

Recent Developments in Synthetic Controls

Regularization, Prediction Intervals, and Multiple Treated Units

Tsung-Chih Lai

Department of Economics
National Chung Cheng University

October 17, 2024

Synthetic Control Method

“Arguably the most important innovation in the policy evaluation literature in the last 15 years.”

— Athey & Imbens (2017, JEP)

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

— The Washington Post

Introduction

Pioneered by Abadie & Gardeazabal (2003, AER), the **synthetic control (SC)** method offers a way to study the causal effect of a treatment on a single or a few units.

Popular in comparative case studies:

- German reunification (Abadie, Diamond & Hainmueller, 2015, AJPS).
- California tobacco control program (Abadie, Diamond & Hainmueller, 2010, JASA).
- Economic liberalization (Cattaneo, Feng, Palomba & Titiunik, 2023).

Single Treated Unit

	Pre-treatment				Post-treatment			
Treated	X	X	...	X	\checkmark	\checkmark	...	\checkmark
	X	X	...	X	X	X	...	X
Control	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	X	X	...	X	X	X	...	X

X : outcome without treatment \checkmark : outcome with treatment

- Idea: use a weighted average of control units to predict the **counterfactual outcome** that the treated unit would have experienced without treatment.

Multiple Treated Units

	Pre-treatment				Post-treatment			
Early adopter	X	X	...	X	\checkmark	\checkmark	...	\checkmark
	\vdots	\vdots		\vdots	X	\checkmark		\checkmark
	\vdots	\vdots		\vdots	\vdots	\vdots		\vdots
Late adopter	X	X	...	X	X	X	...	\checkmark
	X	X	...	X	X	X	...	X
	\vdots	\vdots		\vdots	\vdots	\vdots		\vdots
Never treated	X	X	...	X	X	X	...	X
	\vdots	\vdots		\vdots	\vdots	\vdots		\vdots
	X	X	...	X	X	X	...	X

X : outcome without treatment \checkmark : outcome with treatment

- Simultaneous or **staggered adoption**.

Basic Setup

Let's focus on the canonical single treated unit case.

Suppose we have an $N \times T$ panel data, where the observed outcome Y_{it} for $i = 1, \dots, N$ and $t = 1, \dots, T$ is

$$Y_{it} = \begin{cases} Y_{it}(0) & \text{if } i = 1 \text{ and } t = 1, \dots, T_0 \\ Y_{it}(1) & \text{if } i = 1 \text{ and } t = T_0 + 1, \dots, T \\ Y_{it}(0) & \text{if } i = 2, \dots, N \end{cases}$$

with two **potential outcomes** $Y_{it}(0)$ and $Y_{it}(1)$.

We are interested in the treatment effect on the treated for $t > T_0$:

$$\begin{aligned} \tau_t &\equiv Y_{1t}(1) - Y_{1t}(0) \\ &= Y_{1t} - \mathbf{Y}_{1t}(0) \end{aligned}$$

Canonical SC

Canonical SC assumes that there exists a vector of weights $\mathbf{w} \equiv (w_2, \dots, w_N)'$ such that for all $t = 1, \dots, T$

$$Y_{1t}(0) \approx \sum_{i=2}^N w_i Y_{it}$$

By matching the treated and control units' pre-treatment outcomes

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w} \in \mathcal{W}} \sum_{t=1}^{T_0} \left(Y_{1t} - \sum_{i=2}^N w_i Y_{it} \right)^2$$
$$\mathcal{W} = \{ \mathbf{w} \in \mathbb{R}_+^{N-1} : \|\mathbf{w}\|_1 = 1 \}$$

the SC estimator of $Y_{1t}(0)$ for $t > T_0$ is given by

$$\hat{Y}_{1t}(0) = \sum_{i=2}^N \hat{w}_i Y_{it}$$

Implementation

In words, implementing the SC method involves two steps:

- (i) The treated unit is **matched** to control units using only their pre-treatment outcomes via constrained regressions.
- (ii) The counterfactual outcome of the treated unit is **predicted** by applying the pre-treatment matching weights to the post-treatment outcomes of the control units.

For demonstration, we use the `scpi` software package provided by Cattaneo, Feng, Palomba & Titiunik (2024), which is available in Python, R, and Stata.

Package Comparison

Package name	Statistical platform	Prediction method	Inference method	Multiple treated	Staggered adoption	Misspecification robust	Automatic parallelization	Last update
ArCo	R	LA	Asym			✓		2017-11-05
pgsc	R	SC	Perm	✓				2018-10-28
npsynth	St	SC	Perm					2020-06-23
tidysynth	R	SC	Perm					2021-01-27
scinference	R	SC, LA	Perm			✓		2021-05-13
gsynth	R	FA	Asym	✓	✓		✓	2021-08-06
Synth	Py	SC	Perm					2021-10-07
treebased-sc	Py	TB	Perm			✓		2021-11-01
sytnhdid	R	LS, RI	Asym	✓	✓			2022-03-15
allsynth	St	SC	Perm	✓	✓			2022-06-22
SCtools	R	SC	Perm	✓			✓	2022-06-09
scul	St	LA	Perm	✓				2022-08-21
MSCMT	R	SC	Perm				✓	2023-04-17
Synth	R, St	SC	Perm					2023-06-02
microsynth	R	CA	Perm	✓			✓	2023-06-30
augsynth	R	SC, RI	Perm	✓	✓			2023-09-21
synth2	St	SC	Perm					2023-10-05
SCUL	R	LA	Perm					2023-10-10
scpi	Py, R, St	SC, LA, RI, LS, +	PI, Asym, Perm	✓	✓	✓	✓	2023-11-01

Table 1: Comparison of different packages available on PyPi, CRAN, REPEC, or GitHub. Py = Python (<https://www.python.org/>); R = R (<https://cran.r-project.org/>); St = Stata (<https://www.stata.com/>); LA = Lasso penalty; CA = calibration; FA = factor-augmented models; LS = unconstrained least squares; RI = Ridge penalty; SC = canonical synthetic control; TB = tree-based methods; + = user-specified options (see Table 3 below for more details); Perm = permutation-based inference; Asym = asymptotic-based inference; PI = prediction intervals (non-asymptotic probability guarantees). The symbol ✓ means that the feature is available. The last column reports the date of last update as of November 1, 2023.

Example 1: German Reunification

Using the SC method, Abadie, Diamond & Hainmueller (2015, AJPS) study the economic consequences of German reunification.

- Treatment: German reunification in 1990.
- Outcome variable: GDP per capita (in 2002 USD).
- Treated unit: West Germany.
- Control units: 16 OECD member countries.
- Pre-treatment period: 1960–1990.
- Post-treatment period: 1991–2003.

Replication data is provided in the `scpi` package.

