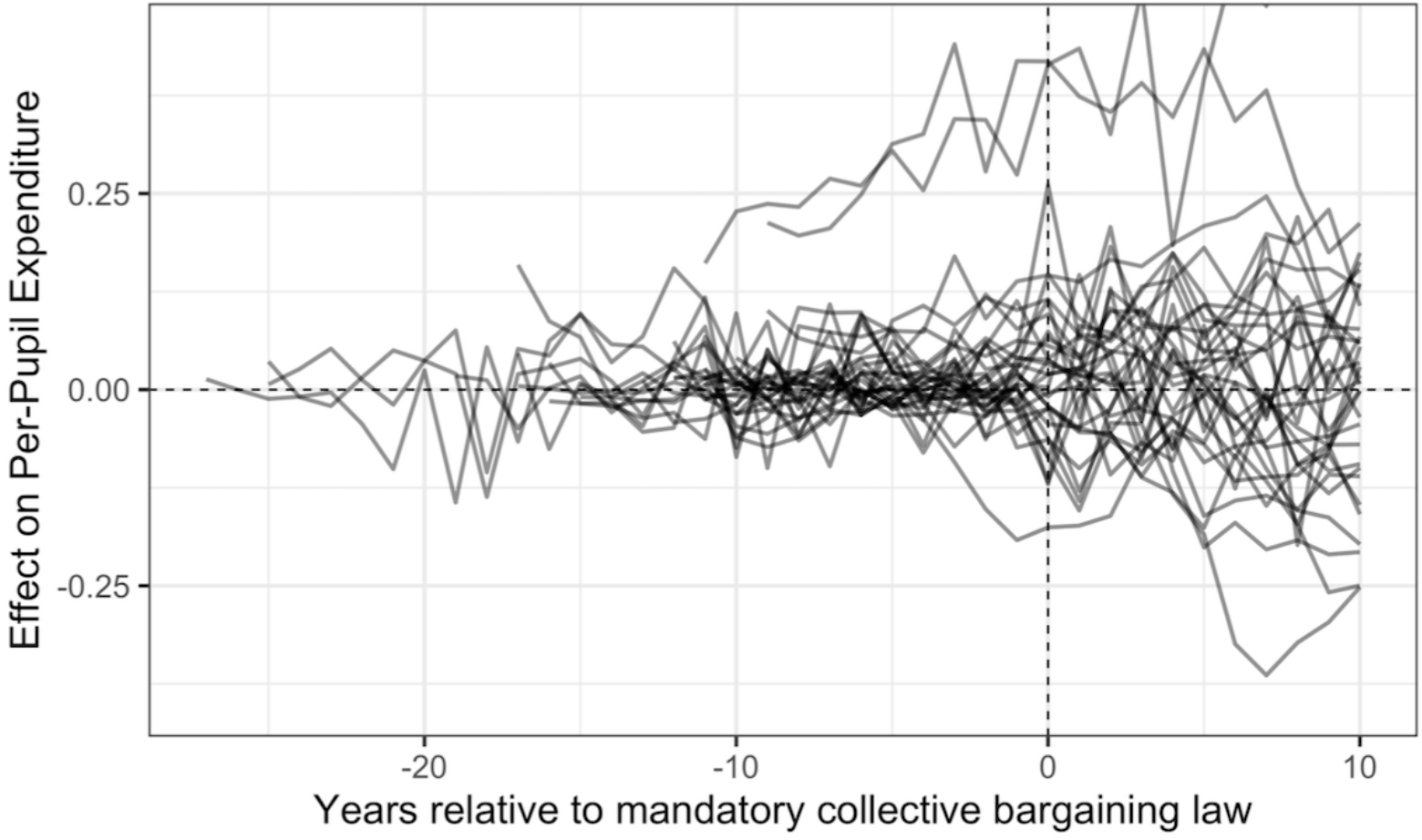
Separate SCM

But we can't fit New York well at all.



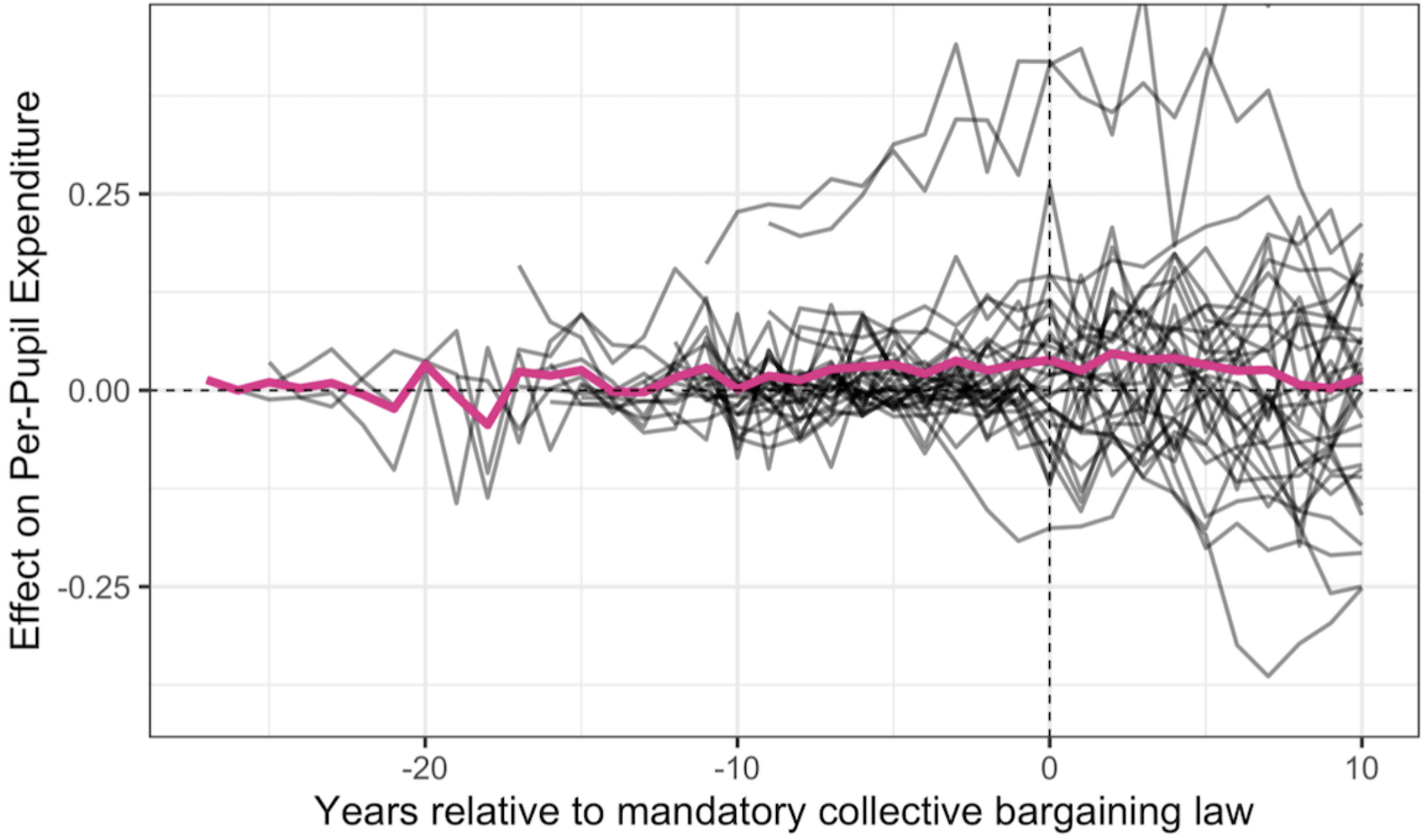
Separate SCM

The full set of separate SCM estimates.



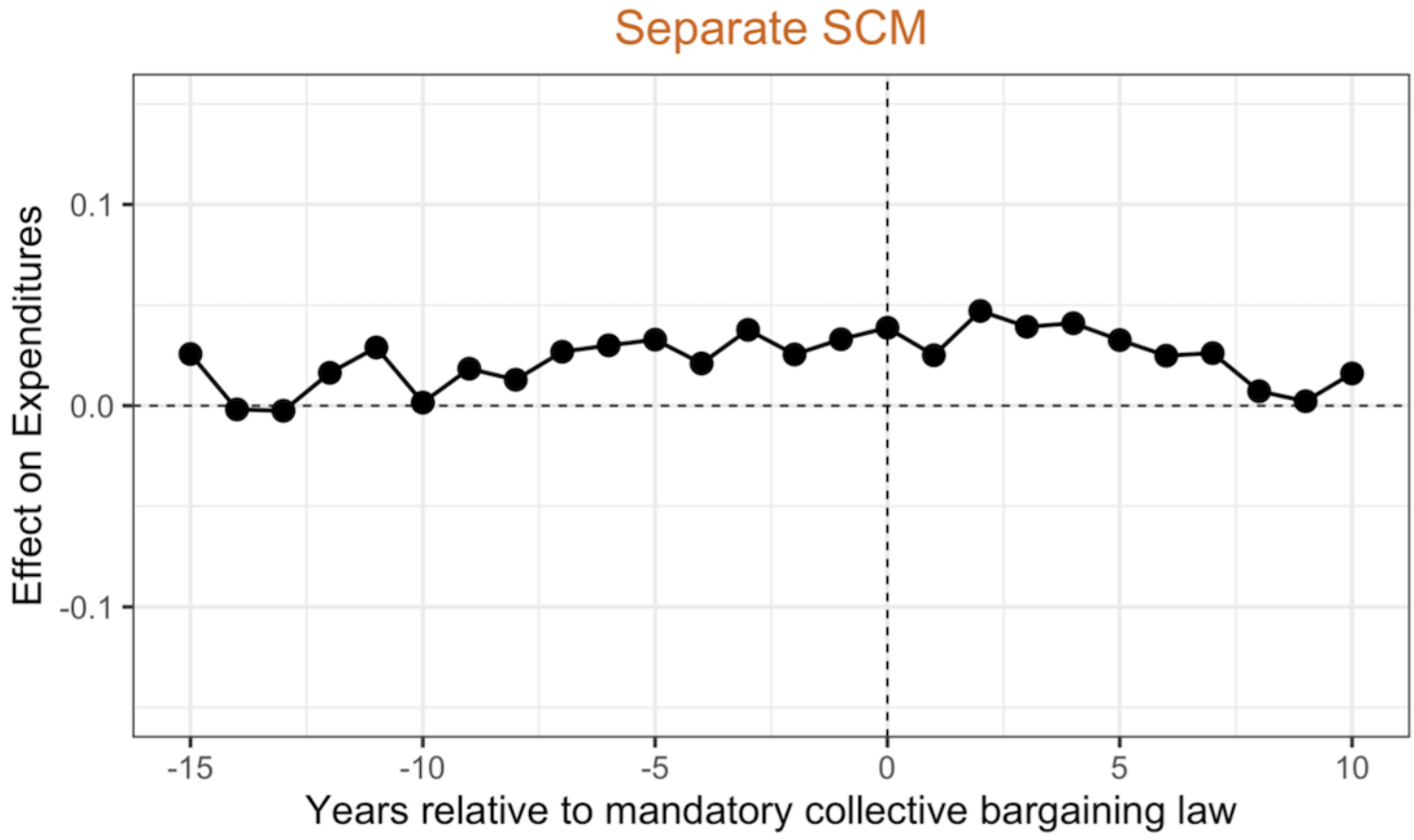
Separate SCM

The "separate SCM" estimator averages these.



