Synthetic Control Methods: Theory and application

Dr. Juan Manuel del Pozo Segura

NFER, PUCP

jmdelpozo@pucp.pe

PUCP Q Lab - June 2023

Outline

1. Motivation
2. Formal aspects
   - Potential outcomes, econometric model and its bias
   - Comparison to regression and estimation of the SC
   - Inference, falsification and robustness
   - Limitations and recommendations for practitioners
3. Empirical application
   - The intervention in Abadie et al. (2010)
   - The command synth and examples
   - Replication of the paper
4. Current developments
   - How reasonable is the Convex Hull condition?
   - The problem with the Cross Validation method for V^*
   - How to correctly specify the model and avoid cherry picking

SCM as a causal estimator

- As we have seen so far in the course, the most important questions in economics usually involve analysing **causal questions**.
  - In some cases, these causal questions can be "easily" answered by an RCT
  - Unfortunately, in most cases, causal questions are not and *cannot* be answered via this golden standard. E.g.
    - Smoking and cancer
    - Immigration and wages

  Consequently, we need to rely on **quasi experiments**
- Depending on the nature of the data and on the phenomenon under analysis, we can use different **identification strategies**
  - This class is concerned with 1 of those: Abadie's **Synthetic Control Method (SCM)**
  - *"The synthetic control approach developed by Abadie et al. [2010, 2015] and Abadie and Gardeazabal [2003] is arguably the most important innovation in the policy evaluation literature in the last 15 years."*
    Susan Athey and Guido Imbens (2017)

The type of interventions of interest here

- We are interested in **comparative case studies** where
  1. The treatment happens on **aggregated entities** (such as states)
  2. We have to compare the evolution of aggregate outcomes
     - for units affected by a particular occurrence of the event or intervention of interest
     - for some control group of unaffected units
- E.g.
  - Card (1990): impact of the 1980 Mariel Boatlift using other cities in the southern United States as comparison group.
  - Card and Krueger (1994): evolution of employment in fast food restaurants in NJ and its PN around the time of an increase in NJ's minimum wage.
  - Abadie and Gardeazabal (2003): evolution in GDP in Basque Country and other Spanish regions due to terrorist conflict in the former
- These studies are feasible because
  1. many policy interventions of interest in the social sciences take place at an *aggregate* level
  2. widespread availability of data for units affected by the event of interest and a set of unaffected units
     - Many times we have data on a sample of disaggregated units
     - In case we do not, we only require aggregate data

Why should we bother learning a method other than DiD?

- We now know that we can analyse case studies like these via **DiD** when we have data both
  1. for the period before and after the quasi-experiment
  2. for a (set of) treated an control units
- Advantages:
  - Simple, powerful and no strict need of panel: works also with RCS
  - Vast literature on how to correctly analyse inference issues (**Cameron & Miller 2015** [and references therein], Bertrand, M. et al. 2002, Brewer, M. et al. 2017)
  - Modifications such as 3D, 4D that relax the PTA assumption for identification
- Disadvantages
  1. What units we choose as controls?
  2. With small G and G_1 (as usually is) we have small power and we need to work hard on inference (Cameron & Miller 2015, Mackinnon, J. & Webb, M. 2016 and see also Hansen, B. (2007) for the policy autocorrelation problem)
  3. Strong reliance on the PTA assumption for identification (Lee 2016, Angrist & Pischke 2008, Imbens & Wooldridge 2015)

Let's see each of these in turn.

What units we choose as controls?

- One of the seminal papers for DiD comes from Card (1990)
  - This studies the impact of the 1980 Mariel Boatlift when approximately 125,000 Cubans emigrated to Florida over this 6 month period of time.
  - Card saw this as an **exogenous shift** in the labor supply curve.
    - treated area: Miami
    - control group: Atlanta, Los Angeles, Houston and Tampa-St. Pbrg. *The choice of these cities is in a footnote in the paper*: "similar based on demographics *and* economic conditions"

  Card estimated a **simple DD** model and found no effect
- However, Angrist & Krueger (1999) show the risk of inference from analyzing an event with a small number of treatment and control units.
  - Analyze as Card (1990) a "non-existent" Mariel Boatlift in *1994*
    - In 1994 Castro again announced that Cubans who wanted to leave, could leave.
    - So, a big inflow of refugees to Miami was about to take place but *did not.*
  - They show that between 1993 (pre-non-shock) and 1995 (post-non-shock) unemployment rate for Black workers
    - in Miami increased by 3.6 pp
    - in control group of cities in Card (1990) decreased by 2.7 pp.
  - Hence, *we would estimate a* **fake treatment effect** *of 6.3* pp

The problem with small G ...

- Cameron & Miller (2015) (and the last chapter of Lee 2016) greatly discuss the inference issues around FE, as well as what to cluster over.
- In the **standard setting** when $G \to \infty$, we can use
  - the Liang & Zeger CRVE $\hat{V}_{clu}\left(\hat{\beta}\right) = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \sum_{g=1}^{G} \mathbf{X}_g' \hat{\mathbf{u}}_g \hat{\mathbf{u}}_g' \mathbf{X}_g \left(\mathbf{X}'\mathbf{X}\right)^{-1}$ (, `vce(cluster clustervar)`) which
    - use T_{G-1} critical values
    - replace $\hat{\mathbf{u}}_g$ by $\hat{\mathbf{u}}_g \times \sqrt{\frac{G}{G-1}\frac{N-1}{N-K}}$

    even after including the regions FEs dummies ◂ DiD Clusters (G)
  - pairs cluster bootstrap
- However, an important part of this paper is devoted to the problem of small G (e.g.: Peru, with $G = 24$). In this case, the asymptotics in L&Z CRVU do not work. We need to painstakingly take other routes
  - Bell & McCaffrey (2002) CR2VE and CR3VE bias corrections
  - Wild cluster bootstrap method (Rademacher and/or Webb weights)
  - Donald & Lang (2007) T_{G-L} distribution
  - Imbens & Kolesar (2012) effective DoF and Carter et al. (2017) effective G

...and to put matters worse: the problem with small G_1

- MacKinnon & Webb (2016) emphasize another usually overlooked aspect in DiD inference with large G: L&Z CRVE assumes
  - A1. $G \to \infty$
  - A2. The within-cluster error correlations are the same for all $g \in G$
  - A3. Each $g \in G$ contains an equal number of observations $(N_g, \forall g)$
- So, *even with large G, L&Z can fail* because the 'rule of $G = 42$' no longer holds when assumption 3 relaxed. They find that in the **DiD case** with dichotomous regressors that
  - N_g matters and in cases with large imbalances, the wild bootstrap works reasonably well
  - But unfortunately, all the methods fail *badly* when G_1 is small or (in some cases) large