Summary Statistics

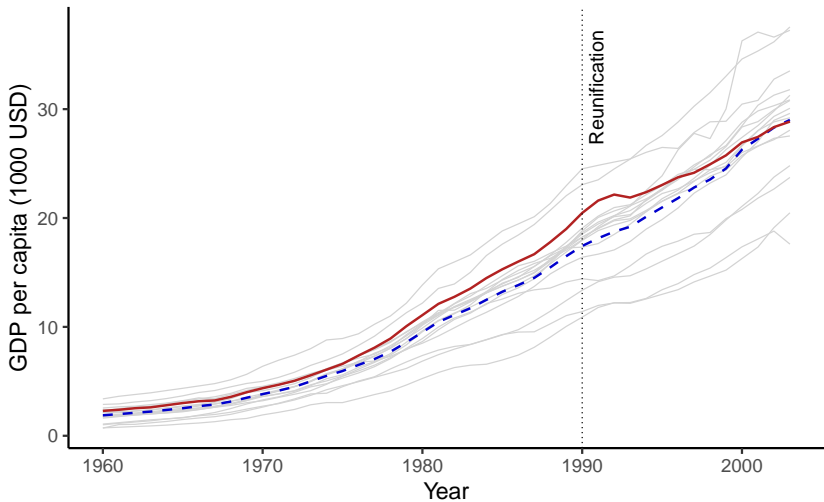
```
# Install & load packages
install.packages("scpi")
library(scpi)
library(dplyr)
library(ggplot2)

# Load data
data <- scpi_germany
summary(data)
```

index	country	year	gdp
Min. : 1.00	Length:748	Min. :1960	Min. : 0.707
1st Qu.: 5.00	Class :character	1st Qu.:1971	1st Qu.: 3.985
Median : 9.00	Mode :character	Median :1982	Median :10.258
Mean :10.29		Mean :1982	Mean :12.144
3rd Qu.:16.00		3rd Qu.:1992	3rd Qu.:18.878
Max. :21.00		Max. :2003	Max. :37.548

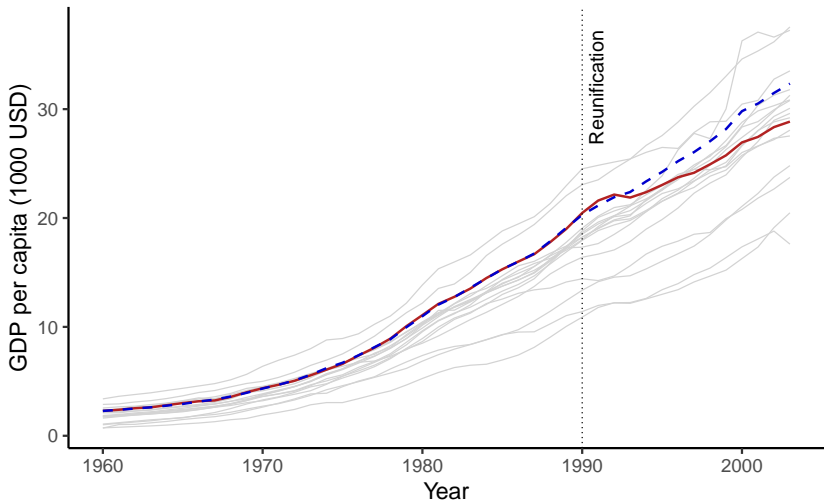
infrate	trade	schooling	industry
Min. : -0.9151	Min. : 9.429	Min. : 3.50	Min. :21.59
1st Qu.: 2.4683	1st Qu.: 33.843	1st Qu.:26.40	1st Qu.:29.35
Median : 4.0800	Median : 49.530	Median :38.00	Median :33.07
Mean : 5.8677	Mean : 53.124	Mean :36.36	Mean :33.24
3rd Qu.: 7.5101	3rd Qu.: 68.511	3rd Qu.:46.90	3rd Qu.:36.38
Max. :28.7833	Max. :149.682	Max. :69.60	Max. :48.00
NA's :21	NA's :102	NA's :597	NA's :207

Time Series Plot



— West Germany — Average of other OECD countries — Other OECD countries

(Synthetic) West Germany



— West Germany — Synthetic West Germany — Other OECD countries

Data Preparation

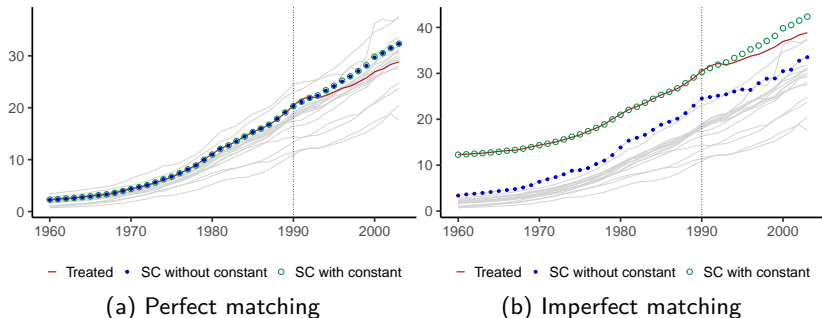
```
df <- scdata(  
  df = data,  
  id.var = "country",  
  time.var = "year",  
  outcome.var = "gdp",  
  unit.tr = "West Germany",  
  unit.co = setdiff(data$country,  
                    "West Germany"),  
  period.pre = 1960:1990,  
  period.post = 1991:2003,  
  
  features = "gdp",  
  constant = TRUE,  
)  
  
summary(df)
```

Synthetic Control - Setup

Treated Unit:	West Germany
Size of the donor pool:	16
Features:	1
Pre-treatment period:	1960 1990
Post-treatment period:	1991 2003
Pre-treatment periods used in estimation:	31
Covariates used for adjustment:	1

- features: features (predictor variables) to be matched.
- constant: allow for a constant shift in the counterfactual outcome of the treated unit, i.e., $Y_{1t}(0) = \alpha + \sum_{i=2}^N w_i Y_{it}$.

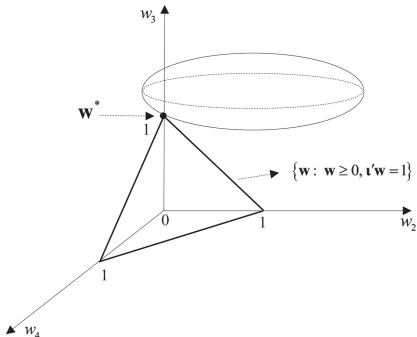
(Im)Perfect Matching



- Under perfect matching, both methods work well.
- Under imperfect matching, only SC with constant is effective.
- See Ferman & Pinto (2021, QE) for more details.