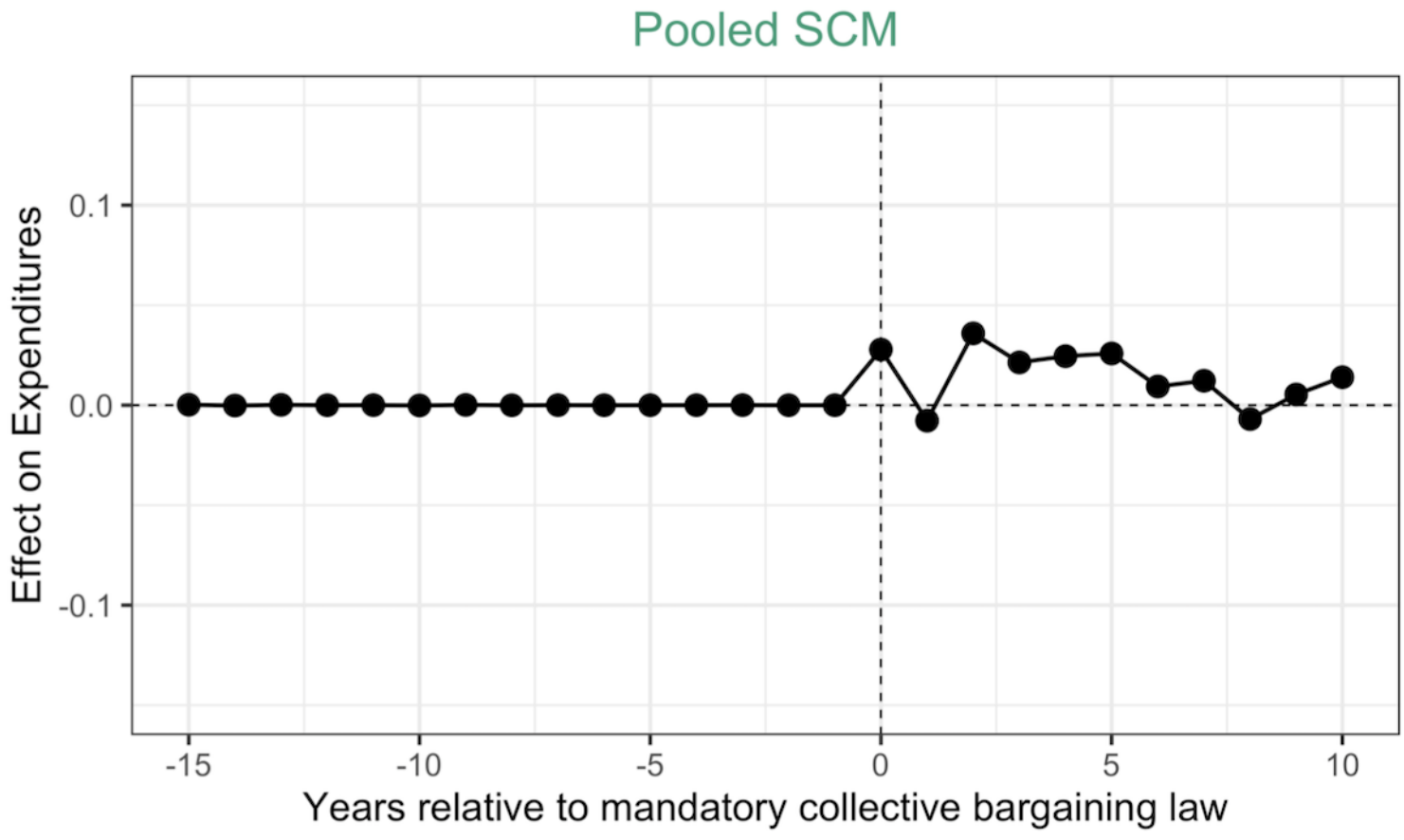
Separate SCM

The "separate SCM" balance doesn't look so good.

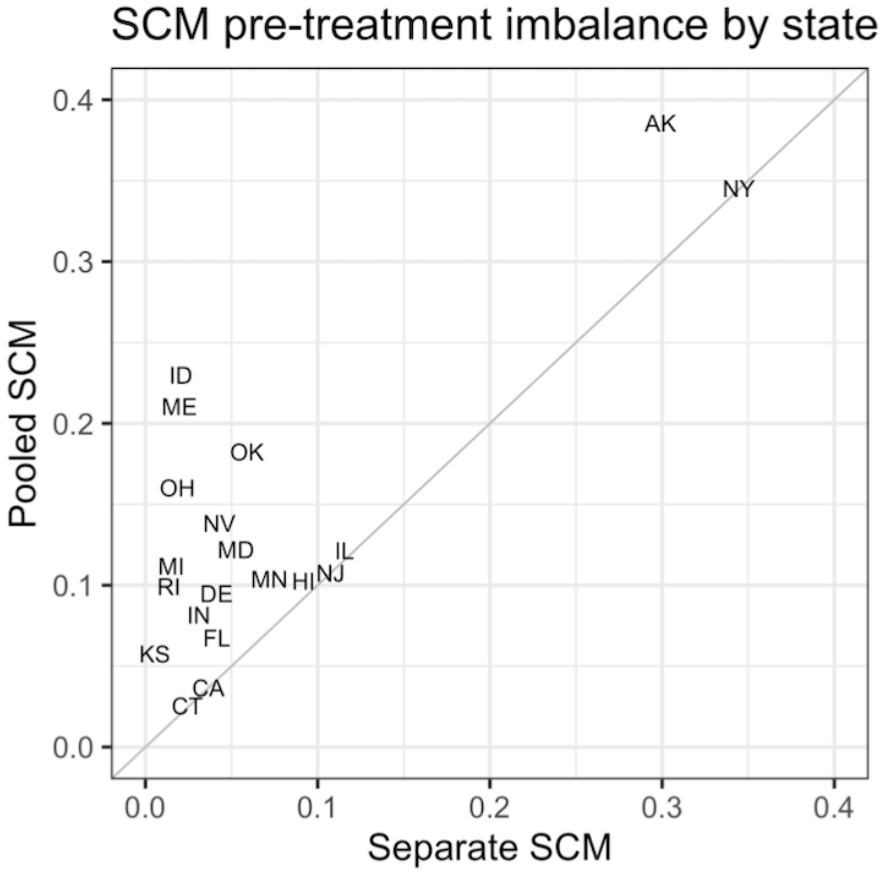


Separate SCM

By contrast, if we balance the pooled treated units, we get perfect fit.



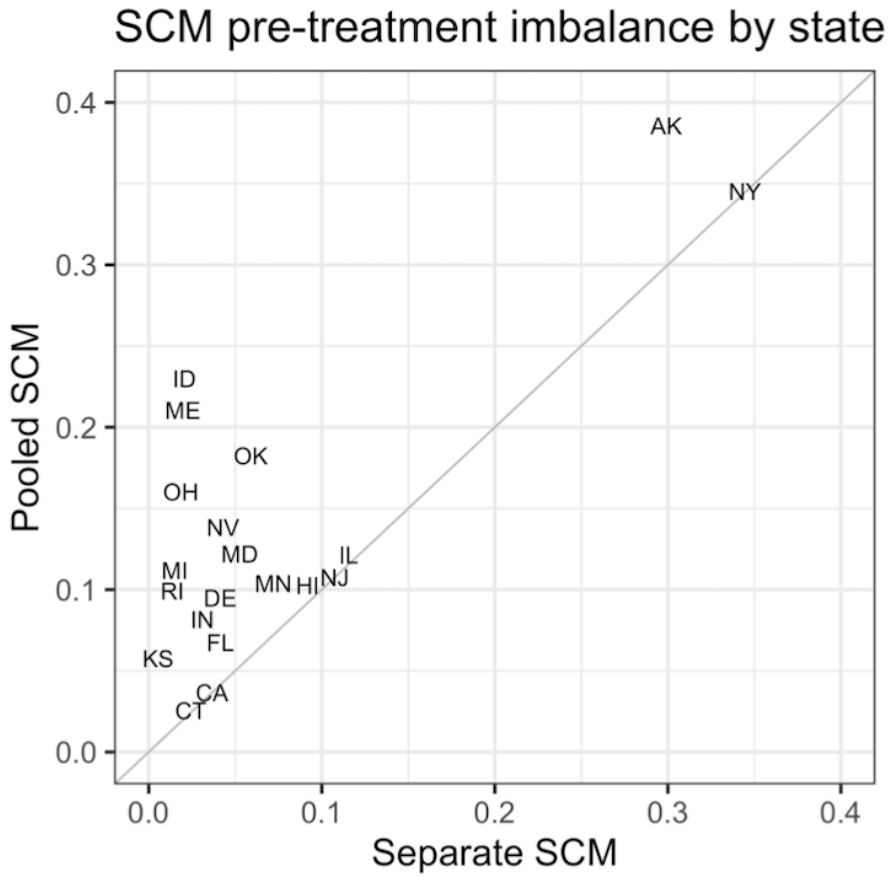
Pooled SCM generally makes state-specific fits worse



If we optimize the pooled imbalance,

- Pooled Balance is better...
- ... but State Balance is worse.
  - ⇝ Bad for **state estimates**

Pooled SCM generally makes state-specific fits worse



If we optimize the pooled imbalance,

- Pooled Balance is better...
- ... but State Balance is worse.
  - $\rightsquigarrow$ Bad for **state estimates**

$\rightarrow$ Also bad for the **average**!