PTA assumption for identification

- The potential outcome for a *non treated observation* in a given state s and time t is

$$E\left(Y_i^0|j,t\right) = \mu_j + \lambda_t$$

in this case, Y will be determined by the sum of

1. a time-invariant state fixed effect, μ_j idiosyncratic to the state j
2. a time effect λ_t that is common across all states

- Under the **conditional independence assumption**, the **average treatment effect** is $E\left(Y_i^1 - Y_i^0|j,t\right) = \beta_d$. So, the observed outcome can be written as

$$Y_{ijt} = \text{non treatment} + \text{treatment effect if treated} + \varepsilon_{ist} \rightarrow Y_i = \underbrace{\mu_j + \lambda_t}_{E\left(Y_i^0|j,t\right)} + \beta_d D_i + \varepsilon_{ijt}$$

$\mu_j + \lambda_t$ represents the PTA; its fulfillment makes the DiD valid ◂ DiD PTA

- Given this, β_d will give us the causal effect if we include, along with D_{st}, dummies for states j and periods t. This *will work* because this will wipe out the FEs μ_j and τ_t in this linear setting
- However, since the validity of this depends on the PTA, and this in turn depends on the fact that FEs enter additively, our 2-way FE will not yield the causal effect if the FEs are in fact, e.g., interacted.

What does this mean for DiD estimator?

- DiD is *still* a very useful and powerful tool for analysis. However, the circumstances where it is weak are staples of economic quasi-experiment. In many cases
  1. The treatment happens on **aggregated entities** (such as states) on a population of interest with small G
  2. In several important cases, the treatment occurs on $G_1 = 1$ or $G_1 =$ very low
  3. In an attempt to escape the small G problem we take all the non treated units as control group
  4. How do we know that the additivity assumption for the FEs actually holds?
- Following Abadie et al. (2010), a seminal paper:
  - There is uncertainty about the ability of the subjectively-chosen control group to reproduce the counterfactual outcome trajectory that the affected units would have experienced in the absence of the intervention or event of interest.
  - This uncertainty is *not* reflected by the standard errors constructed with traditional inferential techniques for comparative case studies

How does SCM help in overcoming this?

- In this seminal paper, Abadie et al. advocate the use of data-driven procedures to construct suitable comparison groups to reduce discretion in the choice of the comparison control units
  - It is *difficult to find 1 unexposed unit* that approximates nest the unit(s) exposed to the event of interest.
  - So, the core the idea behind the SCM is to take as **counterfactual** a **Synthetic Control** (**SC**), a combination of units which provides a *better* comparison for the unit exposed to the intervention than any single unit alone.
- E.g.
  - Abadie and Gardeazabal (2003):they use a combination of 2 Spanish regions to approximate the economic growth that the Basque Country would have experienced in the absence of terrorism.
  - Peri and Yasenov (2019): use a combination of cities in USA to approximate the evolution that the Miami labor market would have experienced in the absence of the Mariel Boatlift.

How does SCM help in overcoming this?

- SCM has 3 attractive features relative to traditional regression methods
  1. **transparency:** Because the SC is a weighted average of the available control units, the SCM makes explicit:
     - the *relative contribution* of each control unit to the counterfactual of interest
     - the *similarities* (or lack thereof) between the unit affected by the intervention of interest and the SC, in terms of

     1.1 preintervention outcomes

     1.2 predictors of postintervention outcomes
  2. safeguard against **extrapolation**:
     - the weights that make up the SC, the counterfactual, sum up to 1
     - so by using as counterfactual a convex hull of control group units, it is based on where data *actually is*
  3. construction of the SC does **not require access to the post-treatment outcomes**.
- Importantly, the model extends the traditional linear difference-in-differences framework. This allows that the effects of unobserved variables on the outcome vary with time
- However, we need to know when this method can actually be implemented. There is no free lunch and it also has some limitations.

Outline

1. Motivation

2. Formal aspects
   - Potential outcomes, econometric model and its bias
   - Comparison to regression and estimation of the SC
   - Inference, falsification and robustness
   - Limitations and recommendations for practitioners

3. Empirical application
   - The intervention in Abadie et al. (2010)
   - The command synth and examples
   - Replication of the paper

4. Current developments
   - How reasonable is the Convex Hull condition?
   - The problem with the Cross Validation method for V^*
   - How to correctly specify the model and avoid cherry picking

A motivating model

- In this section we follow Abadie et al. (2010), and we will supplement it with Abadie et al. (2015) and Abadie (2020)
- To simplify the exposition, we assume that
  - there are $J+1$ aggregated units (we could aggregate the data from the regions exposed to the intervention)
    - **only 1 unit or region is subject to the intervention of interest**
    - there are J unexposed units, called the **donor pool (DP)**
  - there are T periods: $\underbrace{1, 2, ..., T_0}_{\text{pre-treatment}}; \underbrace{T_0+1, T_0+2, ..., T}_{\text{post-treatment}}$
- In terms of the potential outcomes
  1. Y_{jt}^N is the *outcome* that would be observed for
     1.1 $j = 2, ..., J+1$ in $t = 1, ..., T$, i.e. at all periods for the **control units** units
     1.2 $j = 1$ in $t = 1, ..., T_0$, i.e. where there is not intervention for the treated unit
  2. Y_{jt}^I is the *outcome* that would be observed for $j = 1$ in $t = T_0+1, .., T$, i.e. where there is exposure to the intervention for the treated unit

  We assume the intervention has *no effect* on the outcome before T_0+1
- We still hold the **SUTVA**: Y_i, $i = 2, ..., J+1$ are not affected by intervention in $i = 1$

The treatment effect and what it takes to estimate it

- The **effect of the intervention** for unit $j = 1$ at *every year* t is given by

$$\boxed{\alpha_{1t} = Y^I_{1t} - Y^N_{1t}}$$

Since $Y_{1t} = Y^I_{1t}$ for $t > T_0$, then for $t > T_0$, when the treatment took place

$$\alpha_{1t} = Y^I_{1t} - Y^N_{1t} = Y_{1t} - \boxed{Y^N_{1t}}$$

Because Y_{1t} is observed, the challenge is estimating the **counterfactual** Y^N_{1t}. It is a counterfactual because

- At $t > T_0$ the treatment for $j = 1$ *already happened* and then we observe Y^I_{1t}
- So, at $t > T_0$ we estimate Y^N_{1t} for $j = 1$ based on the control units. SCM provides a new way to use this information in the control units to build the counterfactual!