SC Estimation

```
res <- scest(
  data = df,
  w.constr = list(name = "simplex"),
  # w.constr = list(name = "ols"),
  # w.constr = list(name = "lasso"),
  # w.constr = list(name = "ridge"),
  # w.constr = list(name = "L1-L2"),
)
summary(res)
```



Synthetic Control Prediction - Results

Active donors: 6

Coefficients:

	Weights
Australia	0.000
Austria	0.441
Belgium	0.000
Denmark	0.000
France	0.000
Greece	0.000
Italy	0.177
Japan	0.014
Netherlands	0.058
New Zealand	0.000
Norway	0.000
Portugal	0.000
Spain	0.000
Switzerland	0.036
UK	0.000
USA	0.274

Covariates

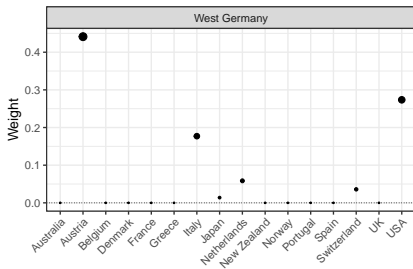
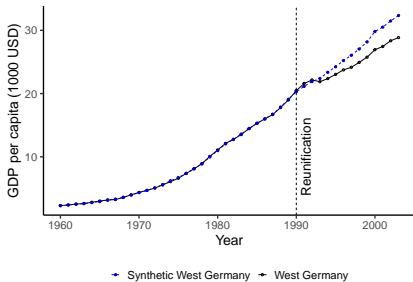
West Germany.0.constant 0.158

SC Plots

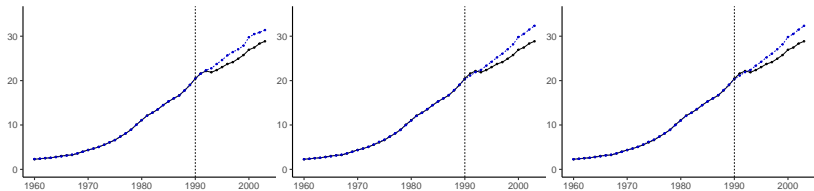
```
plot <- splot(
  result = res,
  col.synth = "mediumblue",
  col.treated = "black",
  label.xy = list(
    x.lab = "Year",
    y.lab = "GDP per capita (1000 USD)",
    x.ticks = seq(1960, 2000, by = 10),
    event.label = list(
      lab = "Reunification", height = 10),
  )
```

```
plot$plot_out +
  ggtitle(NULL) +
  scale_color_manual(
    values = c("mediumblue", "black"),
    labels = c("Synthetic West Germany",
              "West Germany"))
```

```
plot_wt <- coef(res)
plot_wt +
  theme(axis.text.x = element_text(
    angle = 45, vjust = 1, hjust = 1))
```



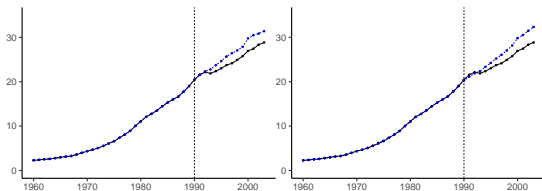
Different Constraints on the Weights



(a) OLS (no constraint)

(b) Simplex

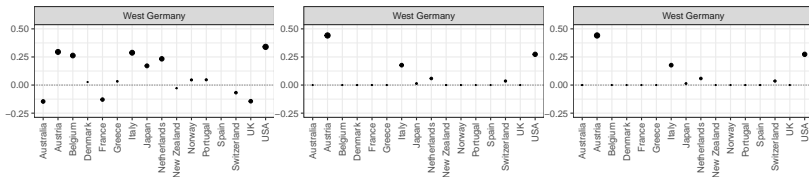
(c) Lasso



(d) Ridge

(e) L1-L2

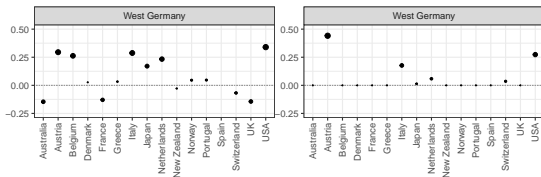
Different Constraints on the Weights



(a) OLS (no constraint)

(b) Simplex

(c) Lasso



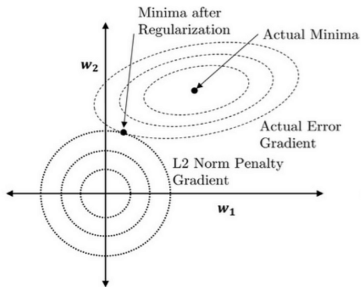
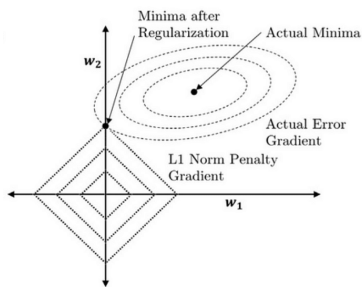
(d) Ridge

(e) L1-L2

Lasso (L1) and Ridge (L2) Regularization

Name	w.constr	Form of \mathcal{W}
OLS	<code>list(name = 'ols')</code>	\mathbb{R}^J
Simplex	<code>list(name = 'simplex', Q = Q)</code>	$\{\mathbf{w} \in \mathbb{R}_+^J : \ \mathbf{w}\ _1 = Q\}$
Lasso	<code>list(name = 'lasso', Q = Q)</code>	$\{\mathbf{w} \in \mathbb{R}^J : \ \mathbf{w}\ _1 \leq Q\}$
Ridge	<code>list(name = 'ridge', Q = Q)</code>	$\{\mathbf{w} \in \mathbb{R}^J : \ \mathbf{w}\ _2 \leq Q\}$
L1-L2	<code>list(name = 'L1-L2', Q = Q, Q2 = Q2)</code>	$\{\mathbf{w} \in \mathbb{R}_+^J : \ \mathbf{w}\ _1 = Q, \ \mathbf{w}\ _2 \leq Q_2\}$

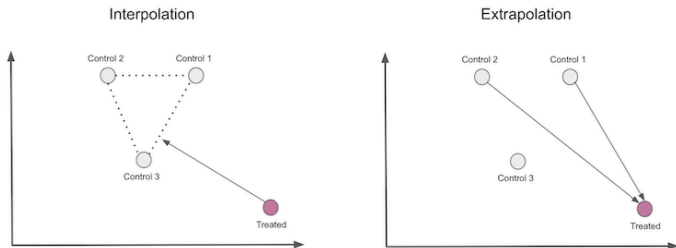
Table 2: Constraints on the weights directly implemented in **scpi**.