- $\rightsquigarrow$ Balancing average X may not balance average $Y_{T_j+k}(0)$s.
- $\rightsquigarrow$ E.g., in factor model – factors are different in different time periods.
  - Once time is re-centered, matching the average need not mean matching the factors.

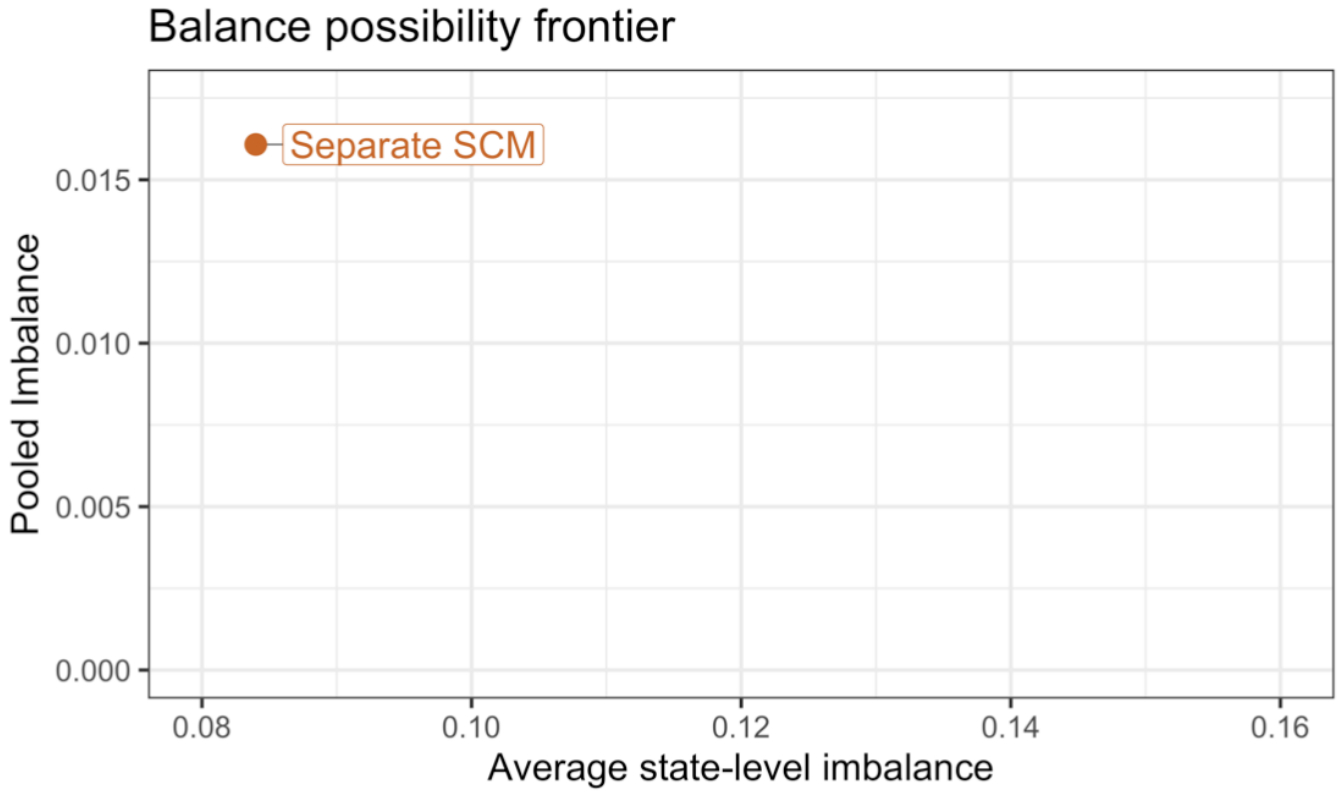
Proposal: Partially pool synth

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

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

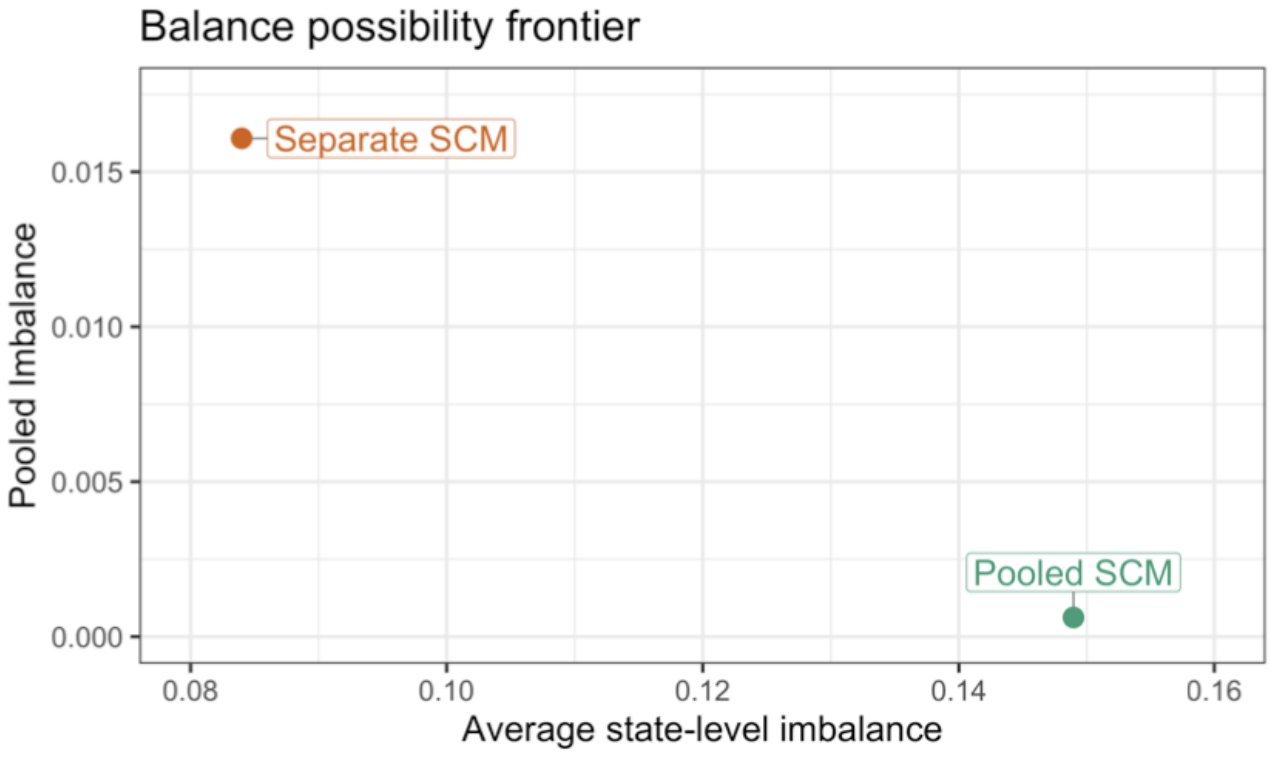
"Returns" to this are highly convex: setting v just a little below 1 yields a big improvement in state-level imbalance with very little cost in pooled imbalance

As we vary ν, we can trace out the unit-level and pooled imbalance



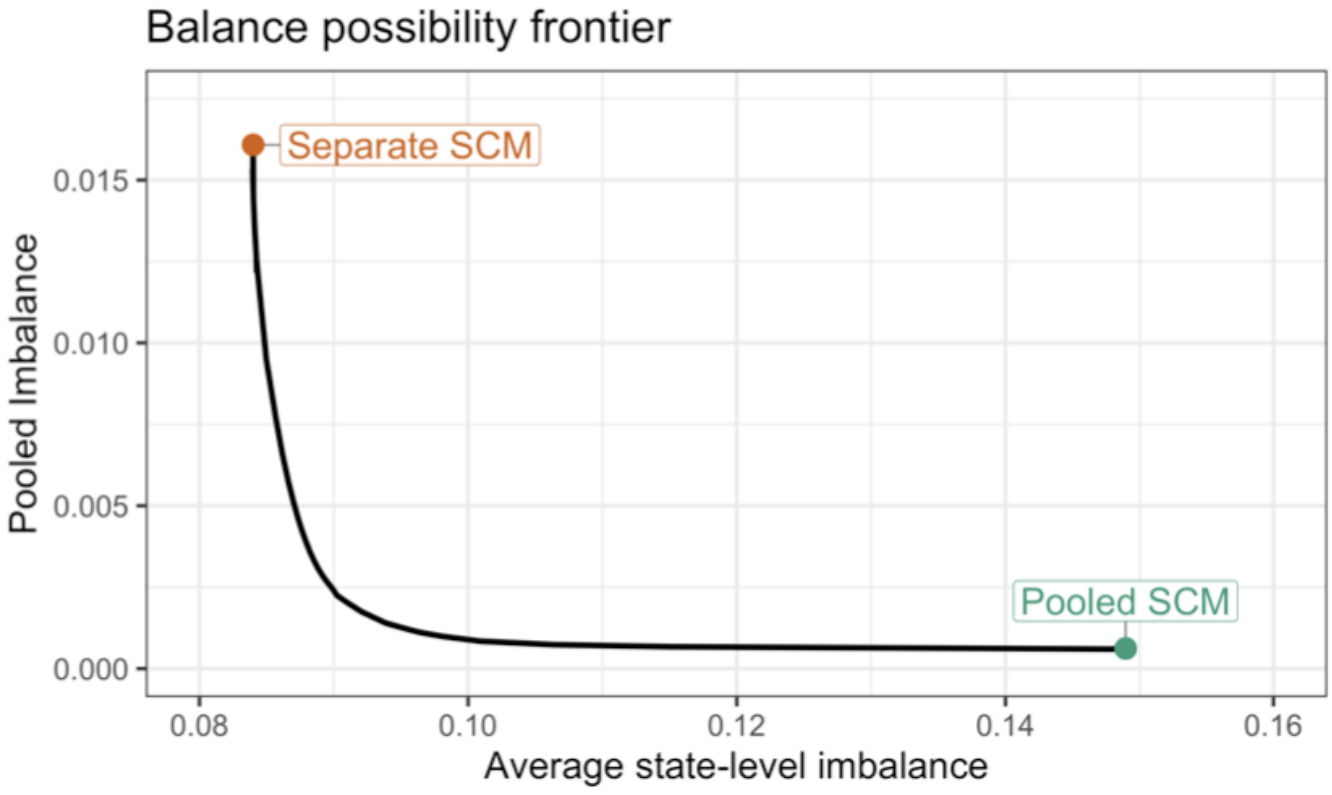
$$\min_{\Gamma \in \Delta^{\text{scm}}} \quad \nu \, \|\text{Pooled Balance}\|_2^2 + (1-\nu)\frac{1}{J}\sum_{j=1}^{J} \|\text{State Balance}_j\|_2^2 + \text{penalty}$$