- Since $\alpha_{1t} = Y_{1t} - Y^N_{1t} \Leftrightarrow Y_{1t} = Y^N_{1t} + \alpha_{1t}$, we can define D_{it} as a dummy:
  - 1 if $j = 1$ and is exposed to the intervention at year t $(t = T_0 + 1, ..., T)$
  - 0 otherwise

So, the observed outcome for unit j at any year t is

$$\boxed{Y_{jt} = \text{non treatment} + \text{treatment if treated} = Y^N_{jt} + \alpha_{jt} D_{jt}}$$

How does the SC estimates the counterfactual

- SCM takes a different route to find the **counterfactual for the treated** compared to regression models. The way this is estimated is via a **weighted** average of the observed outcomes of the Y of the units in the DP

$$Y_{1t}^{N} = \sum_{j=2}^{J+1} w_j^* Y_{jt}^{N}$$

  where
  - Y_{jt}^{N} is the observed values of the control units
  - w_j^* are the optimal weights provided by the SCM (1 for each j in the DP). These allow us to *reasonably* approximate the outcome in $j = 1$
- How does this way of estimating counterfactuals improves upon those provided by DiD? The strategy Abadie follows is
  1. He *proves* that, theoretically, the counterfactual used by SCM, $\sum_{j=2}^{J+1} w_j^* Y_{jt}^{N}$, has a *low bias* when Y_{jt}^{N} has a more realistic behaviour than that implied by the DiD/FE model
  2. Once this is found, he shows how to calculate the w_j^*

The underlying model (the big strength!)

- The model he chooses to generalize the DiD is the **linear factor model**

$$Y_{jt}^N = \mu_j + \lambda_t + \theta_t \boldsymbol{Z}_j + \varepsilon_{jt} \rightarrow \boxed{Y_{jt}^N = \delta_t + \underbrace{\theta_t \boldsymbol{Z}_j}_{\text{this can change every year}} + \underbrace{(\boldsymbol{\lambda}_t)_{(1\times F)} (\boldsymbol{\mu}_j)_{(F\times 1)}}_{\text{unobserved}} + \varepsilon_{jt}}$$

where

- δ_t are time FEs (i.e. 1 dummy per year)
- $\boldsymbol{Z}_j$ is a $r \times 1$ vector of observed variables (not affected by the intervention) and θ_t is a $1 \times r$ vector of coefficients (i.e. different effects of $\mathbf{Z}_i$ for different years)
- ε_{jt} are unobserved transitory shocks at the region level with $E(\varepsilon_{it}) = 0$
- $\boldsymbol{\lambda}_t$ is a vector of unobserved **common factors** that change in time and $\boldsymbol{\mu}_j$ is a vector of unknown **factor loadings** This is where the flexibility comes from!

- This model is used only for theoretical purposes: after showing that the bias of the estimator is low, he shows how to estimate w^*

What does the factor model actually allows for?

- The key to understand the advantage of this model lies in $(\boldsymbol{\lambda}_t)_{(1\times F)}(\boldsymbol{\mu}_j)_{(F\times 1)}$. Following Bai (2009), we can express this as

$$(\boldsymbol{\lambda}_t)_{(1\times F)}(\boldsymbol{\mu}_j)_{(F\times 1)} = \lambda_{1t}\mu_{1j} + \lambda_{2t}\mu_{2j} + ... + \lambda_{Ft}\mu_{Fj}$$

- On the one hand, if $F = 2$, so that $\boldsymbol{\lambda}_t = \begin{bmatrix} 1 & \tau_t \end{bmatrix}$ and $\boldsymbol{\mu}_j = \begin{bmatrix} 1 & \alpha_j \end{bmatrix}'$ then $(\boldsymbol{\lambda}_t)_{(1\times 2)}(\boldsymbol{\mu}_j)_{(2\times 1)} = \alpha_i + \tau_t$ and so we are in the canonical DiD model

$$Y_{jt}^N = \delta_t + (\boldsymbol{\theta}_t)_{(1\times r)}(\boldsymbol{Z}_j)_{r\times 1} + \alpha_j + \tau_t + \varepsilon_{it}$$

  and *only* if this is the case, then these FEs can be eliminated by taking time differences. I.e. the canonical FE model

  1. allows for the presence of unobserved confounders
  2. *restricts* the effect of those confounders to be constant in time

- On the other hand, in the factor model the FEs can enter in the model interactively and not only additively. Hence, if this is the case, these FEs *cannot* be eliminated by taking time differences. I.e. the factor model

  1. allows for the presence of unobserved confounders
  2. allows the effects of confounding unobserved characteristics to *vary with time*

The counterfactual from the factor model

- To see how good the SCM is when Y_{jt}^N follows such general behaviour, we need to form the counterfactual $\sum_{j=2}^{J+1} w_j^* Y_{jt}^N$, also known as the Synthetic Control (**SC**). To do so, we need a vector of weights $\boldsymbol{W}$ which will allow creating the SC

$$\boldsymbol{W} = (w_2, ..., w_{J+1})' \text{ s.t.1) } w_j \geq 0 \text{ for } j = 2, ..., J+1$$
$$2) \; w_2 + \cdots + w_{J+1} = 1$$

  note that these there are 2 restrictions on the w_j^*s. These assure **sparse SC**, i.e. made of small number of js in the DP
- Once we get these $w_j; j = 2, ..., J+1$ we create the SC calculating using $j = 2, ..., J+1$ in the DP

$$\sum_{j=2}^{J+1} w_j Y_{jt}^N = \sum_{j=2}^{J+1} w_j \left(\delta_t + \boldsymbol{\theta}_t \boldsymbol{Z}_j + \boldsymbol{\lambda}_t \boldsymbol{\mu}_j + \varepsilon_{it}\right)$$
$$= \sum_{j=2}^{J+1} w_j \delta_t + \sum_{j=2}^{J+1} w_j \boldsymbol{\theta}_t \boldsymbol{Z}_j + \sum_{j=2}^{J+1} w_j \boldsymbol{\lambda}_t \boldsymbol{\mu}_j + \sum_{j=2}^{J+1} w_j \varepsilon_{jt}$$
$$= \delta_t \underbrace{\sum_{j=2}^{J+1} w_j}_{=1} + \boldsymbol{\theta}_t \sum_{j=2}^{J+1} w_j \boldsymbol{Z}_j + \boldsymbol{\lambda}_t \sum_{j=2}^{J+1} w_j \boldsymbol{\mu}_j + \sum_{j=2}^{J+1} w_j \varepsilon_{jt} = \delta_t + \boldsymbol{\theta}_t \sum_{j=2}^{J+1} w_j \boldsymbol{Z}_j + \boldsymbol{\lambda}_t \sum_{j=2}^{J+1} w_j \boldsymbol{\mu}_j + \sum_{j=2}^{J+1} w_j \varepsilon_{jt}$$