Interpolation and Extrapolation

Article	\mathcal{W}	\mathcal{R}	w. constr			constant
			name	Q	Q2	
Hsiao <i>et al.</i> (2012)	\mathbb{R}^J	\mathbb{R}	"ols"	NULL	NULL	TRUE
Abadie <i>et al.</i> (2010)	$\{\mathbf{w} \in \mathbb{R}_+^J : \ \mathbf{w}\ _1 = 1\}$	$\{0\}$	"simplex"	1	NULL	FALSE
Ferman and Pinto (2021)	$\{\mathbf{w} \in \mathbb{R}_+^J : \ \mathbf{w}\ _1 = 1\}$	\mathbb{R}	"simplex"	1	NULL	TRUE
Chernozhukov <i>et al.</i> (2021)	$\{\mathbf{w} \in \mathbb{R}_+^J : \ \mathbf{w}\ _1 \leq 1\}$	\mathbb{R}	"lasso"	1	NULL	TRUE
Amjad <i>et al.</i> (2018)	$\{\mathbf{w} \in \mathbb{R}_+^J : \ \mathbf{w}\ _2 \leq Q\}$	$\{0\}$	"ridge"	Q	NULL	FALSE
Arkhangelsky <i>et al.</i> (2021)	$\{\mathbf{w} \in \mathbb{R}_+^J : \ \mathbf{w}\ _1 = 1, \ \mathbf{w}\ _2 \leq Q_2\}$	\mathbb{R}	"L1-L2"	1	Q	TRUE

Table 3: Examples of \mathcal{W} and \mathcal{R} in the synthetic control literature ($M = 1$).



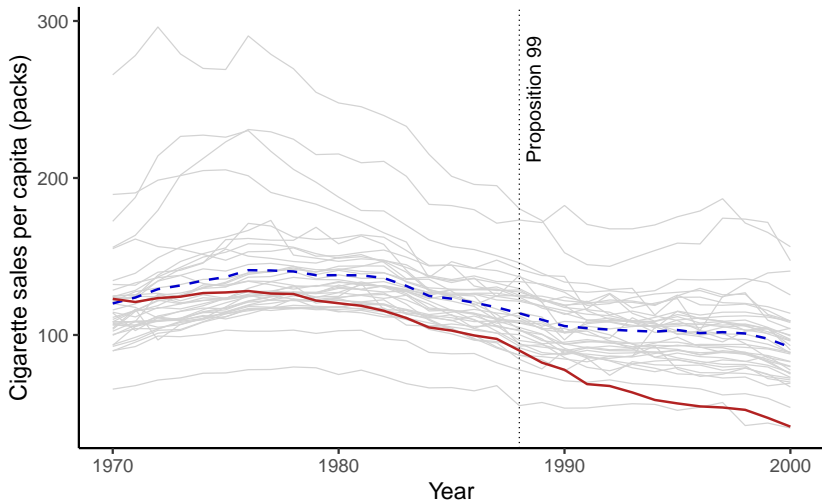
Example 2: California's Proposition 99

Another seminal study by Abadie, Diamond & Hainmueller (2010, JASA) evaluates the effectiveness of tobacco control in California.

- Treatment: Proposition 99, a large-scale tobacco control program adopted in California in 1988.
- Outcome variable: cigarette sales per capita (in packs).
- Treated unit: California.
- Control units: 38 states without similar programs.
- Pre-treatment period: 1970–1988.
- Post-treatment period: 1989–2000.

Replication data is provided in the `synthdid` package.

Time Series Plot



— California — Average of other states — Other states

Data Preparation

```
df <- sdata(  
  df = data,  
  id.var = "State",  
  time.var = "Year",  
  outcome.var = "PacksPerCapita",  
  unit.tr = "California",  
  unit.co = setdiff(data$State,  
                    "California"),  
  period.pre = 1970:1988,  
  period.post = 1989:2000,  
  features = "PacksPerCapita",  
  constant = FALSE,  
  
  cointegrated.data = TRUE,  
)  
  
summary(df)
```

Synthetic Control - Setup

Treated Unit:	California
Size of the donor pool:	38
Features:	1
Pre-treatment period:	1970 1988
Post-treatment period:	1989 2000
Pre-treatment periods used in estimation:	19
Covariates used for adjustment:	0

- `cointegrated.data`: whether the outcomes from the treated and control units form a cointegrated system, i.e., $Y_{it} \sim I(1)$ for all i and $Y_{1t} - \sum_{i \geq 2} w_i Y_{it} \sim I(0)$.
- It only affects the inference procedure as explained below.

SC Estimation (Inference?)

```

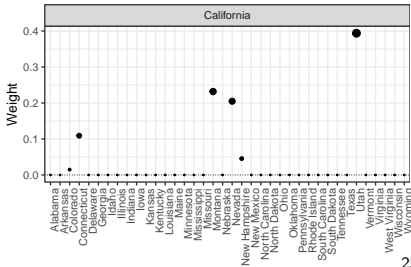
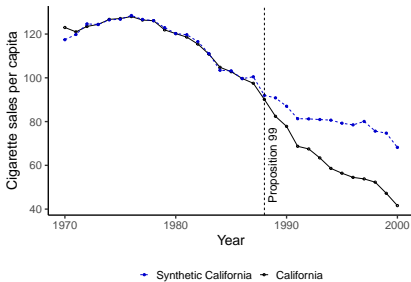
res_est <- scest(
  data = df,
  w.constr = list(name = "simplex"),
  # w.constr = list(name = "ols"),
  # w.constr = list(name = "lasso"),
  # w.constr = list(name = "ridge"),
  # w.constr = list(name = "L1-L2"),
)

plot_est <- scplot(
  result = res_est,
  col.synth = "mediumblue",
  col.treated = "black",
  label.xy = list(
    x.lab = "Year",
    y.lab = "Cigarette sales per capita"),
  x.ticks = seq(1970, 2000, by = 10),
  event.label = list(
    lab = "Proposition 99", height = 60),
)

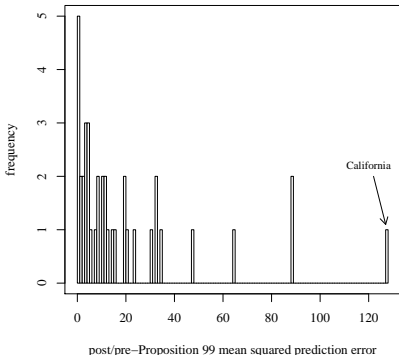
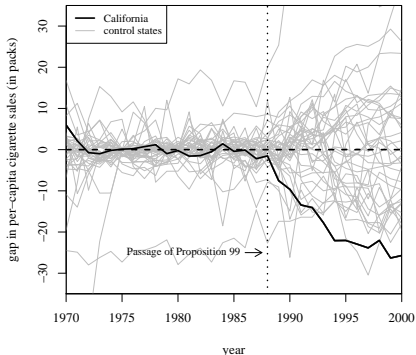
plot_est$plot_out +
  ggtitle(NULL) +
  scale_color_manual(
    values = c("mediumblue", "black"),
    labels = c("Synthetic California",
              "California"))

coef(res_est)