As we vary ν, we can trace out the unit-level and pooled imbalance



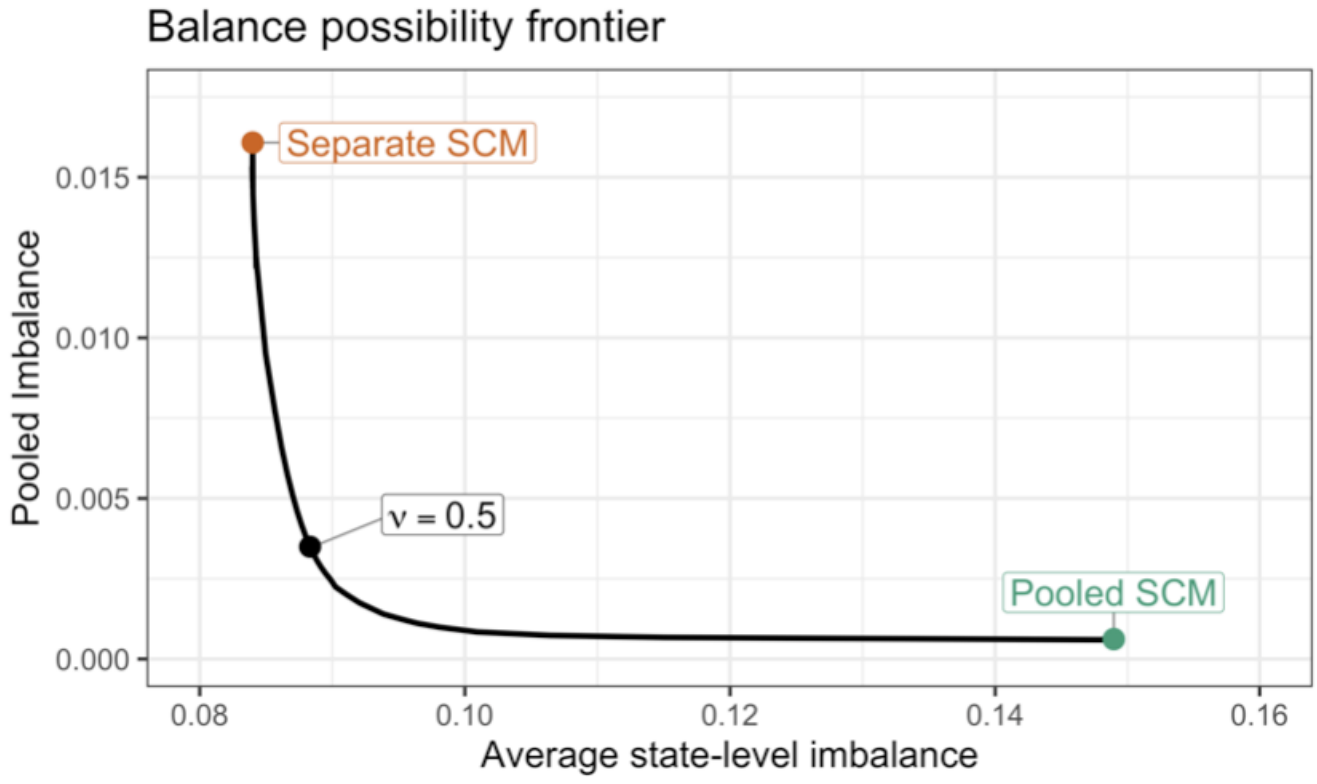
$$\min_{\Gamma \in \Delta^{\text{scm}}} \quad \nu \, \|\text{Pooled Balance}\|_2^2 + (1-\nu)\frac{1}{J}\sum_{j=1}^{J} \|\text{State Balance}_j\|_2^2 + \text{penalty}$$

As we vary ν, we can trace out the unit-level and pooled imbalance



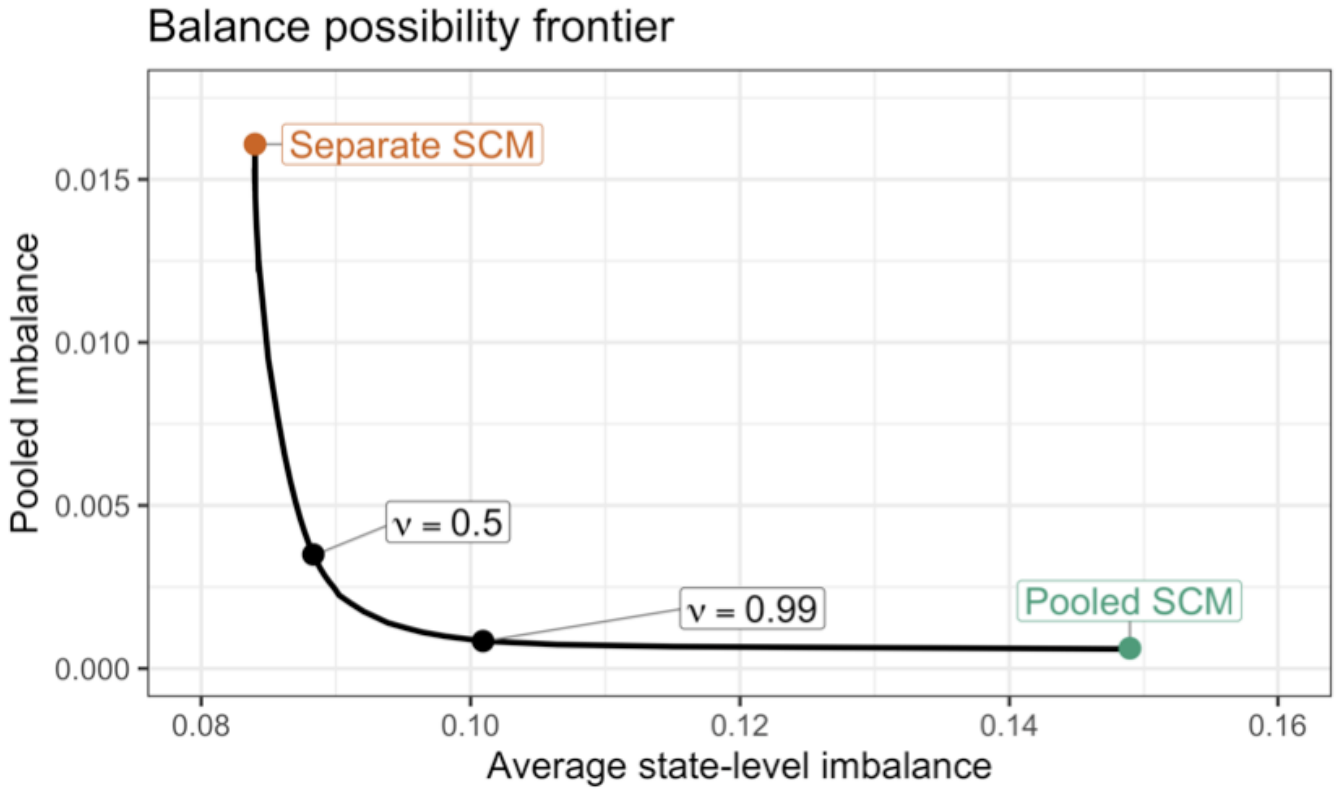
$$\min_{\Gamma \in \Delta^{\text{scm}}} \quad \nu \, \|\text{Pooled Balance}\|_2^2 + (1-\nu)\frac{1}{J}\sum_{j=1}^{J} \|\text{State Balance}_j\|_2^2 + \text{penalty}$$

As we vary ν, we can trace out the unit-level and pooled imbalance



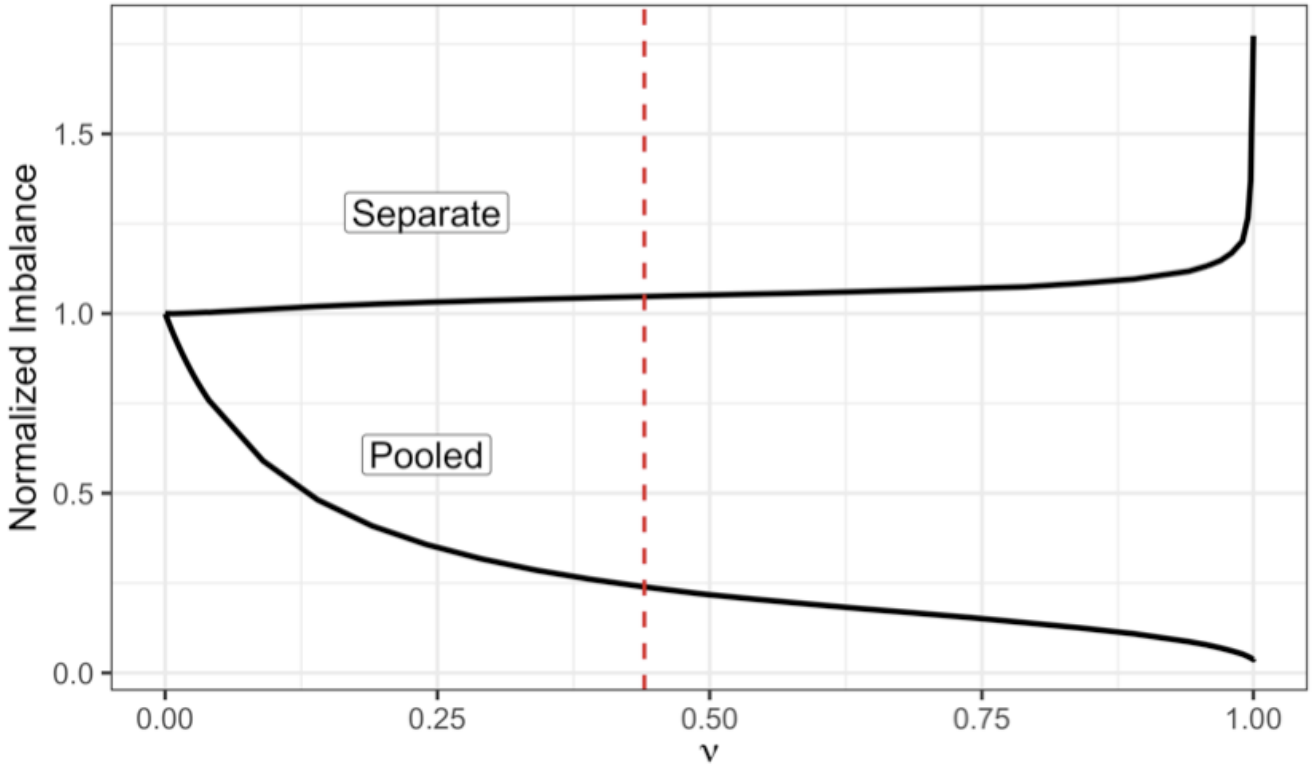
$$\min_{\Gamma \in \Delta^{\text{scm}}} \quad \nu \, \|\text{Pooled Balance}\|_2^2 + (1-\nu)\frac{1}{J}\sum_{j=1}^{J} \|\text{State Balance}_j\|_2^2 + \text{penalty}$$