The bias of the estimator for the counterfactual

- To derive **bias**, $Y_{1t}^N - E\left(\sum_{j=2}^{J+1} w_j Y_{jt}^N\right) = E\left(Y_{1t}^N - \sum_{j=2}^{J+1} w_j Y_{jt}^N\right)$, he calculates

$$\underbrace{Y_{1t}^N}_{\text{counterfactual}} - \underbrace{\sum_{j=2}^{J+1} w_j Y_{jt}^N}_{\text{estimator}} = (\delta_t + \boldsymbol{\theta}_t \boldsymbol{Z}_1 + \boldsymbol{\lambda}_t \boldsymbol{\mu}_1 + \varepsilon_{1t}) - \left(\delta_t + \boldsymbol{\theta}_t \sum_{j=2}^{J+1} w_j \boldsymbol{Z}_j + \boldsymbol{\lambda}_t \sum_{j=2}^{J+1} w_j \boldsymbol{\mu}_j + \sum_{j=2}^{J+1} w_j \varepsilon_{jt}\right)$$

$$= \boldsymbol{\theta}_t \left(\boldsymbol{Z}_1 - \sum_{j=2}^{J+1} w_j \boldsymbol{Z}_j\right) + \boldsymbol{\lambda}_t \left(\boldsymbol{\mu}_1 - \sum_{j=2}^{J+1} w_j \boldsymbol{\mu}_j\right) + \left(\underbrace{(w_2 + \ldots + w_{J+1})}_{=1} \varepsilon_{1t} - \sum_{j=2}^{J+1} w_j \varepsilon_{jt}\right)$$

$$= \boldsymbol{\theta}_t \left(\boldsymbol{Z}_1 - \sum_{j=2}^{J+1} w_j \boldsymbol{Z}_j\right) + \boldsymbol{\lambda}_t \left(\boldsymbol{\mu}_1 - \sum_{j=2}^{J+1} w_j \boldsymbol{\mu}_j\right) + \sum_{j=2}^{J+1} w_j \left(\varepsilon_{1t} - \varepsilon_{jt}\right)$$

- After reducing this expression, and in order to progress, we need to assume the existence of **optimal weights** $\left(w_2^*, \ldots, w_{J+1}^*\right)$ such that for the pre-intervention period

$$\underbrace{\sum_{j=2}^{J+1} w_j^* Y_{jt}}_{\text{donors}} = w_2^* Y_{2t} + w_3^* Y_{3t} + \ldots + w_{J+1}^* Y_{J+1,t} = \underbrace{Y_{1t}}_{\text{treated}} ; t = \underbrace{1, \ldots, T_0}_{\text{pre-intervention}}$$

$$\underbrace{\sum_{j=2}^{J+1} w_j^* \boldsymbol{Z}_j}_{\text{donors}} = w_2^* \boldsymbol{Z}_2 + w_3^* \boldsymbol{Z}_3 + \ldots + w_{J+1}^* \boldsymbol{Z}_{J+1} = \underbrace{\boldsymbol{Z}_1}_{\text{treated}} ; t = \underbrace{1, \ldots, T_0}_{\text{pre-intervention}}$$

this is also known as the **convex hull assumption**

The bias of the estimator for the counterfactual

- This convex hull assumption allows us to write $\boldsymbol{\theta}_t\left(\boldsymbol{Z}_1-\sum_{j=2}^{J+1}w_j\boldsymbol{Z}_j\right)+\boldsymbol{\lambda}_t\left(\boldsymbol{\mu}_1-\sum_{j=2}^{J+1}w_j\boldsymbol{\mu}_j\right)+\sum_{j=2}^{J+1}w_j\left(\varepsilon_{1t}-\varepsilon_{jt}\right)$ as

$$\underbrace{Y_{1t}^N}_{\text{counterfactual}}-\underbrace{\sum_{j=2}^{J+1}w_j^*Y_{jt}^N}_{\text{estimator}}=\underbrace{\boldsymbol{\lambda}_t\left(\boldsymbol{\lambda}^{P\prime}\boldsymbol{\lambda}^P\right)^{-1}\boldsymbol{\lambda}^{P\prime}\sum_{j=2}^{J+1}w_j^*\varepsilon_j^P}_{=R_{1t}}-\underbrace{\boldsymbol{\lambda}_t\left(\boldsymbol{\lambda}^{P\prime}\boldsymbol{\lambda}^P\right)^{-1}\boldsymbol{\lambda}^{P\prime}\varepsilon_1^P}_{=R_{2t}}+\underbrace{\sum_{j=2}^{J+1}w_j^*\left(\varepsilon_{1t}-\varepsilon_{jt}\right)}_{=R_{3t}}$$

- So, when $w_j=w_j^*, j=2,\ldots,J+1$, the **bias** for $t>T_0$ is

$$\underbrace{E\left[Y_{1t}^N-\sum_{j=2}^{J+1}w_j^*Y_{jt}^N\right]}_{\text{Bias}}=\underbrace{E\left(\boldsymbol{\lambda}_t\left(\boldsymbol{\lambda}^{P\prime}\boldsymbol{\lambda}^P\right)^{-1}\boldsymbol{\lambda}^{P\prime}\sum_{j=2}^{J+1}w_j^*\varepsilon_j^P\right)}_{=E(R_{1t})\neq 0}-\underbrace{E\left(\boldsymbol{\lambda}_t\left(\boldsymbol{\lambda}^{P\prime}\boldsymbol{\lambda}^P\right)^{-1}\boldsymbol{\lambda}^{P\prime}\varepsilon_1^P\right)}_{=E(R_{2t})=0}+\underbrace{E\left(\sum_{j=2}^{J+1}w_j^*\left(\varepsilon_{1t}-\varepsilon_{jt}\right)\right)}_{=E(R_{3t})=0}$$