```

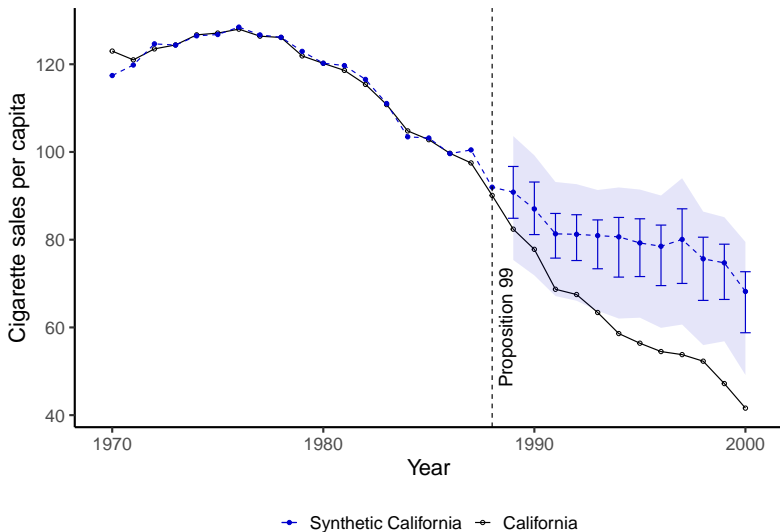


Permutation-Based Inference



- Proposed by Abadie, Diamond & Hainmueller (2010, JASA), also known as the **placebo test**.

Pointwise and Simultaneous Prediction Intervals



- Developed by Cattaneo, Feng & Titiunik (2021, JASA).

Two Sources of Uncertainty

Recall that the quantity of interest is

$$\tau_t = Y_{1t} - Y_{1t}(0) \quad (1)$$

- It is better treated as **random** rather than a fixed parameter.

Canonical SC assumes that for the pre-treatment period $t \leq T_0$, there exists a set of pseudo-true weights $\{w_i^0 : i \geq 2\}$ such that

$$Y_{1t}(0) = \sum_{i \geq 2} w_i^0 Y_{it} + u_t \quad (2)$$

Under correct specification (say $u_t = 0$), $\{w_i^0\}$ can be effectively estimated by the pre-treatment matching weights $\{\hat{w}_i\}$, yielding the SC estimator of (1):

$$\hat{\tau}_t = Y_{1t} - \sum_{i \geq 2} \hat{w}_i Y_{it} \quad (3)$$

Two Sources of Uncertainty

Similar to (2), we may assume that for $t > T_0$

$$Y_{1t}(0) = \sum_{i \geq 2} w_i^0 Y_{it} + e_t \quad (4)$$

where e_t represents misspecification error **plus** any additional noise occurring at the post-treatment period.

Using (1), (3) and (4), it follows that for $t > T_0$

$$\begin{aligned} \hat{\tau}_t - \tau_t &= Y_{1t}(0) - \sum_{i \geq 2} \hat{w}_i Y_{it} \\ &= \underbrace{- \sum_{i \geq 2} (\hat{w}_i - w_i^0) Y_{it}}_{\text{in-sample uncertainty}} + \underbrace{e_t}_{\text{out-of-sample uncertainty}} \end{aligned}$$

In-Sample Uncertainty: Stationary Case

Let $\mathbf{x}_t \equiv (Y_{2t}, \dots, Y_{Nt})'$ and $\mathcal{H} \equiv \{\mathbf{x}_t : t \geq 1\}$ the conditioning set.

For the stationary case:

$$\sqrt{T_0}(\widehat{\mathbf{w}} - \mathbf{w}_0) = \arg \min_{\boldsymbol{\delta} \in \Delta} \underbrace{\boldsymbol{\delta}' \left(\frac{1}{T_0} \sum_{t=1}^{T_0} \mathbf{x}_t \mathbf{x}_t' \right)}_{\widehat{\mathbf{Q}}} \boldsymbol{\delta} - 2 \underbrace{\left(\frac{1}{\sqrt{T_0}} \sum_{t=1}^{T_0} \mathbf{x}_t' u_t \right)}_{\widehat{\boldsymbol{\gamma}}} \boldsymbol{\delta}$$

- $\widehat{\mathbf{Q}}$ is fixed conditional on \mathcal{H} .
- Approximate $\widehat{\boldsymbol{\gamma}}$ (possibly misspecified) by Berry-Esseen bound:

$$\mathbb{P}(\widehat{\boldsymbol{\gamma}} - \mathbb{E}(\widehat{\boldsymbol{\gamma}}|\mathcal{H}) \in \mathcal{G}) \approx \mathbb{P}(\mathbf{G} \in \mathcal{G})$$

where $\mathbf{G} \sim \mathcal{N}(0, \mathbb{V}(\widehat{\boldsymbol{\gamma}}|\mathcal{H}))$.

In-Sample Uncertainty: Cointegrated Case

Let $\mathbf{x}_t \equiv (Y_{2t}, \dots, Y_{Nt})'$ and $\mathcal{H} \equiv \{\mathbf{x}_t : t \geq 1\}$ the conditioning set.

For the cointegrated case:

$$T_0(\widehat{\mathbf{w}} - \mathbf{w}_0) = \arg \min_{\delta \in \Delta} \underbrace{\delta' \left(\frac{1}{T_0^2} \sum_{t=1}^{T_0} \mathbf{x}_t \mathbf{x}_t' \right) \delta}_{\widehat{\mathbf{Q}}} - 2 \underbrace{\left(\frac{1}{T_0} \sum_{t=1}^{T_0} \mathbf{x}_t' u_t \right) \delta}_{\widehat{\gamma}}$$

- $\widehat{\mathbf{Q}}$ is fixed conditional on \mathcal{H} .
- Approximate $\widehat{\gamma}$ (possibly misspecified) by Berry-Esseen bound:

$$\mathbb{P}(\widehat{\gamma} - \mathbb{E}(\widehat{\gamma}|\mathcal{H}) \in \mathcal{G}) \approx \mathbb{P}(\mathbf{G} \in \mathcal{G})$$

where $\mathbf{G} \sim \mathcal{N}(0, \mathbb{V}(\widehat{\gamma}|\mathcal{H}))$.