As we vary ν, we can trace out the unit-level and pooled imbalance



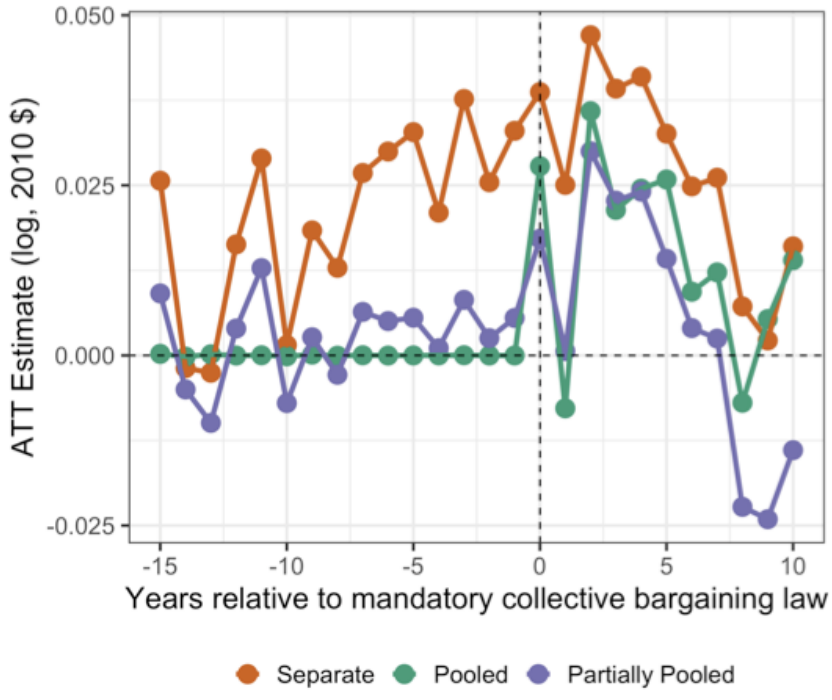
$$\min_{\Gamma \in \Delta^{\text{scm}}} \quad \nu \, \|\text{Pooled Balance}\|_2^2 + (1-\nu)\frac{1}{J}\sum_{j=1}^{J} \|\text{State Balance}_j\|_2^2 + \text{penalty}$$

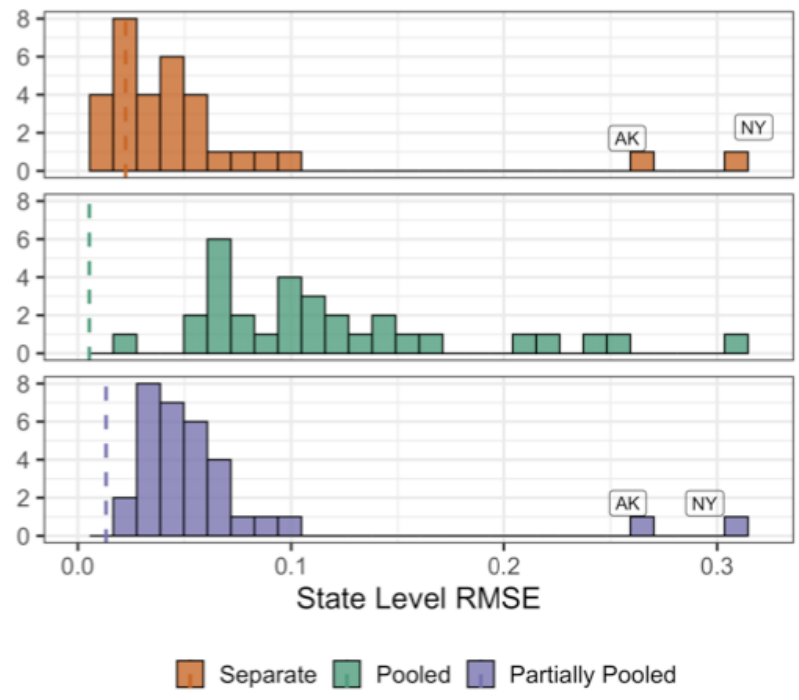
Another view of the tradeoff



Heuristic for $\nu = \dfrac{\|\text{Pooled Balance}\|_2}{\frac{1}{\sqrt{J}} \sum_{j=1}^{J} \|\text{State Balance}_j\|_2}$ fit with $\nu = 0$

Comparing separate, pooled, and partially pooled SCM





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: $\overline{Y}^{pre}_{i,T_j}$
- Estimate synth using residuals (Doudchenko and Imbens 2017; Ferman and Pinto 2018)

