Synthetic Differences in Differences

Susan Athey (Stanford University)

(with Dmitry Arkhangelsky, Nick Doudchenko, Skip Hirshberg)

(Guido Imbens, and Stefan Wager)

References:

Arkhangelsky, Dmitry, Susan Athey, David A. Hirshberg, Guido W. Imbens, and Stefan Wager. Synthetic difference in differences, 2019.

Athey, Susan, Mohsen Bayati, Mohsen, Nick Doudchenko, Guido Imbens, and Khashayar Khosravi, (2018). Matrix completion methods for causal panel data models.

- California's anti-smoking legislation (Proposition 99) took effect in 1989.

- **What is the causal effect of the legislation on smoking rates in California in 1989?**

- We **observe** smoking rates in California in 1989 given the legislation. We need to **impute** the **counterfactual** smoking rates in California in 1989 had the legislation not been enacted.

- We have data in the absence of smoking legislation in California prior to 1989, and for other states both before and in 1989. (and other variables, but not of essence)

Set Up: we observe (in addition to covariates):

$$\mathbf{Y} = \begin{pmatrix} Y_{11} & Y_{12} & Y_{13} & \dots & Y_{1T} \\ Y_{21} & Y_{22} & Y_{23} & \dots & Y_{2T} \\ Y_{31} & Y_{32} & Y_{33} & \dots & Y_{3T} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_{N1} & Y_{N2} & Y_{N3} & \dots & Y_{NT} \end{pmatrix} \qquad \text{(realized outcome).}$$

$$\mathbf{W} = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 1 \\ 0 & 0 & 0 & \dots & 1 & 1 \end{pmatrix} \qquad \text{(binary treatment).}$$

• rows of $\mathbf{Y}$ and $\mathbf{W}$ correspond to units (e.g., states), columns correspond to time periods (years).

In terms of potential outcome matrices $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$:

$$\mathbf{Y}(0) = \begin{pmatrix} \checkmark & \checkmark & \dots & \checkmark & \checkmark \\ \checkmark & \checkmark & \dots & \checkmark & \checkmark \\ \checkmark & \checkmark & \dots & \checkmark & \checkmark \\ \vdots & \vdots & \ddots & \vdots & \\ \checkmark & \checkmark & \dots & ? & ? \\ \checkmark & \checkmark & \dots & ? & ? \end{pmatrix} \qquad \mathbf{Y}(1) = \begin{pmatrix} ? & ? & \dots & ? & ? \\ ? & ? & \dots & ? & ? \\ ? & ? & \dots & ? & ? \\ \vdots & \vdots & \ddots & \vdots & \\ ? & ? & \dots & \checkmark & \checkmark \\ ? & ? & \dots & \checkmark & \checkmark \end{pmatrix}.$$

$Y_{it} = (1 - W_{it})Y_{it}(0) + W_{it}Y_{it}.$

In order to estimate the average treatment effect for the treated, (or other average, e.g., overall average effect)

$$\tau = \frac{\sum_{i,t} W_{it}\Big(Y_{it}(1) - Y_{it}(0)\Big)}{\sum_{it} W_{it}},$$

we **impute** the missing potential outcomes in $\mathbf{Y}(0)$.

Part of the talk I will focus on case with a single treated unit/time-period

$$\mathbf{W} = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix}$$

Challenge:

Trying to predict $Y_{NT}(0)$ based on observed values $Y_{it}(0)$ for $(i,t) \neq (N,T)$.

In empirical studies there is a wide range of values for

- N_0, the number of control units
- N_1, the number of treated units
- T_0, the number of pre-treatment periods
- T_1, the number of post-treatment periods

This is important for guiding choice of analyses.

1. Mariel Boatlift (Card,1990), $N_1 = 1, N_0 = 44, T_0 = 7, T_1 = 6$

2. Minimum wage (Card-Krueger 1994), $N_1 = 321, N_0 = 78, T_0 = 1, T_1 = 1$

3. California smoking (Abadie, Diamond, Hainmueller, 2010) $N_1 = 1, N_0 = 29, T_0 = 17, T_1 = 13$

4. German unification (Abadie, Diamond, Hainmueller, 2014) $N_1 = 1, N_0 = 16, T_0 = 30, T_1 = 14$

5. Lalonde (1986) $N_1 = 185, N_0 = 15992, T_0 = 2, T_1 = 1$

Three related literatures on causal inference for this setting:

1. causal literature with unconfoundedness / horizontal regression
2. synthetic control literature / vertical regression
3. difference-in-differences and factor models

Here: **doubly robust** methods that combine **weighting** and **outcome modeling**

Unconfoundedness Methods / Horizontal Regression

Typical setting: N_0 and N_1 large, T_0 modest, $T_1 = 1$.

$$\mathbf{W} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Linear Model

$$\hat{\tau}_{\text{UNC}} = \frac{1}{N_1} \sum_{i:W_{iT}=1} \left(Y_{iT}(1) - \hat{Y}_{iT}(0) \right)$$

where

$$\hat{Y}_{iT}(0) = \hat{\alpha} + \sum_{t=1}^{T-1} \hat{\lambda}_t Y_{it}$$

and $\hat{\alpha}$ and $\hat{\lambda}$ are estimated by least squares:

$$\min_{\alpha,\lambda} \sum_{i=1}^{N_0-1} \left(Y_{iT} - \alpha - \sum_{t=1}^{T-1} \lambda_t Y_{it} \right)^2 \quad \text{"horizontal" regression}$$

Note: regression with N_0 observations, and T_0 regressors. May need regularization if T_0 is big.

Abadie-Diamond-Hainmueller Synthetic Control Method

Typical setting: T_0 and T_1 modest, N_0 small, $N_1 = 1$.

$$\mathbf{W} = \begin{pmatrix} 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 1 & \dots & 1 \end{pmatrix}$$

For simplicity focus on case with $T_1 = 1$, $T_0 = T - 1$.

$$\hat{\tau}_{\text{DI}} = Y_{NT} - \hat{Y}_{NT}(0), \qquad \hat{Y}_{NT}(0) = \alpha + \sum_{i=1}^{N-1} \omega_i Y_{iT}$$

where

$$\min_{\alpha,\omega} \sum_{t=1}^{T-1} \left(Y_{Nt} - \alpha - \sum_{i=1}^{N-1} \omega_i Y_{it} \right)^2 \qquad \text{"vertical" regression}$$

Note: regression with T_0 observations, and N_0 regressors.

Comparison Unconfoundedness vs Synthetic Controls in Case with $N_1 = T_1 = 1$

- Unconfoundedness req. $N_0 > T_0 \Longrightarrow$ horizontal regression

- Synthetic Control requires $N_0 < T_0 \Longrightarrow$ vertical regression

But, with **regularization** on regression coefficients we can use either unconfoundedness or synthetic control methods, irrespective of relative magnitude of N_0 and T_0.

Difference-In-Differences / Factor Models

Model $Y_{it}(0)$:

$$Y_{it}(0) = \alpha_i + \gamma_t + \varepsilon_{it}$$

leading to

$$\min_{\alpha,\gamma} \sum_{i=1}^{N} \sum_{t=1}^{T} (1 - W_{it}) \left(Y_{it} - \gamma_t - \alpha_i\right)^2$$

$$\hat{\tau} = \frac{1}{N_1 T_1} \sum_{i=N_0+1}^{N} \sum_{t=T_0+1}^{T} Y_{it} - \frac{1}{N_1 T_0} \sum_{i=N_0+1}^{N} \sum_{t=1}^{T_0} Y_{it}$$

$$- \left(\frac{1}{N_0 T_1} \sum_{i=1}^{N_0} \sum_{t=T_0+1}^{T} Y_{it} - \frac{1}{N_0 T_0} \sum_{i=1}^{N_0} \sum_{t=1}^{T_0} Y_{it} \right)$$

More general, factor models:

$$Y_{it}(0) = \sum_{r=1}^{R} \gamma_{tr}\alpha_{ir} + \varepsilon_{it}$$

(Athey, Bayati, Doudchenko, Imbens, Khosravi, 2018)

$$\arg\min_{\alpha,\gamma,\mathbf{L}} \sum_{i=1}^{N} \sum_{t=1}^{T} (1 - W_{it}) \left(Y_{it} - \alpha_i - \gamma_t - L_{it}\right)^2 + \lambda\|\mathbf{L}\|$$

with **nuclear normal** regularization on $\mathbf{L}$ to lead to low rank solution.

- Challenge: How to choose between these methods (vertical/horizontal regression, factor models), or how to tie them together?