Out-of-Sample Uncertainty

Three approaches to quantify out-of-sample uncertainty $e_t|\mathcal{H}$:

(i) Sub-Gaussian bounds: suppose that $e_t|\mathcal{H}$ is sub-Gaussian

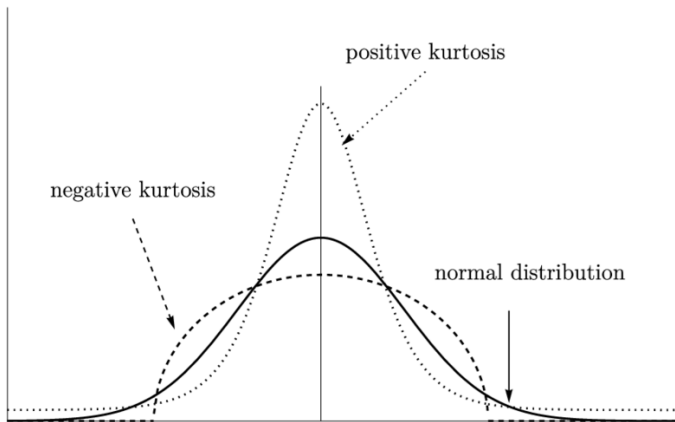
$$\mathbb{P}(|e_t - \mathbb{E}[e_t|\mathcal{H}]| \geq \varpi_e|\mathcal{H}) \leq 2 \exp\left(-\frac{\varpi_e^2}{2\sigma_{\mathcal{H}}^2}\right)$$

(ii) Location-scale model: assume that $\varepsilon_t \perp \mathcal{H}$ and

$$e_t = \mathbb{E}[u_t|\mathcal{H}] + (\mathbb{V}[u_t|\mathcal{H}])^{1/2}\varepsilon_t$$

(iii) Quantile regression: $\mathbb{Q}_{u_t|\mathcal{H}}(\alpha/2)$ and $\mathbb{Q}_{u_t|\mathcal{H}}(1 - \alpha/2)$.

Sub-Gaussianity



- If the (excess) kurtosis of $e_t|\mathcal{H}$ is negative, then $e_t|\mathcal{H}$ is also said to be sub-Gaussian.

Non-Asymptotic Conditional Prediction Intervals

Cattaneo, Feng & Titiunik (2021, JASA, Lemma 1)

Suppose that there exist $M_{1,L}, M_{1,U}, M_{2,L}, M_{2,U}$ satisfying

$$\mathbb{P} \left[M_{1,L} \leq - \sum_{i \geq 2} (\hat{w}_i - w_i^0) Y_{it} \leq M_{1,U} \middle| \mathcal{H} \right] \geq 1 - \alpha_1$$

$$\mathbb{P} \left[M_{2,L} \leq e_t \leq M_{2,U} \middle| \mathcal{H} \right] \geq 1 - \alpha_2$$

with high probability over \mathcal{H} . Then,

$$\mathbb{P} \left[\hat{\tau}_t - M_{1,U} - M_{2,U} \leq \tau_t \leq \hat{\tau}_t - M_{1,L} - M_{2,L} \middle| \mathcal{H} \right] \geq 1 - \alpha_1 - \alpha_2$$

with high probability over \mathcal{H} .

Simultaneous Prediction Intervals

For some $L \geq 1$, one can also construct a sequence of intervals that has high **simultaneous** coverage over multiple periods:

$$\mathbb{P}[\tau_t \in \mathcal{J}_t, \text{ for all } T_0 + 1 \leq t \leq T_0 + L | \mathcal{H}] \geq 1 - \alpha_1 - \alpha_2$$

with high probability over \mathcal{H} .

To quantify out-of-sample uncertainty in this case, the `scpi` package generalizes the sub-Gaussian method to

$$M_{2,L,t} = \mathbb{E}[e_t | \mathcal{H}] - \sqrt{2\sigma_{\mathcal{H}}^2 \log(2L/\alpha_2)}$$
$$M_{2,U,t} = \mathbb{E}[e_t | \mathcal{H}] + \sqrt{2\sigma_{\mathcal{H}}^2 \log(2L/\alpha_2)}$$

yielding wider simultaneous PIs than their pointwise counterparts.

SC Prediction Intervals

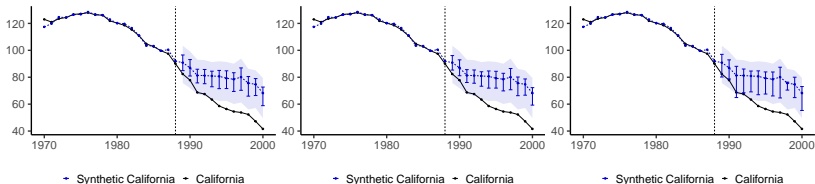
```
res_pi <- scpi(
  data = df,
  w.constr = list(name = "simplex"),

  e.method = "gaussian", # sub-Gaussian bounds
  # e.method = "ls",     # location-scale model
  # e.method = "qreg",   # quantile regression
  u.alpha = 0.05, # in-sample confidence level
  e.alpha = 0.05, # out-of-sample conf. level
)
```

```
plot_pi <- scplot(
  result = res_pi,
  label.xy = list(x.lab = NULL, y.lab = NULL),
  x.ticks = seq(1970, 2000, by = 10),

  joint = TRUE, # include simultaneous PI
)

plot_pi$plot_out +
  ...
```

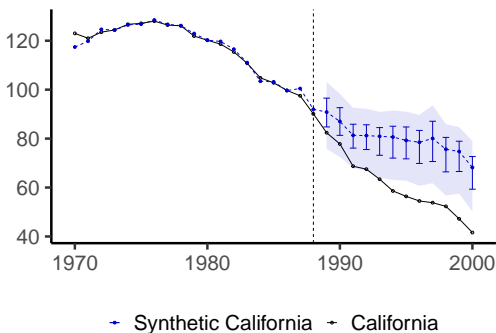


(a) Sub-Gaussian bounds (b) Location-scale model (c) Quantile regression

- These PIs are of $1 - 0.05 - 0.05 = 90\%$ coverage probability.

More Options

```
res_pi <- scpi(  
  data = df,  
  w.constr = list(name = "simplex"),  
  e.method = "gaussian",  
  u.alpha = 0.05,  
  e.alpha = 0.05,  
  
  u.misssp = TRUE, # misspecification  
  sims = 200,     # simulations  
  cores = 1,     # parallelization  
)  
  
plot_pi <- scplot(  
  ...  
  
plot_pi$plot_out +  
  ...
```



- `u.misssp`: whether the SC model is treated as misspecified.
- `sims`: number of simulations used to quantify in-sample error.
- `cores`: number of cores used for parallel computing.

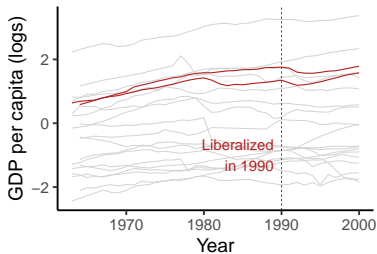
Example 3: Economic Liberalization

Cattaneo, Feng, Palomba & Titiunik (2023) study the effects of economic liberalization on GDP for emerging European countries.

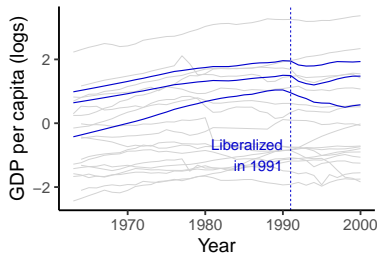
- Treatment: Economic liberalization episodes in the 1990s.
- Outcome variable: GDP per capita (in logs).
- Treated units: 7 emerging European countries.
- Control units: 18 countries worldwide without liberalization.
- Pre-treatment periods: 1963–1990, 1991, 1992.
- Post-treatment periods: 1990, 1991, 1992–2000.