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

where $Y(0) = Y^0$

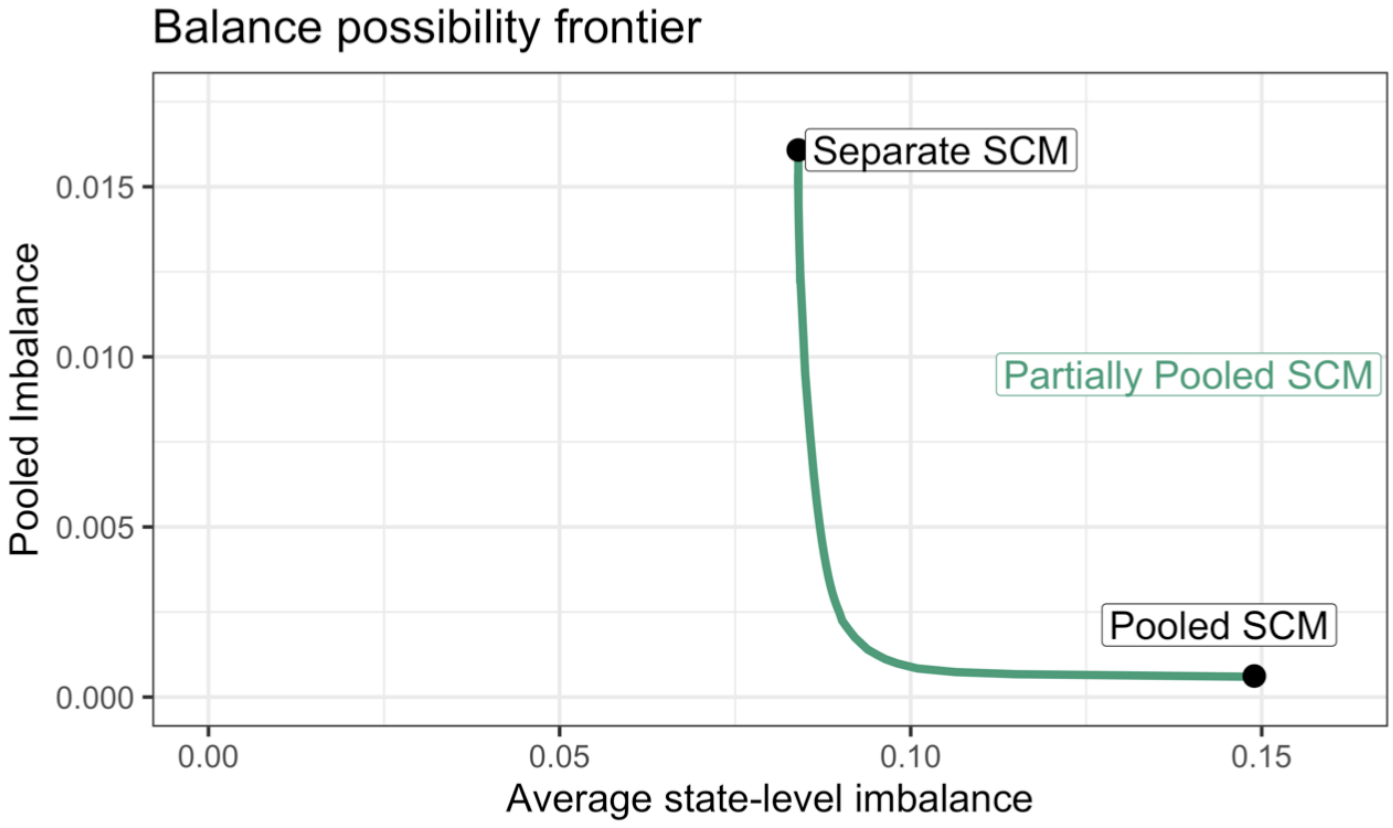
Special case: weighted event study

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

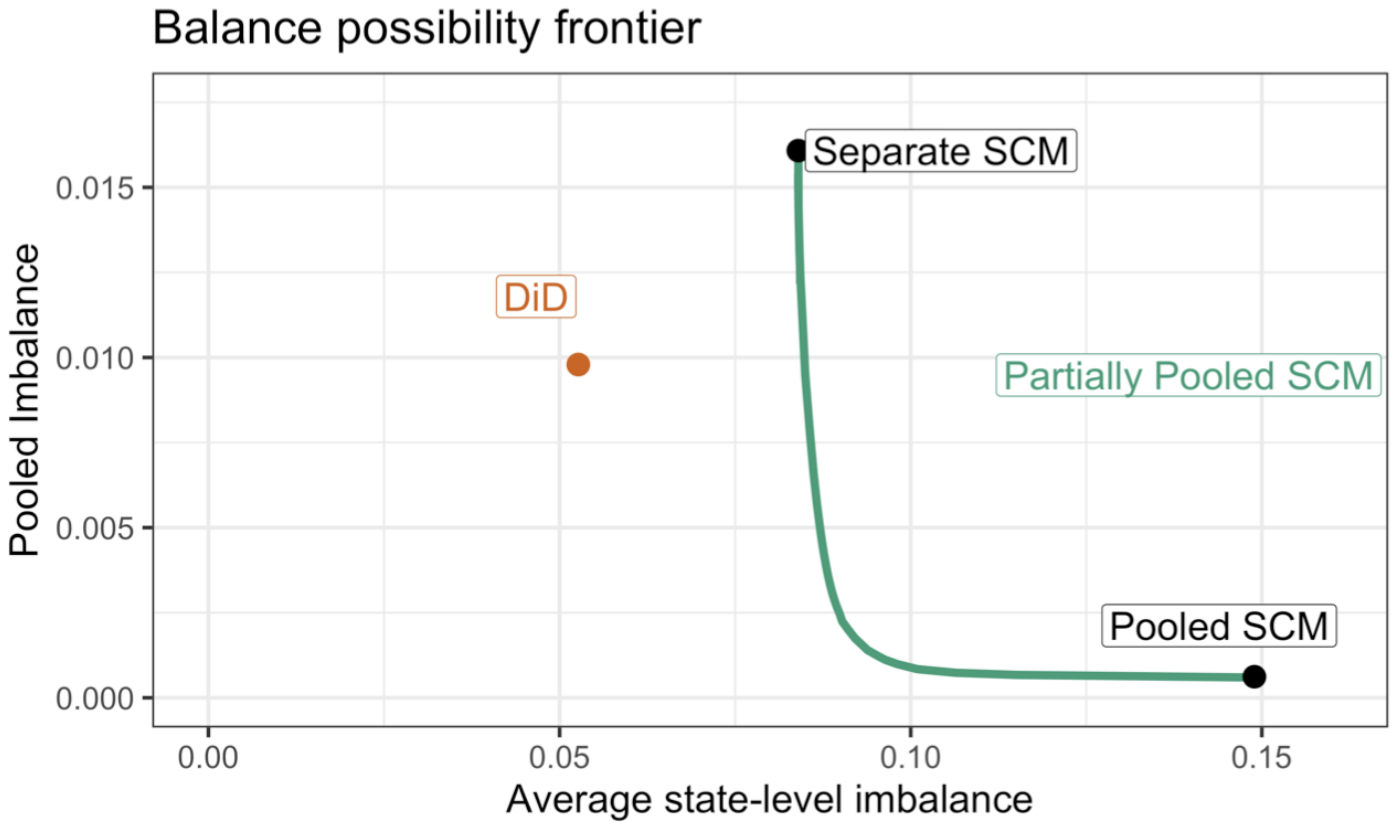
$$\widehat{\delta}_{jk}^{aug} = \left(Y_{j,T_j+k} - \overline{Y}_{j,T_j}^{pre} \right) - \sum_{i=1}^{N} \widehat{w}_{ij} \left(Y_{i,T_j+k} - \overline{Y}_{i,T_j}^{pre} \right)$$

Uniform weights correspond to “standard DiD”

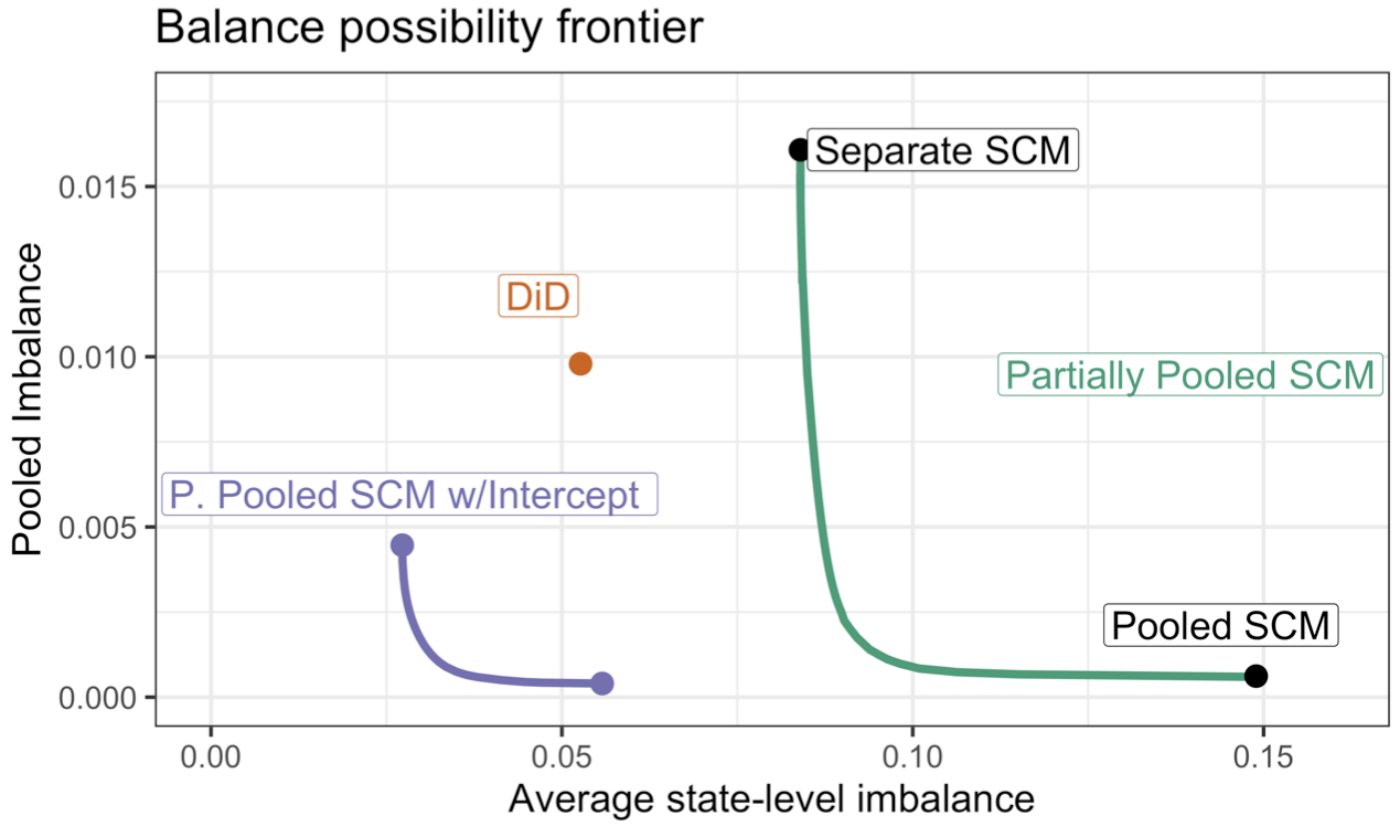
Weighted event studies shift the balance possibility frontier far inward



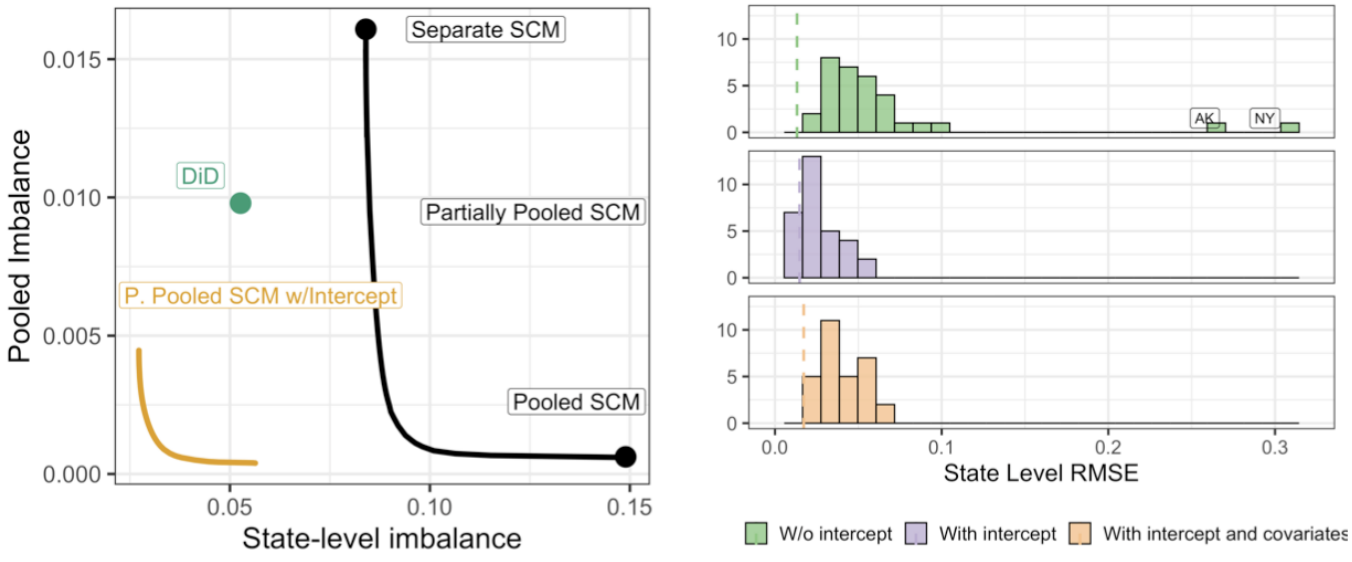
Weighted event studies shift the balance possibility frontier far inward



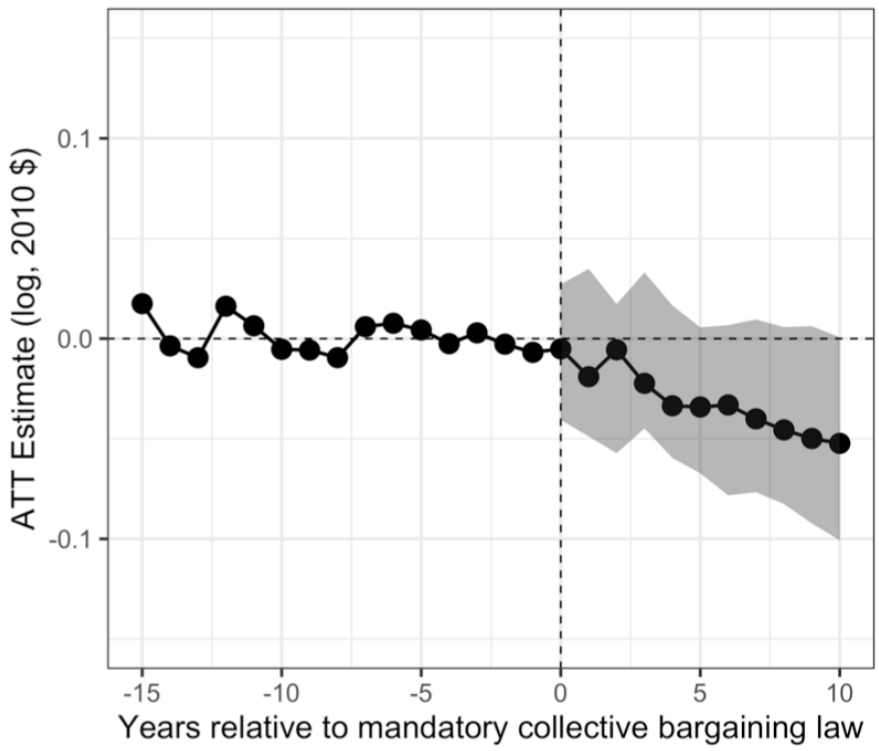
Weighted event studies shift the balance possibility frontier far inward