- Applying the Cauchy–Schwarz, the Hölder and the Rosenthal inequality, this bias is **bounded by**, i.e. *is at most*

$$\underbrace{E\left[Y_{1t}^N-\sum_{j=2}^{J+1}w_j^*Y_{jt}^N\right]}_{\text{Bias}}=E\left(R_{1t}\right)\leq\left[C(P)^{\frac{1}{p}}\left(\frac{\overline{\lambda}^2F}{\xi}\right)\boxed{J^{\frac{1}{p}}}\right]\times max\left\{\frac{\overline{m}_p^{\frac{1}{p}}}{T_0^{1-\frac{1}{p}}},\frac{\overline{\sigma}}{T_0^{\frac{1}{2}}}\right\}$$

where $\overline{m}_p=\underset{j=2,\ldots,J+1}{max}\left(\frac{1}{T_0}\sum_{t=1}^{T_0}E|\varepsilon_{jt}|^p\right)$ is the **scale** of the idiosyncratic errors . Note how this depends on T_0, $\overline{m}_p$ and J

What does this bias bound expression tell us

- We want to use an estimator that has a low bias. This expression above tell us the *maximum amount* of bias of our SCM estimator. Hence, we want the RHS to be as small as possible so that the bias is small
- To find what we need for this to happen, note that this expression depends mainly on the second term

$$max\left\{\frac{\boxed{\overline{m}_p^{\frac{1}{p}}}}{\boxed{T_0^{1-\frac{1}{p}}}}, \frac{\overline{\sigma}}{\boxed{T_0^{\frac{1}{2}}}}\right\} \equiv max\left\{\frac{\left[\max_{j=2,\ldots,J+1}\left(\frac{1}{T_0}\sum_{t=1}^{T_0} E|\varepsilon_{jt}|^p\right)\right]^{\frac{1}{p}}}{\boxed{T_0^{1-\frac{1}{p}}}}, \frac{\overline{\sigma}}{\boxed{T_0^{\frac{1}{2}}}}\right\}$$

so, it will be small (close to 0) if T_0 is large relative to the scale of ε_{jt}

- Only in the case we can be sure of that the bias is almost 0, and so the SCM method works. The unbiased estimator of the TE, α_{1t}, is given by

$$\boxed{\hat{\alpha}_{1t} = Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt}; t \in \{T_0+1, \ldots, T\}}$$

What we need for this nice result to happen

- Remember that for this result to hold, we required to *assume* that there exists optimal weights $\left(w_2^*, ..., w_{J+1}^*\right)$ that create a SC that **fits perfectly** the Y and covariates $\boldsymbol{Z}$ of $j = 1$ in the pre-intervention period: $\sum_{j=2}^{J+1} w_j^* Y_{jt} = Y_{1t}$ and $\sum_{j=2}^{J+1} w_j^* \boldsymbol{Z}_j = \boldsymbol{Z}_1$ for $t = 1, ..., T_0$
- However, in practice this perfect fit is unlikely
  - it is *often the case* that $\sum_{j=2}^{J+1} w_j^* Y_{jt} \approx Y_{1t}$ and $\sum_{j=2}^{J+1} w_j^* \boldsymbol{Z}_j \approx \boldsymbol{Z}_1$ but still the bias bound *kicks in*
  - in some cases
    - $\sum_{j=2}^{J+1} w_j^* Y_{jt} << Y_{1t}$ or $\sum_{j=2}^{J+1} w_j^* Y_{jt} >> Y_{1t}$, and
    - $\sum_{j=2}^{J+1} w_j^* \boldsymbol{Z}_j << \boldsymbol{Z}_1$ or $\sum_{j=2}^{J+1} w_j^* \boldsymbol{Z}_j >> \boldsymbol{Z}_1$

    and so there is a *large bias* in the estimation so that *it is not recommended using the SCM*
- Hence, it is important to compare Y_{1t} and $\boldsymbol{Z}_1$ (for $j = 1$) with $\sum_{j=2}^{J+1} w_j^* Y_{jt}$ and $\sum_{j=2}^{J+1} w_j^* \boldsymbol{Z}_j$ (for the SC) to decide if the characteristics of the treated unit are sufficiently matched by the SC in the pre-treatment

◂ Example of good pre-treat. covariate balance

How the method can control for unmeasured factors

- Abadie (2020) states that, under the factor model, a SC that reproduces $\mathbf{Z}_1$ *and* $\boldsymbol{\mu}_1$ would provide an *unbiased* estimator of the ATET
◂ Linear factor model
. However, $\boldsymbol{\mu}_1$ is not observed so it cannot be matched directly in the data
- Still, a SC *that*
  1. reproduces the trajectory of the Y_1 over an extended period of time, and
  2. reproduces the values of $\mathbf{Z}_1$

  indicates of low bias because this suggests that the remaining term, $\boldsymbol{\mu}_1$, is similarly matched
◂ Example of good pre-treatment fit
- Note that
  1. Large T_0 automatically cannot drive down the bias if the fit is bad
  2. The SC can closely match pre-treatment outcomes without actually matching the values of μ_1

J and the bias

- Under a factor model for Y_{it}^N , larger J
  - Makes it easier to fit pre-treatment outcomes even when there are substantial discrepancies in μs between the $j = 1$ and the SC
  - However, the bias bound depends positively on J because a large number of units in the DP may create or exacerbate the bias of the estimator, especially if the μ_j in the DP greatly differ from μ_1 ◂ Bias bound
- A practical implication of this is that *each of the units in the DP have to be chosen judiciously* to provide a reasonable control for the treated unit. So that units in the DP should have
  - similar values of the observed attributes $\mathbf{Z}_j$ relative to the treated unit
  - similar values of *the unobserved attributes* μ_j relative to the treated unit

Outline

1. Motivation
2. Formal aspects
   - Potential outcomes, econometric model and its bias
   - Comparison to regression and estimation of the SC
   - Inference, falsification and robustness
   - Limitations and recommendations for practitioners
3. Empirical application
   - The intervention in Abadie et al. (2010)
   - The command synth and examples
   - Replication of the paper
4. Current developments
   - How reasonable is the Convex Hull condition?
   - The problem with the Cross Validation method for V^*
   - How to correctly specify the model and avoid cherry picking