Multiple treated units with simultaneous and **staggered adoption**.

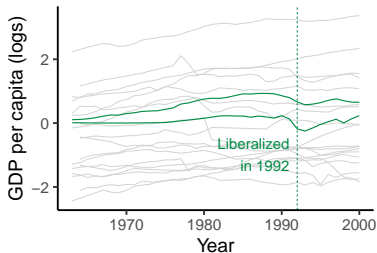
Time Series Plots



(a) Hungary and Poland



(b) Bulgaria, Czech and Slovak



(c) Albania and Romania

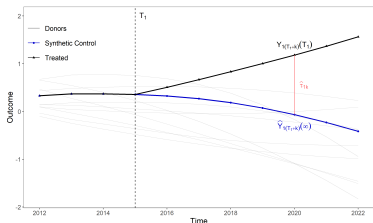
Multiple Treated Units Case

```
df <- sdataMulti(  
  df = data,  
  id.var = "countryname",  
  time.var = "year",  
  outcome.var = "lgdp",  
  treatment.var = "liberalization",  
  units.est = treated_units,  
  post.est = 10,  
  features = list("lgdp"),  
  constant = TRUE,  
  cointegrated.data = FALSE,  
  
  effect = "unit-time", # individual treatment effect  
  # effect = "unit",    # average post-treatment effect  
  # effect = "time",   # average treatment effect  
  # on the treated (ATT)  
)
```

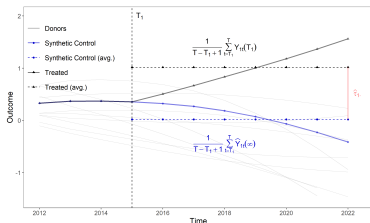
```
res <- scpi(  
  data = df,  
  w.constr = list(name = "L1-L2"),  
  e.method = "gaussian",  
  
  V = "separate", # separate fit  
  # V = "pooled", # pooled fit  
)  
  
plot <- scplotMulti(  
  result = res,  
  joint = TRUE,  
  type = "series",  
)  
  
plot$plot_out +  
  ...
```

- What are these treatment effects?
- What are separate and pooled fits?

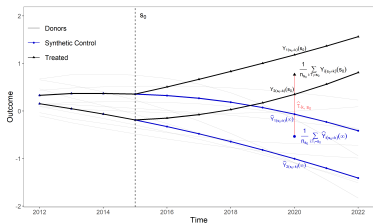
Quantities of Interest



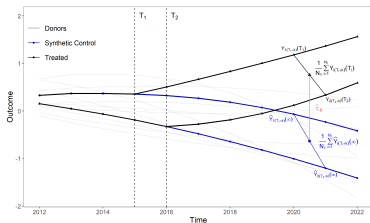
(a) Individual treatment effect



(b) Average post-treatment effect

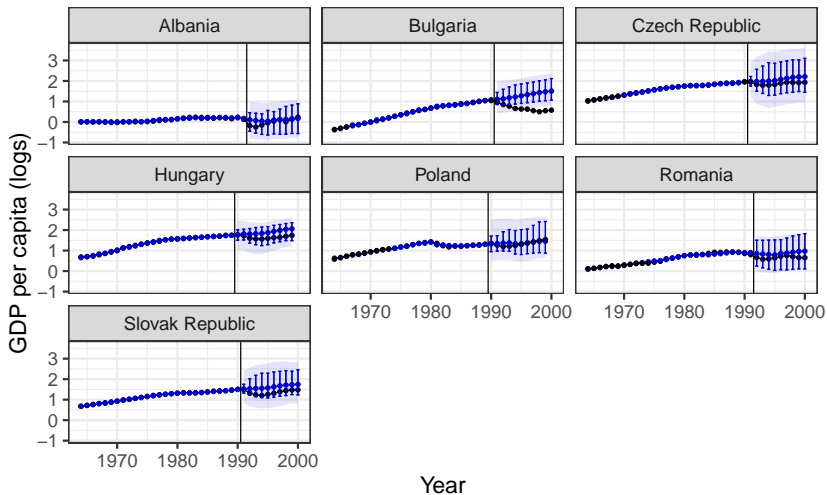


(c) ATT (simultaneous adoption)



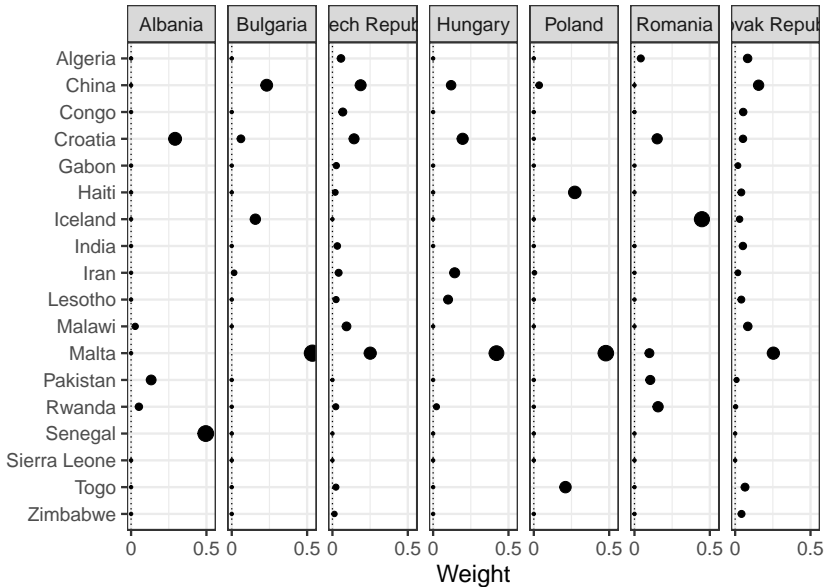
(d) ATT (staggered adoption)