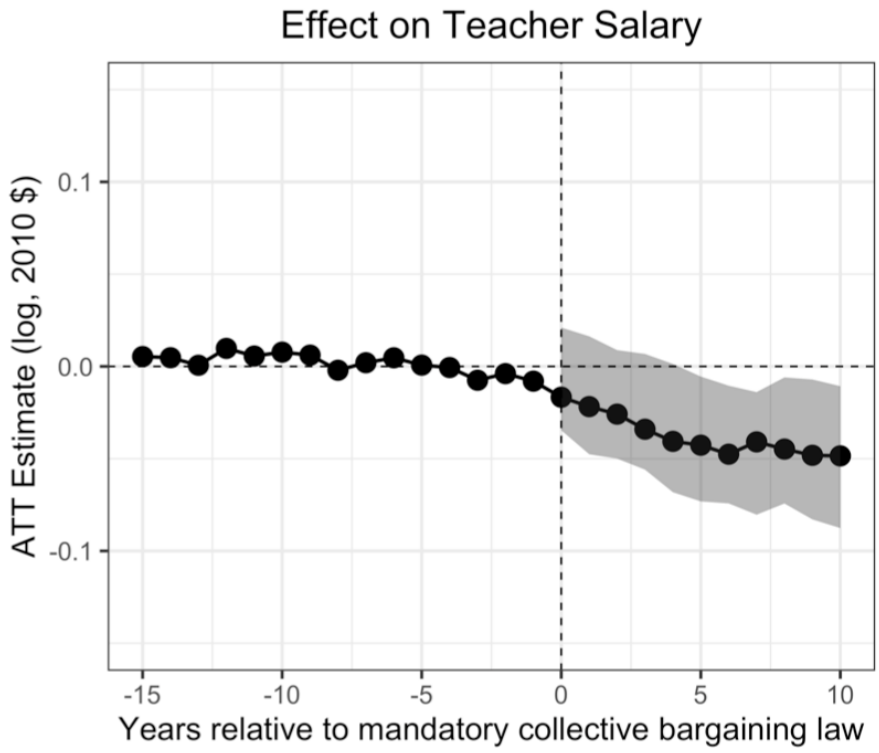
Weighted event studies shift the balance possibility frontier far inward



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

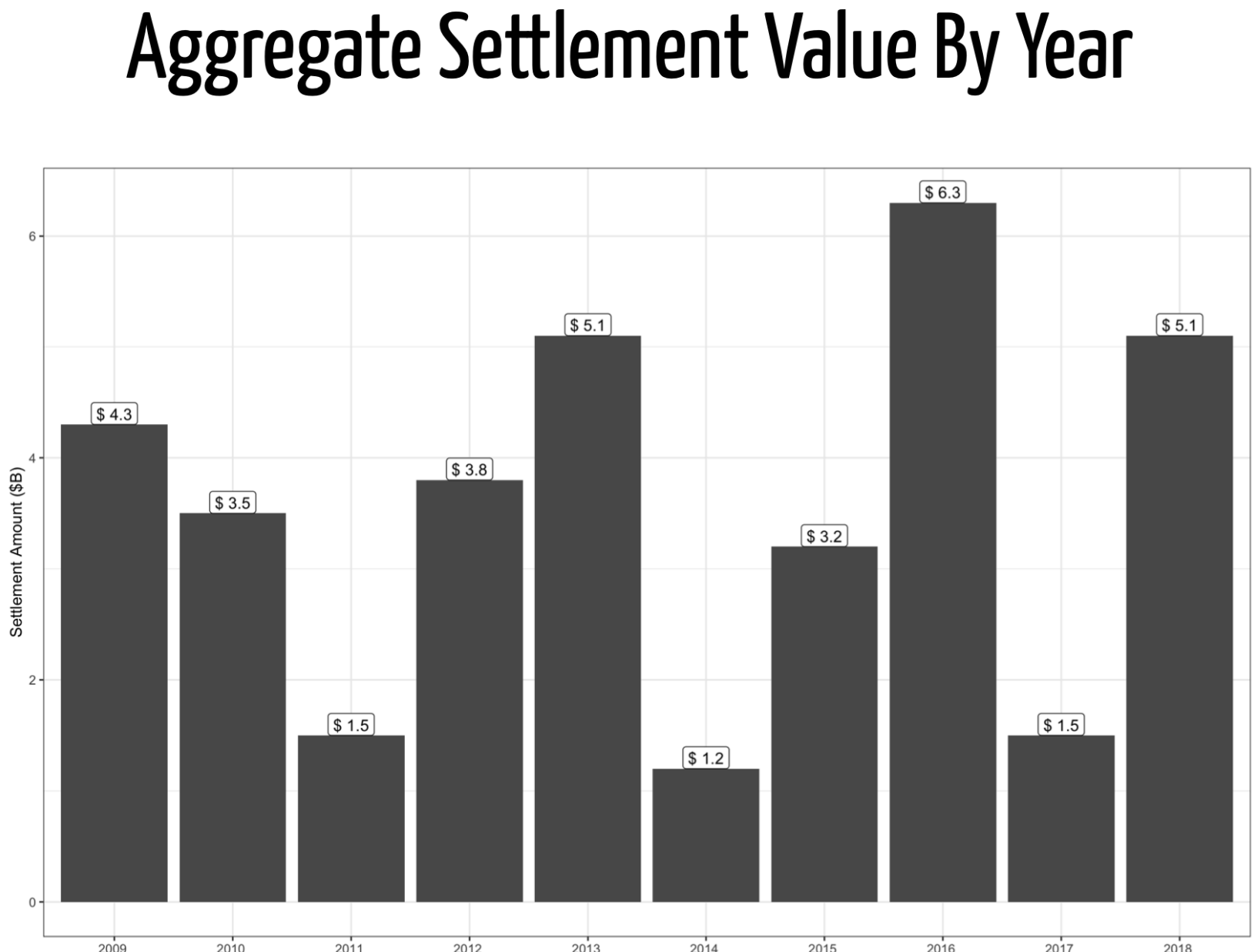
Possibilities for detecting corruption

- 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
Settlement Amount ($B)
0
2
4
6
$ 4.3
$ 3.5
$ 1.5
$ 3.8
$ 5.1
$ 1.2
$ 3.2
$ 6.3
$ 1.5
$ 5.1
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018

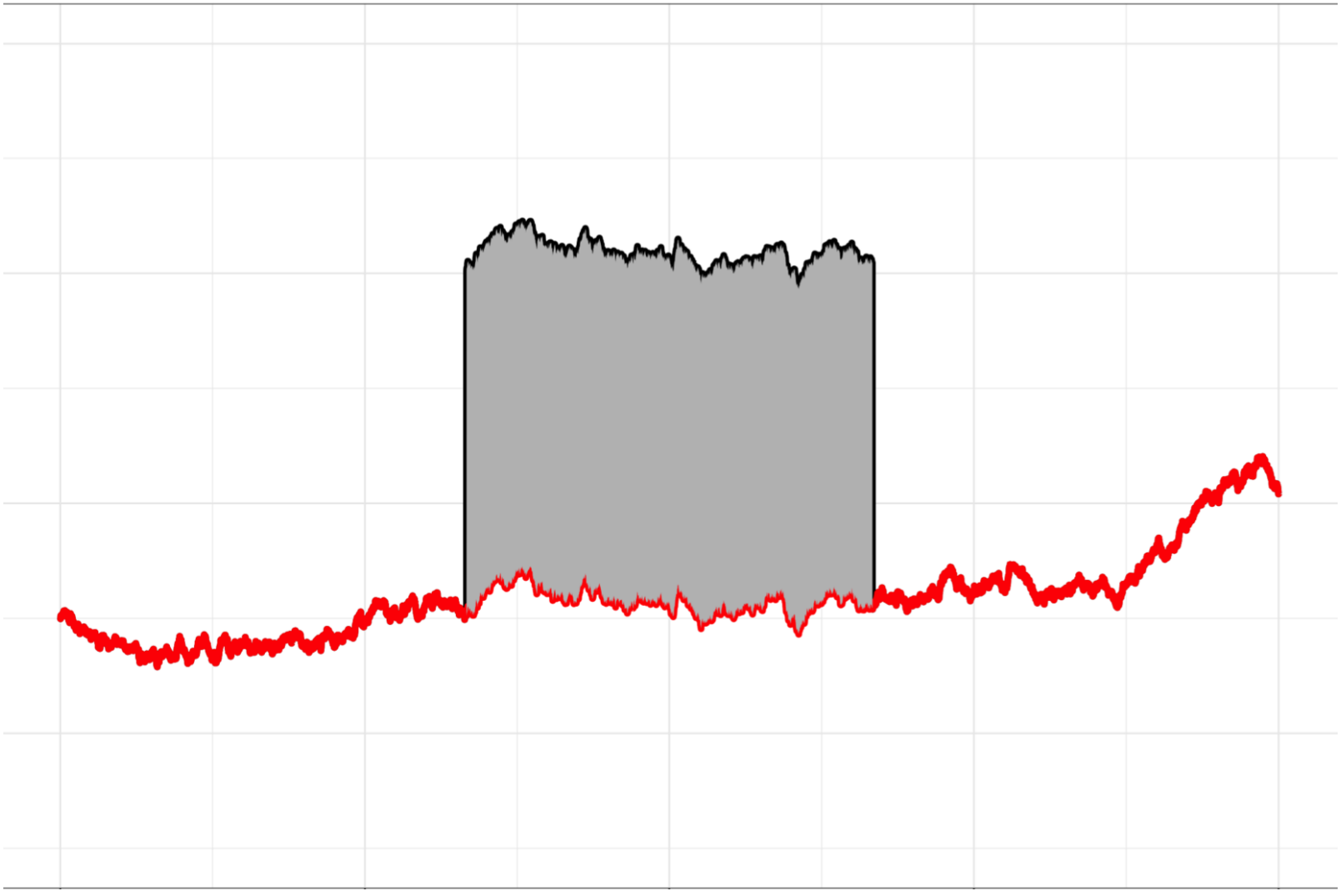
Event studies and securities litigation

- 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. It's effect on daily return is $r^1_{i,t} - r^0_{i,t}$ and we want to estimate $r^0_{i,t}$ with $\widehat{r}^0_{i,t}$
- Construct a predicted residual as $\widehat{\varepsilon}_{i,t} = r_{i,t} - \widehat{r}^0_{i,t}$
- 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 $\widehat{\varepsilon}_{i,t} = r_{i,t} - \widehat{\beta} X_{i,t}$ 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