- Relative merits of these methods

Comparison of

1. unconfoundedness (horizontal) regression with elastic net regularization (EN-H)

2. synthetic control (vertical) regression with elastic net regularization and no restrictions (EN-V)

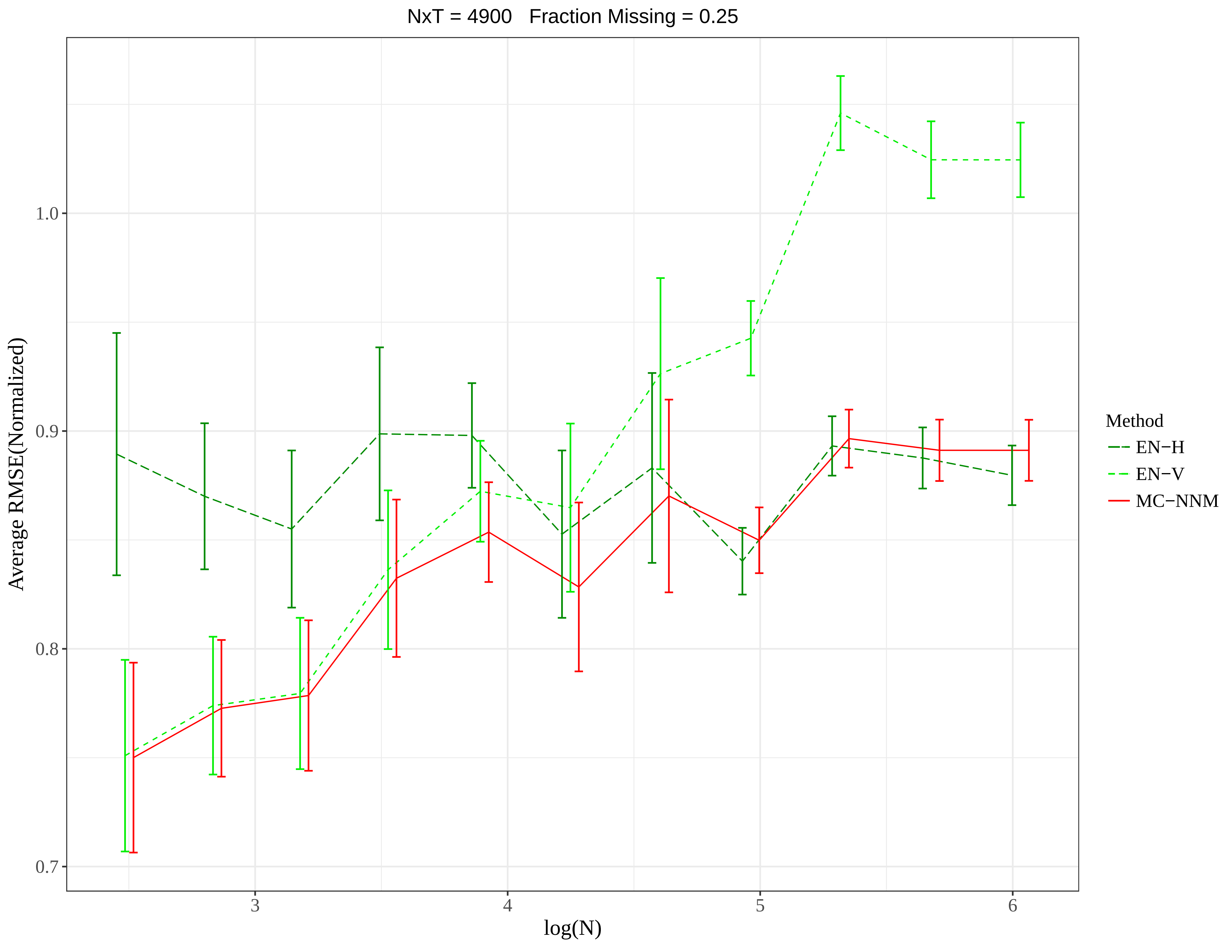
3. matrix completion with nuclear normal (MC-NNM)

Illustration: Stock Market Data

We use daily returns for 2453 stocks over 10 years (3082 days). We create sub-samples by looking at the first T daily returns of N randomly sampled stocks for pairs of (N, T) such that $N \times T = 4900$, ranging from fat to thin: $(N, T) = (10, 490), \ldots, (70, 70), \ldots, (490, 10)$.