The matrices with pre-intervention characteristics

- $\mathbf{W} = (w_2, ..., w_{J+1})'$ s.t. $w_j \geq 0$ for $j = 2, ..., J+1$ and $w_2 + \cdots + w_{J+1} = 1$
  - As before, each possible $\mathbf{W}$ represents a **SC**, i.e. a *weighted average of the units in the DP*
  - The 2 constraints imply that the SC is a **convex combination** of untreated units so to avoid **extrapolation**
- Let
  - $\boldsymbol{Z}_j$ is a $(r \times 1)$ vector of **observed covariates** for j
  - $\left(\overline{Y}_j^{\mathbf{K}_1}, ..., \overline{Y}_j^{\mathbf{K}_M}\right)$ is a $(M \times 1)$ vector of **observed values** for Y for j

  so
  - $\boldsymbol{X}_1 = \underset{k \times 1 = (r+M) \times 1}{\begin{pmatrix} \boldsymbol{Z}_1 \\ \hline \overline{Y}_1^{\mathbf{K}_1} \\ ... \\ \overline{Y}_1^{\mathbf{K}_M} \end{pmatrix}}$ a vector of ***preintervention*** characteristics for $j = 1$
  - $\boldsymbol{X}_0 = \underset{k \times J = (r+M) \times J}{\begin{pmatrix} \boldsymbol{Z}_2 & \boldsymbol{Z}_3 & ... & \boldsymbol{Z}_{J+1} \\ \hline \overline{Y}_2^{\mathbf{K}_1} & \overline{Y}_3^{\mathbf{K}_1} & ... & \overline{Y}_{J+1}^{\mathbf{K}_1} \\ ... & ... & & ... \\ \overline{Y}_2^{\mathbf{K}_M} & \overline{Y}_3^{\mathbf{K}_M} & ... & \overline{Y}_{J+1}^{\mathbf{K}_M} \end{pmatrix}}$ is a matrix of ***preintervention*** characteristics for js in the DP

◂ Example of X matrices

Comparison to regression-based counterfactuals

- SCM constructs a SC as a linear combination of units in DP under 2 restrictions. However, *also* **regression** creates weighted counterfactuals!

$$\left(\hat{\beta}'\right)_{T_1\times k}(\mathbf{X}_1)_{k\times 1} = \left(\left(\mathbf{X}_0\mathbf{X}_0'\right)^{-1}\mathbf{X}_0\mathbf{Y}_0'\right)'\mathbf{X}_1 = \mathbf{Y}_0\left[\mathbf{X}_0'\left(\mathbf{X}_0\mathbf{X}_0'\right)^{-1}\mathbf{X}_1\right] = (\mathbf{Y}_0)_{T_1\times J}(\mathbf{W}^{reg})_{J\times 1}$$

  where
  - $\mathbf{X}_0$ is a $k\times J$ matrix with *preintervention* characteristics for the units in DP
  - $\mathbf{X}_1$ is a $k\times 1$ vector with *preintervention* characteristics for $j=1$
- The sum of the weights of this R-B counterfactual is given by

$$\left(\iota'\right)_{1\times J}(\mathbf{W}^{reg})_{J\times 1} = \iota'\mathbf{X}_0'\left(\mathbf{X}_0\mathbf{X}_0'\right)^{-1}\mathbf{X}_1 = \left(\left(\mathbf{X}_0\mathbf{X}_0'\right)^{-1}\mathbf{X}_0\iota\right)'\mathbf{X}_1$$

- Assume that, as usual, regression includes an intercept so the first row of $\mathbf{X}_0$ is a vector of 1s. Then, because $\left(\mathbf{X}_0\mathbf{X}_0'\right)^{-1}\mathbf{X}_0\iota$ can be seen as the $\hat{\beta}$s of the regression ι (a vector of 1s) on $\mathbf{x}_0$, the only non-0 coefficient is the intercept (=1). So, *mechanically*,

$$\left(\left(\mathbf{X}_0\mathbf{X}_0'\right)^{-1}\mathbf{X}_0\iota\right)_{K\times 1} = \begin{pmatrix} 1 & 0 & \dots & 0 \end{pmatrix}$$

Comparison to regression-based counterfactuals

- The latter implies that the sum of weights is

$$\left(\iota^{'}\right)_{1\times J}(\mathbf{W}^{reg})_{J\times 1} = \iota^{'}\mathbf{X}_0^{'}\left(\mathbf{X}_0\mathbf{X}_0^{'}\right)^{-1}\mathbf{X}_1 = \left(\left(\mathbf{X}_0\mathbf{X}_0^{'}\right)^{-1}\mathbf{X}_0\iota\right)^{'}\mathbf{X}_1 = \begin{pmatrix} 1 & 0 & \dots & 0 \end{pmatrix}\mathbf{X}_1 = 1$$

  so, the regression-based counterfactuals are created using
  - weights that sum to 1
  - *not restricted to be between 0 and 1*: may take negative values
- As a result, regression-based counterf. can lead to extrapolation outside the support of the comparison units. This implies that
  - regression weights **extrapolate** to produce a perfect fit *even if* the Xs of $j = 1$ cannot be approximated by a weighted average of the Xs of js in the DP
  - Technically, even if $\mathbf{X}_1$ is far from the convex hull of the columns of $\mathbf{X}_0$, regression weights extrapolate to produce

$$\mathbf{X}_0\mathbf{W}_{reg} = \mathbf{X}_0\mathbf{X}_0^{'}\left(\mathbf{X}_0\mathbf{X}_0^{'}\right)^{-1}\mathbf{X}_1 = \mathbf{X}_1$$

- So, usually,
  - regression-based counterfactual relies on **extrapolation**
  - Instead, the SCM-based counterfactual
    1. closely fits the values of the characteristics of the units
    2. does not extrapolate outside of the support of the data

The minimization process to find W^*

- Given this, the vector $\boldsymbol{W}^*$ is chosen to minimize the MSE of the SC

$$\boldsymbol{W}^* = \underset{W}{\operatorname{argmin}} \|\boldsymbol{X}_1 - \boldsymbol{X}_0 \boldsymbol{W}\|_V \equiv (\boldsymbol{X}_1 - \boldsymbol{X}_0 \boldsymbol{W})' \boldsymbol{V} (\boldsymbol{X}_1 - \boldsymbol{X}_0 \boldsymbol{W}) = \sum_{m=1}^{k} v_m \left(\boldsymbol{X}_{1m} - \sum_{j=2}^{J+1} w_j \boldsymbol{X}_{jm} \right)^2$$

$$\text{s.t.} w_2 \geq 0, ..., w_{J+1} \geq 0$$

$$w_2 + ... + w_{J+1} = 1$$

where $\boldsymbol{V} = [v_m], m = 1, ..., k$ is a $k \times k = (r + M) \times (r + M)$ **diagonal** symmetric and positive semidefinite matrix. Each v_m assigns a weight to every covariate m ($\boldsymbol{Z}$ or $\overline{Y}_2^{\mathbf{K}_1} ... \overline{Y}_2^{\mathbf{K}_M}$) to indicate its importance in forming the SC

- Since each potential choice of $\mathbf{V} = (v_1, ..., v_k)$ produces a synthetic control $\boldsymbol{W}^* \equiv \boldsymbol{W}^*(\boldsymbol{V})$, we need to take care on finding a reasonable $\mathbf{V}^*$