Individual Treatment Effects

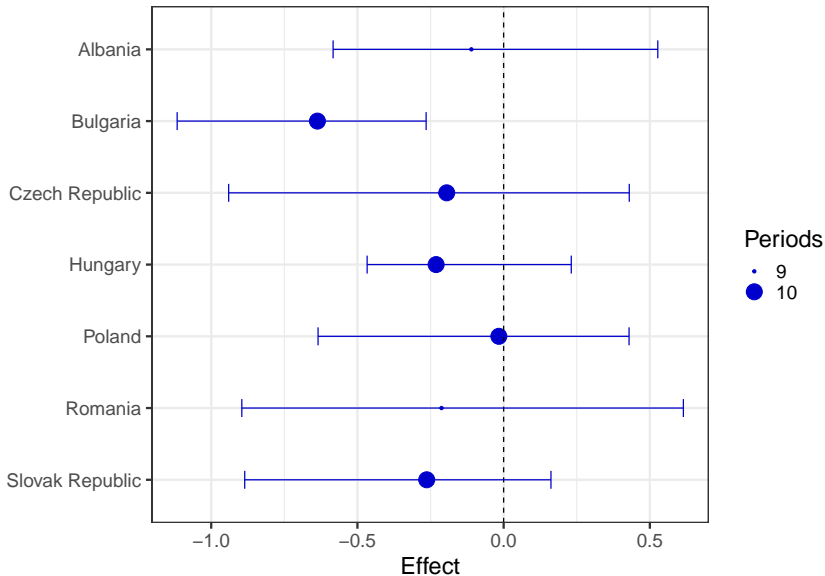


—●— Treated —●— Synthetic Control

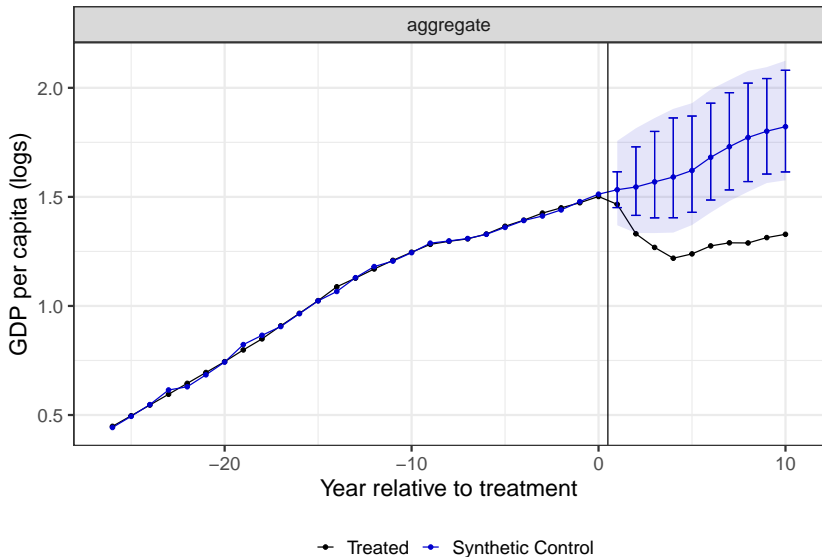
Unit Weights



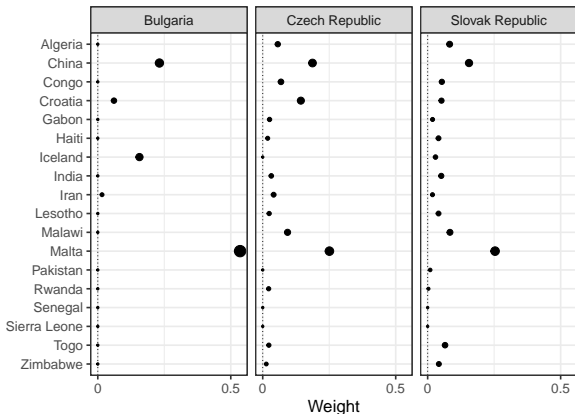
Average Post-Treatment Effects



ATT for Countries Liberalized in 1991



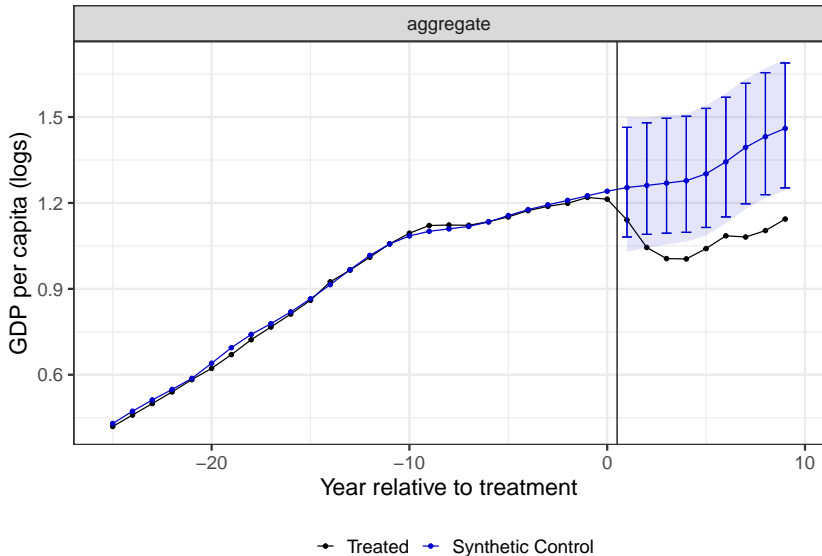
Separate Fit



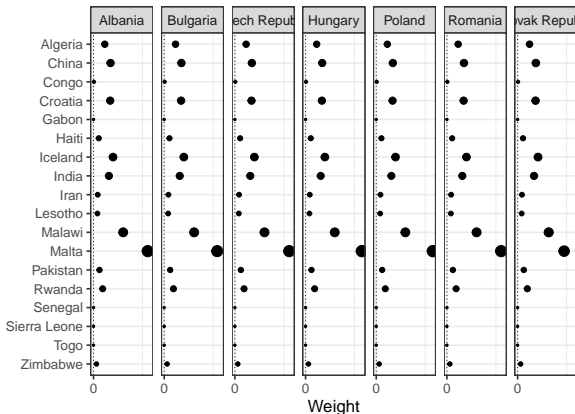
- $V = \text{“separate”}$ to optimize separate fit for each treated unit:

$$\arg \min_{w^i \in \mathcal{W}} \sum_{t=1}^{T_0} \left(Y_{it} - \sum_{j=N_1+1}^{N_1+N_0} w_j^i Y_{jt} \right)^2$$

ATT for All Liberalized Countries



Pooled Fit



- $V = \text{“pooled”}$ to optimize pooled fit for avg. of treated units:

$$\arg \min_{w^i \in \mathcal{W}} \sum_{t=1}^{T_0} \left[\frac{1}{N_1} \sum_{i=1}^{N_1} \left(Y_{it} - \sum_{j=N_1+1}^{N_1+N_0} w_j^i Y_{jt} \right) \right]^2$$

Conclusion

Synthetic control method is a powerful tool for policy evaluation and causal inference in panel data settings.

Recent advancements include:

- Different constraints on weight selection.
- Uncertainty quantification through prediction intervals.
- Handling multiple treated units with staggered adoption.

Future Research

Some of my ongoing projects:

- Can the house hoarding tax reduce Taipei's vacancy rate? A synthetic control approach.
- High-dimensional regression association: A matching and synthetic control approach.
- Mediation analysis synthetic control with multiple mediators.
- Mediation analysis synthetic control with multiple treated units.