Given the sample, we pretend that half the stocks are treated at the mid point over time, so that 25% of the entries in the matrix are missing.

$$\mathbf{Y}_{N\times T} = \begin{pmatrix} \checkmark & \checkmark & \checkmark & \checkmark & \ldots & \checkmark \\ \checkmark & \checkmark & \checkmark & \checkmark & \ldots & \checkmark \\ \checkmark & \checkmark & \checkmark & \checkmark & \ldots & \checkmark \\ \checkmark & \checkmark & \checkmark & ? & \ldots & ? \\ \checkmark & \checkmark & \checkmark & ? & \ldots & ? \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \checkmark & \checkmark & \checkmark & ? & \ldots & ? \end{pmatrix}$$


NxT = 4900 Fraction Missing = 0.25
Average RMSE(Normalized)
log(N)
1.0
0.9
0.8
0.7
3
4
5
6
Method
EN−H
EN−V
MC−NNM

Results

- MC-NNM does better than EN-H and EN-V, adapts to shape of matrix

- ADH restrictions (non-negativity of weights, and summing to one, and no intercept) sometimes improve things relative to Elastic-Net estimator, more so for the vertical regressions than for the horizontal regressions.

Combining Synthetic Control Methods and Matrix Completion: Observation I

Synthetic Control is weighted linear regression without unit fixed effects:

$$\hat{\tau}^{\text{ADH}} = \arg\min_{\tau,\gamma} \sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{it} - \gamma_t - \tau W_{it})^2 \times \omega_i^{\text{ADH}}$$

• regression with time fixed effects and ADH weights (easy to include covariates).

• under some conditions standard errors can be based on regression interpretation taking weights as given (even though the weights depend on outcome data).

Combining Synthetic Control Methods and Matrix Completion: Observation II

DID is unweighted regression with unit and time fixed effects:

$$\hat{\tau}^{\text{DID}} = \arg\min_{\tau,\gamma,\alpha} \sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{it} - \gamma_t - \alpha_i - \tau W_{it})^2$$

• regression with time fixed effects and unit fixed effects, no weights.

Synthetic Difference In Differences

$$\hat{\tau}^{\text{SDID}} = \arg\min_{\tau,\gamma,\alpha} \sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{it} - \gamma_t - \alpha_i - \tau W_{it})^2 \times \omega_i^{\text{ADH}} \times \lambda_t^{\text{ADH}}$$

Regression with unit and time fixed effects, and with unit and time weights.
Time weights satisfy:

$$\lambda = \arg\min_{\lambda} \sum_{i=1}^{N-1} \left(Y_{iT} - \sum_{t=1}^{T-1} \lambda_t Y_{it} \right)^2 + \text{regularization term},$$

subject to

$$\lambda_t \geq 0, \quad \sum_{t=1}^{T-1} \lambda_t = 1.$$

(or down-weight observations from distant past.)

Generalization: Synthetic Factor Models (SFM)

$$\hat{\tau}^{\text{SFM}} =$$

$$\arg\min_{\mathbf{L},\alpha,\gamma,\tau} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(Y_{it} - \alpha_i - \gamma_t - L_{it} - \tau W_{it}\right)^2 \omega_i^{\text{ADH}} \lambda_t^{\text{ADH}}$$

$$+\lambda \|\mathbf{L}\|,$$

Double Robustness

• If a factor model holds, but the weights are good (e.g., ADH weights), SDID is consistent.

• If the DID model holds, but we use arbitrary weights, SDID is consistent.

Conditions for Asymptotic Normality

With more units than time periods and a small treated block, it suffices that:

1. The noise ε has negligible long-range autocorrelation.
2. The signal L has low effective rank.
3. You can balance the units' noiseless pre-treatment outcomes with synthetic control weights that are not too concentrated.
4. You can balance the time periods' noiseless control outcomes with time weights approximating the noise autoregression.