The minimization process to find V^*

- The choice of **V**can be subjective (not recommended) or data-driven
  1. **minimization**: Abadie & Gardeazabal (2003) suggest **V** such that the SC (defined by the$\mathbf{W}^*(\mathbf{V})$ we just found) approximates the trajectory of Y_1 *only* in the *preintervention* periods $t = 1, ..., T_0$

$$\boldsymbol{V}^* = \underset{V}{\operatorname{argmin}} \left(\boldsymbol{Y}_1 - \boldsymbol{Y}_0 \boldsymbol{W}^*(\boldsymbol{V})\right)' \left(\boldsymbol{Y}_1 - \boldsymbol{Y}_0 \boldsymbol{W}^*(\boldsymbol{V})\right) = \sum_{t=1}^{T_0} \left(Y_{1t} - \sum_{j=2}^{J+1} w_j^*(V)\, Y_{jt} \right)^2$$

     where **V** is the set of all non-negative diagonal $(K \times K)$ matrices. We use this to derive again the W^* matrix
  2. **regression-based**: we for every t, we regress Y_j on $\sum_{k=1}^{r+M} \beta_k Z_{kj}$ and then we apply Kaul et al. (2018) formula
  3. **cross validation**: if T_0 *is large enough* we can divide pre-intervention periods into
     - 3.1 Initial **training period**: given a **V**, we can compute using only data from this period $\mathbf{W}^*(\mathbf{V})$
     - 3.2 Subsequent **validation period**: **V** minimizes MSPE produced by the weights $\mathbf{W}^*(\mathbf{V})$ during the validation period.

Outline

1. Motivation
2. Formal aspects
   - Potential outcomes, econometric model and its bias
   - Comparison to regression and estimation of the SC
   - Inference, falsification and robustness
   - Limitations and recommendations for practitioners
3. Empirical application
   - The intervention in Abadie et al. (2010)
   - The command synth and examples
   - Replication of the paper
4. Current developments
   - How reasonable is the Convex Hull condition?
   - The problem with the Cross Validation method for V^*
   - How to correctly specify the model and avoid cherry picking

What does inference mean in this aggregate case

- The SEs commonly reported in regression-based comparative case studies reflect only the unavailability of aggregate data
- However, despite that aggregate data is used for estimation in SCM, there is still uncertainty about the value of the parameters of interest.
  - Still there is uncertainty from ignorance about the ability of the control group to reproduce the counterfactual
  - So, using individual micro data increases the total amount of uncertainty if the outcome of interest is an aggregate.
- Large sample inferential techniques are *not well suited to comparative case studies* when the number of units in the comparison group is small.
  - Abadie proposes **exact inferential techniques**, akin to **permutation tests**, in which the distribution of a test statistic is computed under random permutations assigning units to intervention and nonintervention groups
  - This inferential exercise is *exact* in the sense that regardless of the J, T or whether the data are individual or aggregate, it is always possible to calculate the exact distribution of the estimated effect of the placebo interventions.
- Under the H_0 of no intervention effect, an abnormal estimated effects for $j = 1$ relative to the distribution of placebo effects signals its statistical significance

Permutations tests (in-space placebos)

- Take a j in the DP and compute ◂ Example of permutations test
  1. the W^*matrix taking j as the treated and shifting $j=1$ to the DP and, based on $\hat{\alpha}_{1t} = Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt}$, calculate

$$\hat{\alpha}_{jt} = Y_{jt} - \sum_{k=1}^{J+1} w_k^* Y_{kt} \text{ with } k = 1, ..., J+1 \text{ without the current} j$$

  2. the RMSPE for the *pre-treatment period*:

$$RMSPE_j^{pre} = \sqrt{\left(\frac{1}{T_0} \sum_{t=1}^{T_0} \left(Y_{jt} - \sum_{k=1}^{J+1} w_k^* Y_{kt} \right)^2 \right)} \text{ with } k = 1, ..., J+1 \text{ without the current} j$$

  measures *lack of fit* between path of Y_{jt} and its SC in *pre-intervention period*

  Repeating this $\forall j$ in the DP provides the **distribution** of estimated effects
- Based on this, we assess *graphically* if the effect estimated for $j=1$ is large relative to the distrib. of the placebo effects (for the js in the DP).
- But since series with poor fit in pre-treatment period *do not provide good information* to measure the significance of the effect for $j=1$, we need
  1. to create a graphic with the effects of all js ($j=1$ and the placebos)
  2. to create graphics with the effects of *only those* js *with* $\frac{RMSPE_j^{pre}}{RMSPE^{pre}} < cutoff$

$$\boxed{1}$$

P-values

- Abadie et al. (2010) also shows how to construct **exact p-values** which obviates choosing a cut-off for the exclusion of ill-fitting placebos.
  - This measures the *qua*lity of the fit of the SC a j in the post-treatment relative to the quality of the fit in the pre-treatment period
  - Consists on looking at the distribution of **ratios of post and pre treatment period MSPEs** because a *large postintervention* RMSPE is *not* indicative of a large effect of the intervention if the preintervention RMSPE is also large, i.e. if the SC fits poorly Y_{jt} *in the pre-treatment period*
- To do so ◂ Example of p-values
  1. Take a j in the DP and compute
     - 1.1 the W^*matrix taking j as the treated and putting $j = 1$ into the DP
     - 1.2 the ratio of the post-to-pre-treatment RMSPE, with

$$Ratio_j = \frac{RMSPE_j^{post}}{RMSPE_j^{pre}} = \sqrt{\left(\frac{1}{T-T_0}\sum_{t=T_0+t}^{T}\left(Y_{jt} - \sum_{k=1}^{J+1} w_k^* Y_{kt}\right)^2\right)} \Big/ \sqrt{\left(\frac{1}{T_0}\sum_{t=1}^{T_0}\left(Y_{jt} - \sum_{k=1}^{J+1} w_k^* Y_{kt}\right)^2\right)} \text{ with } k = 1, ..., J+$$

  2. Repeat this $\forall j$ in the DP to get the distribution of estimated effects
  3. Sort the ratios in descending order from greatest to highest
  4. Calculate the **treatment unit's ratio** in the distribution as

$$p = \frac{RANK}{TOTAL\ OF\ OBSERVATIONS}$$

Falsification: back casting (in-time placebos)