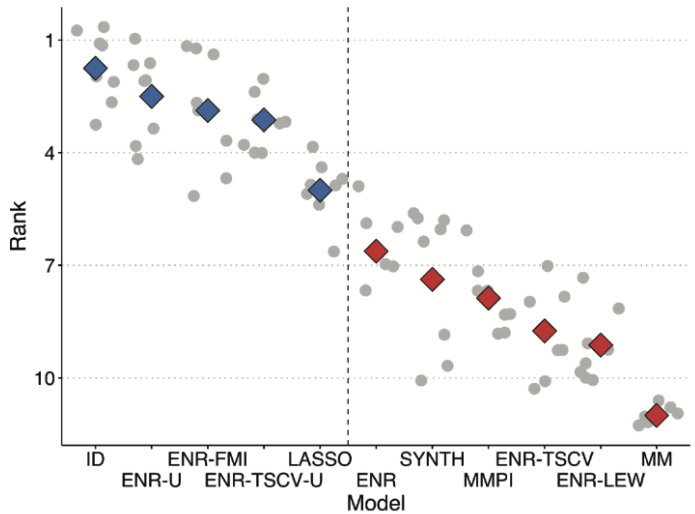


Figure 1: Distribution of Specification Ranks Across Models and Tests.

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 ($\widehat{R}^k_{oos}$ and $\widehat{R}^k_{het}$, 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.

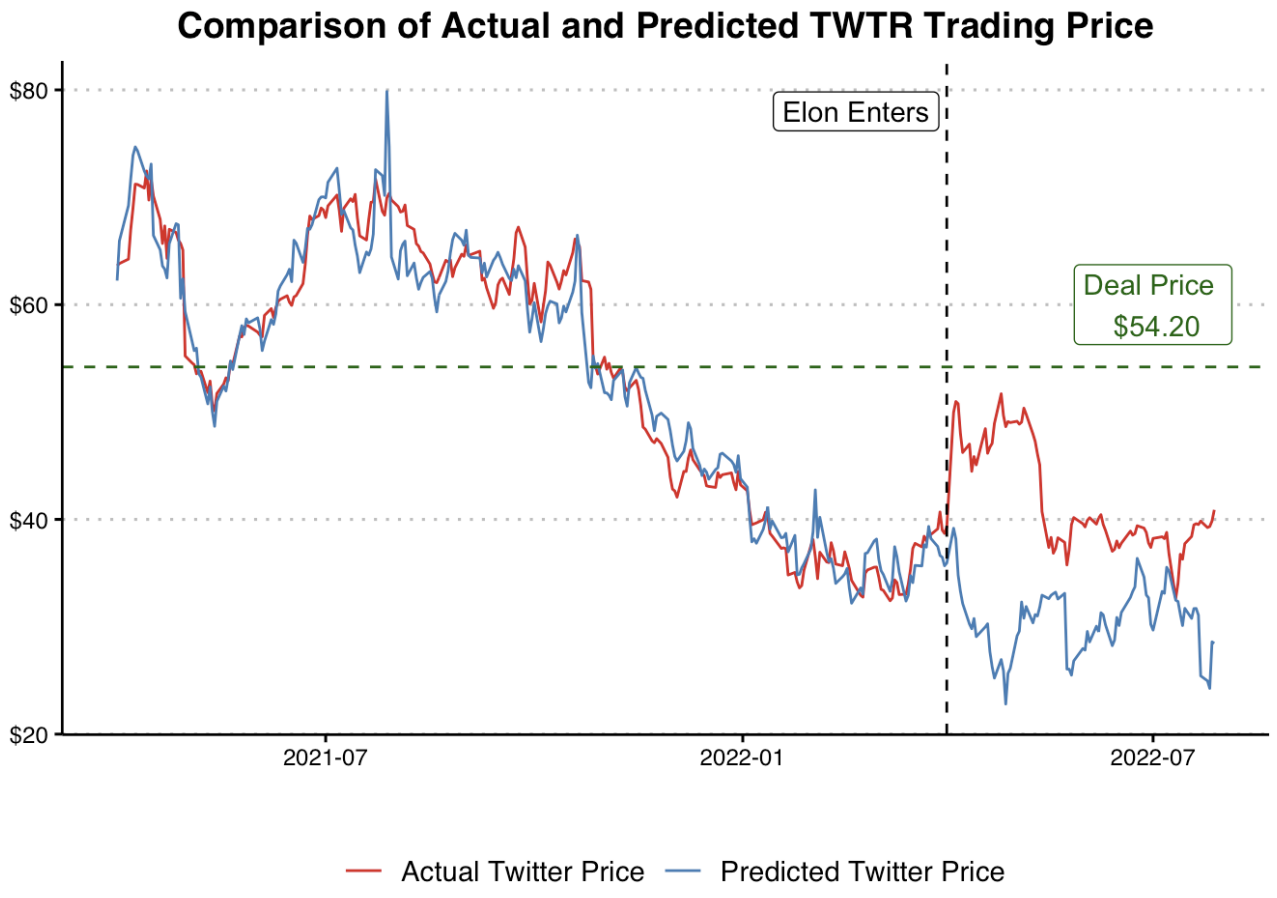
Elon Musk example

- 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

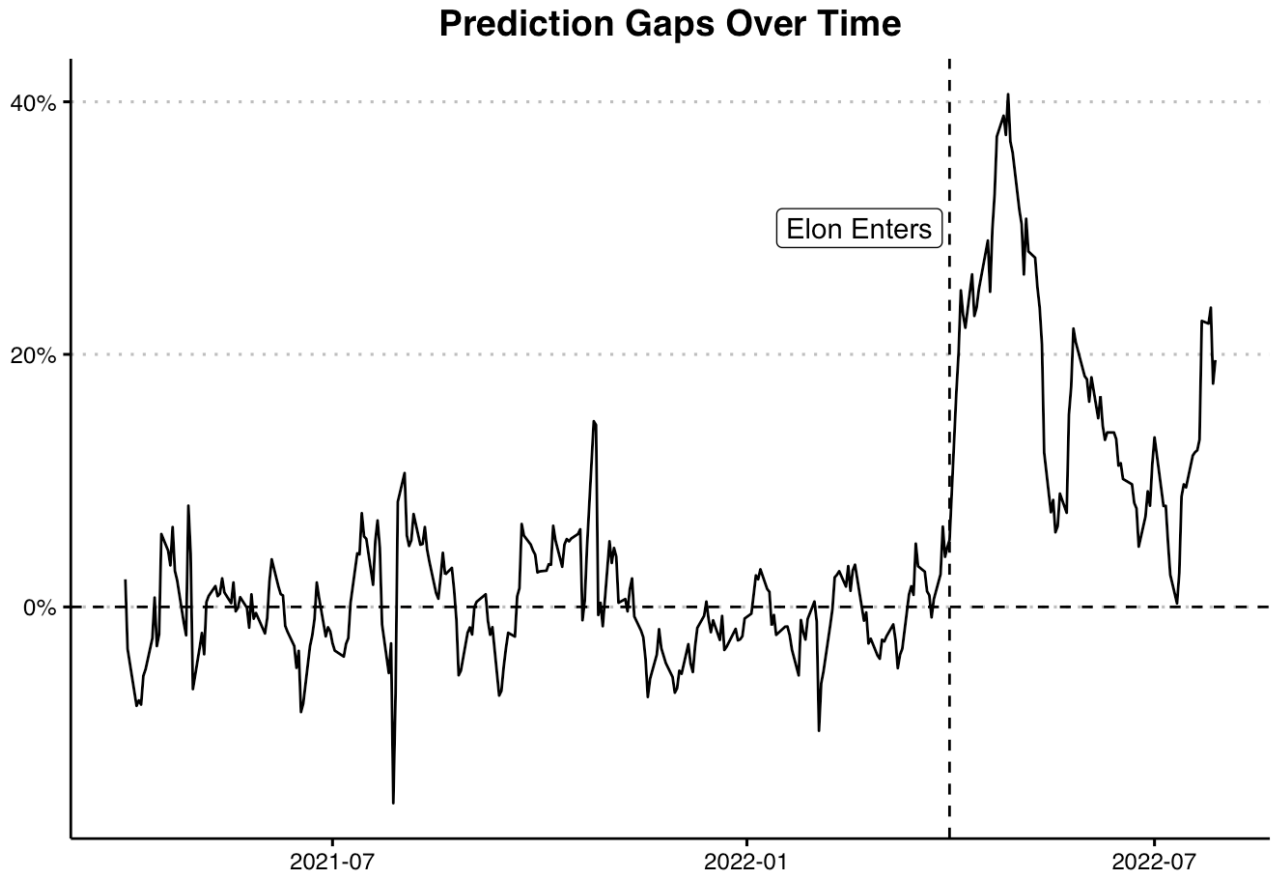
Purpose of the exercise

“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


Comparison of Actual and Predicted TWTR Trading Price
$80
$60
$40
$20
Elon Enters
Deal Price
$54.20
2021-07
2022-01
2022-07
Actual Twitter Price
Predicted Twitter Price

Basic idea


Prediction Gaps Over Time
40%
20%
0%
Elon Enters
2021-07
2022-01
2022-07

Roadmap

Synthetic control

- Abadie, Diamond and Hainmueller
- Matrix completion with nuclear norm
- Synthetic difference-in-differences
- Augmented Synthetic Control
- Augmented Synthetic Control with Staggered Rollout
- Event studies in finance

Concluding remarks

Summarizing

- Randomized treatments are great but not always available
- Causal inference methods can utilize naturally occurring variation, but still must make adequate adjustments to find suitable controls
- Synthetic control and recent work on event studies can be possibilities
- Thank you!