With $T_0 \ll N_0,\ N_1 \ll \sqrt{N_0},\ T_1 \ll \sqrt{T_0}$, it suffices that:

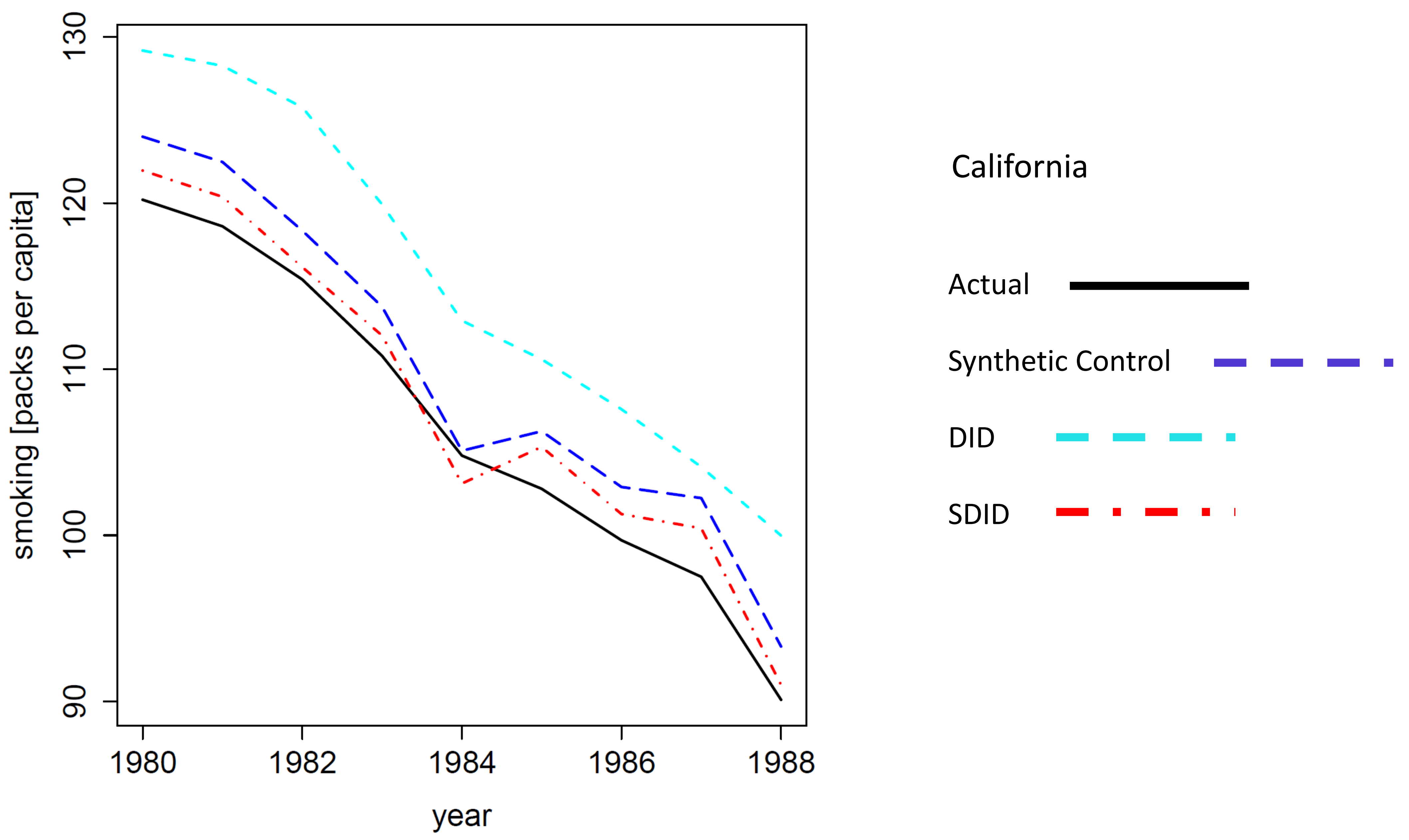
1. $\|\mathrm{Cor}(\varepsilon_i)\| = O(1)$
2. $\sigma_{\sqrt{k}}(L) \lesssim \sqrt{k},\ \ k = T_0, N_0$
3. $\|\omega_\star\| \lesssim T_0^{-1/4} N_1^{-1/2},\ \ N_1^{-1}$
4. $\|\lambda_\star - \psi\| \lesssim N_0^{-1/4} T_1^{-1/2},\ \ T_1^{-1}$

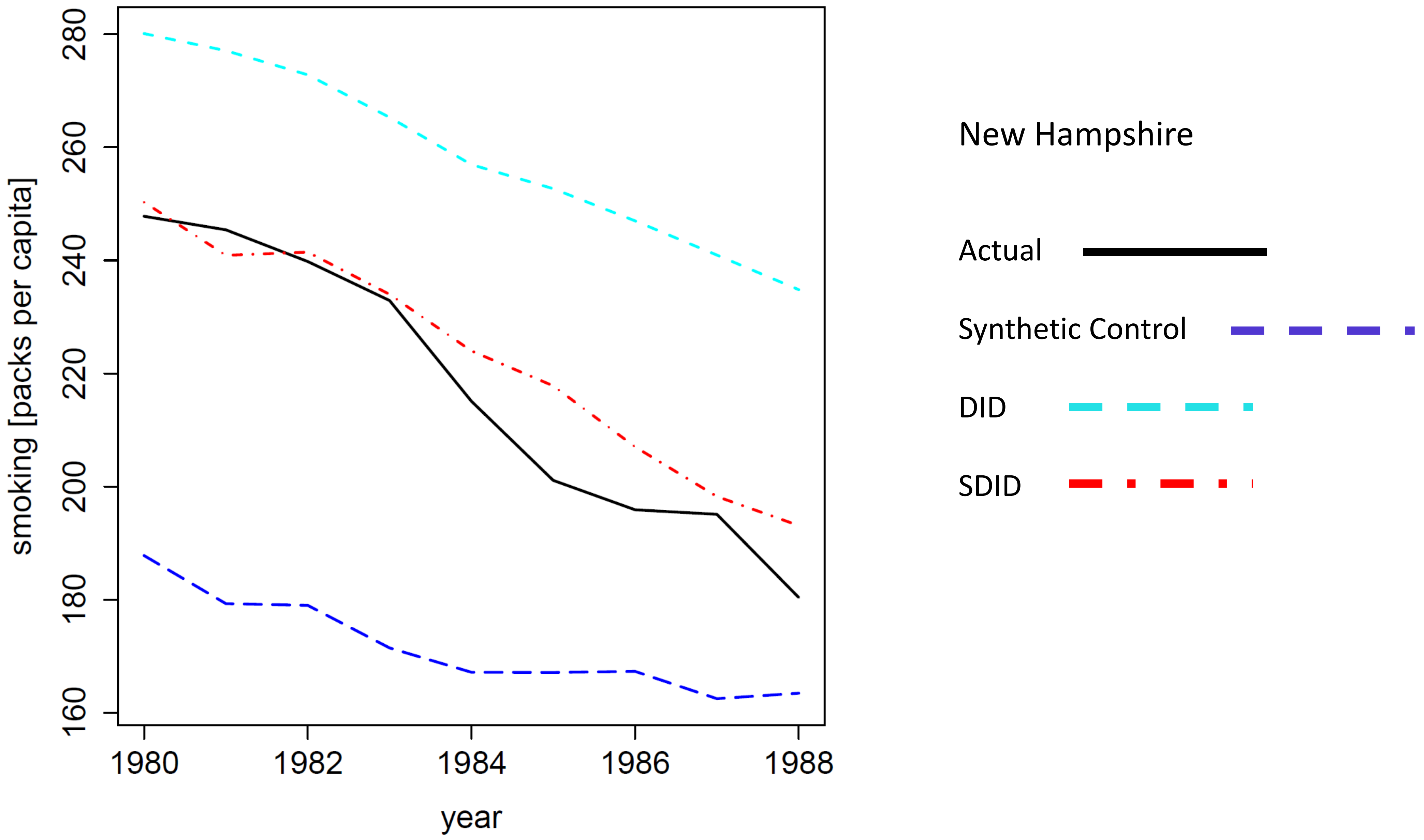
California smoking data calculations

Take pre-1988 data for all states, so we observe all $Y_{it}(0)$ for all unit/time pairs.

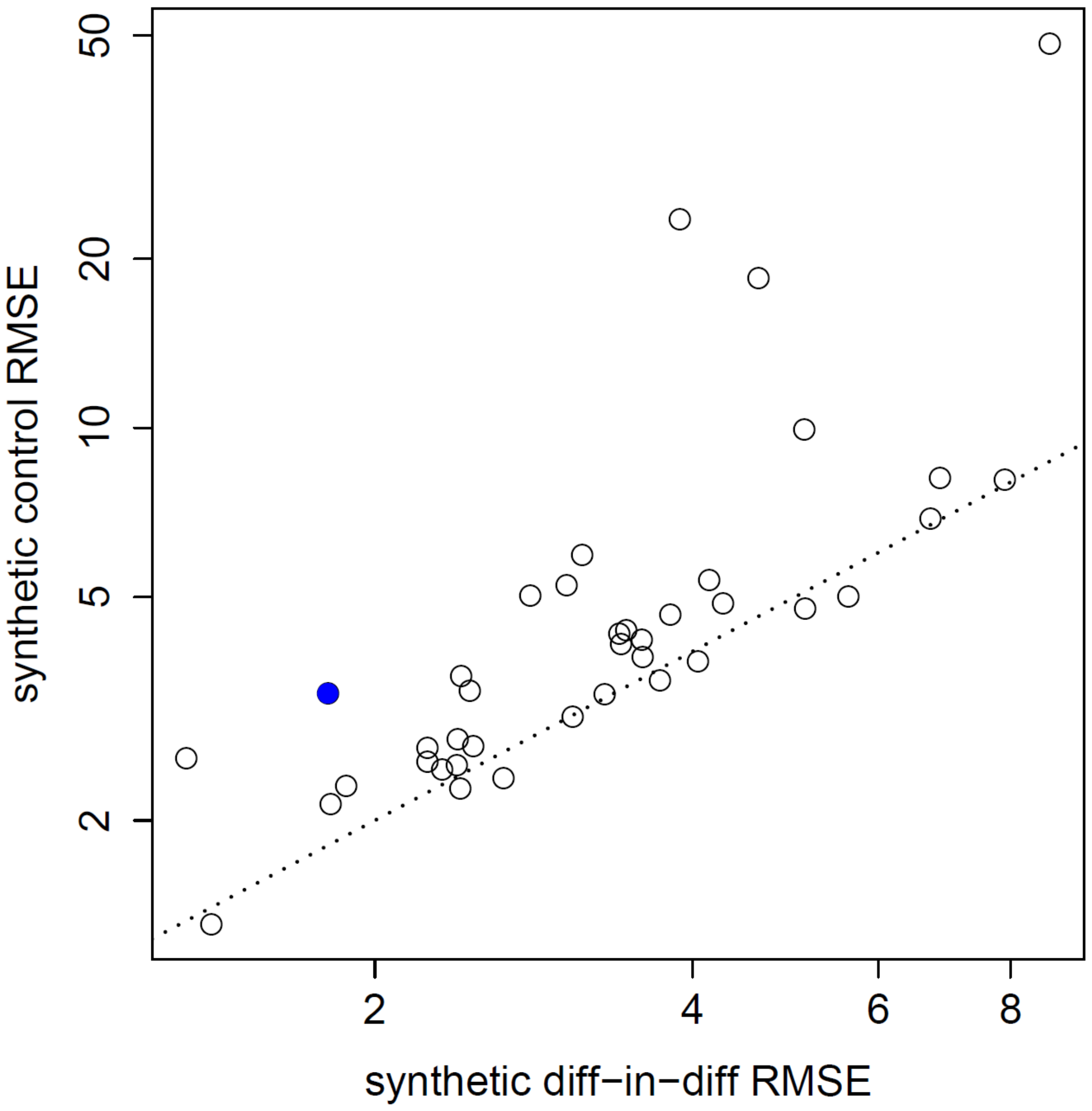
We **pretend** unit i was treated in periods $T_0+1,\ldots,T$, impute the "missing" values and compare them to actual values using SC (blue), DID (teal), SDID (red).

We average squared error by state for 8 periods ($T-T_0=8$) to get RMSEs for each state.


California
Actual
Synthetic Control
DID
SDID
smoking [packs per capita]
130
120
110
100
90
1980
1982
1984
1986
1988
year


New Hampshire
Actual
Synthetic Control
DID
SDID
smoking [packs per capita]
year
160
180
200
220
240
260
280
1980
1982
1984
1986
1988

Synthetic DID does better than SC across states



Replicating Bertrand, Duflo, Mullainathan

- CPS data.
- Log wages by state and year
- 51 states, 21 years
- Pseudo experiments:
  - Randomly select 25 “treated” states
  - Randomly select initial treatment period
- SDID has Lowest RMSE

	RMSE
DID	.0173
GLS (rho=.5)	.0149
SC	.0161
SDID	.0142

R Packages available from Abadie-Diamond-Hainmueller, Athey-Bayati-Doudchenko-Imbens-Khosravi, Arkhangelsky-Athey-Hirshberg-Imbens-Wager.