- As usual, an important part of the applied work is concerned with *validating* our identification strategy. One way to do this concerns the timing of the intervention (Abadie et al. 2015)
- Suppose, that the SCM estimates a sizable effect for a certain intervention of interest.
  - This result would *not* be valid if the method *also* estimated large effects when applied to dates $t < T_0$ when the intervention did not occur (similar to the "pre-program test" in Heckman and Hotz (1989)
  - We refer to these falsification exercises as "**in-time placebos**". We expect there are no estimated effects prior to the intervention, i.e. that the treatment appear *only around* the $T_0 + 1$
- These tests are feasible if there are available data for a sufficiently large T when no structural shocks to the outcome variable occurred

Robustness: model specification and leave-one-out

- In order to assess the robustness of our results,
  1. we can include additional predictors of the outcome to construct the synthetic control. We expect the results to not change much regardless of which and how many predictor variables we included.
  2. **leave-one-out**: testing the sensitivity of results to changes in js that provide $w > 0$ in W^*. E.g.
     - Say that the SC is estimated as a weighted average of $j = 2, 3, 4$ out of $J = 10$ (i.e. received a positive weight)
     - So, we can *iteratively* re-estimate the baseline model to construct a SC omitting *everytime* 1 of the js in $j = 2, 3, 4$.

     By excluding each of these j we sacrifice some goodness of fit but *still* this allows us to evaluate to what extent our results are driven by any particular control country.

Outline

1. Motivation
2. Formal aspects
   - Potential outcomes, econometric model and its bias
   - Comparison to regression and estimation of the SC
   - Inference, falsification and robustness
   - Limitations and recommendations for practitioners
3. Empirical application
   - The intervention in Abadie et al. (2010)
   - The command synth and examples
   - Replication of the paper
4. Current developments
   - How reasonable is the Convex Hull condition?
   - The problem with the Cross Validation method for V^*
   - How to correctly specify the model and avoid cherry picking

Warnings

- The SCM facilitates comparative case studies when *no single untreated unit* provides a *good comparison* for the unit affected by the treatment or event of interest. This is often the case
  - When the treatment affects large aggregates like regions or countries
  - This results that a limited number of untreated units are available.
- Should be excluded from the DP those units
  1. affected by the intervention of interest or of similar nature
  2. that suffered large idiosyncratic shocks in Y if these shocks would have NOT affected $j = 1$ if the treatment did not happen
  3. with different characteristics $\boldsymbol{X}$ to the treated unit (i.e. $j = 1$ and the DP should behave similarly over extended periods of time *prior to the intervention*) so to avoid
     - 3.1 interpolation biases
     - 3.2 **overfitting**, i.e. when the characteristics of $j = 1$ are artificially matched by combining idiosyncratic variations in the sample of js in the DP

Specifying the variables in $\mathbf{X}$

- Abadie et al. 2020 mentions that the credibility of a SC depends on its ability to track the pre-intervention trajectory of Y_1. So how do we choose what to include in in $\mathbf{X}_1$ and $\mathbf{X}_0$? We have some leeway
  - We can match the entire trajectory of Y_1 by including only the average of the Y_j in the DP in the pre-treatment period for
  - This is because the co-movement of Y across the js is *exactly what synthetic controls are designed to exploit*,
- The advantage from such *a* **summary** *of* Y in the pre-intervention period (*as opposed to including all annual values of* Y *as predictors*) to calculate the SC resides in a higher sparsity of the resulting SC
  - Sparse SC (that is, synthetic controls made of a small number of comparison units) are easy to interpret and evaluate
- Does this mean that we should only include pre-intervention outcomes and ignore the information of other predictors, $\mathbf{Z}_j$? No!
  - In $Y_{jt}^N = \delta_t + \boldsymbol{\theta}_t \mathbf{Z}_j + \boldsymbol{\lambda}_t \boldsymbol{\mu}_j + \varepsilon_{jt}$ covariates excluded from $\mathbf{Z}_j$ are mechanically absorbed into $\boldsymbol{\mu}_j$, which increases the number of components of $\boldsymbol{\mu}_j$ and, therefore, the **bound on the bias**

Requirements (1)

1. *Aggregate data on predictors and outcomes*
   - Specifically
     - 1.1 outcomes and predictors of the outcome for the unit or units exposed to the intervention of interest
     - 1.2 outcomes and predictors of the outcome for a set of comparison units
   - Sometimes, when aggregate data do not exist we can employ *aggregates* of micro data
2. *Sufficient* T_0: bias of the SC estimator is *bounded* by a function that is inversely proportional to T_0, during which the SC closely tracks the trajectory of the outcome variable for the affected unit
   - With a small T_0 close or even perfect fit of the predictor values for the treated unit may be spuriously attained, and so the resulting SC may fail to reproduce the trajectory of the outcome for $j = 1$ in the absence of the intervention.
   - The severity of this problem *can be diminished if* $\mathbf{X}$ *includes good predictors of post-intervention values of* Y_{jt} other than pre-intervention values of Y
   - However, a caveat is the possibility of **structural breaks**

Requirements (2)

1. **Sufficient post-intervention information**.
   - The evaluation data must include *outcome measures that*
     1.1 *are affected by the intervention*
     1.2 *are relevant for the policy decision that is the object of the study*.
   - This may be problematic if
     1.1 the effect of an intervention is expected to arise *gradually* over time
     1.2 no forward looking measures of the outcome are available
   - Extensive postintervention information allows a *more complete picture of the effects of the intervention*, in time and across the various outcomes of interest.

It is not recommended using this method when the pretreatment fit is poor or T_0 is small

Outline

1. Motivation
2. Formal aspects
   - Potential outcomes, econometric model and its bias
   - Comparison to regression and estimation of the SC
   - Inference, falsification and robustness
   - Limitations and recommendations for practitioners
3. Empirical application
   - The intervention in Abadie et al. (2010)
   - The command synth and examples
   - Replication of the paper
4. Current developments
   - How reasonable is the Convex Hull condition?
   - The problem with the Cross Validation method for V^*
   - How to correctly specify the model and avoid cherry picking

The Proposition 99

- In 1988, the **Proposition 99 in California** was the first modern-time large-scale tobacco control program in USA. This
  - Triggered a wave of local clean-air ordinances in California
  - Launched a new wave of state and federal anti-tobacco laws
  - Resulted in the tobacco industry response increase in its political activity in California at both the state and local levels.
- *Proposition 99 was widely perceived to have successfully cut smoking in California*. From the passage of Proposition 99 through 1999
  1. adult smoking prevalence fell in California by more than 30%,
  2. youth smoking levels dropped to the lowest in the country
  3. per capita cigarette consumption more than halved (California Department of Health Services 2006).
- Following early reports of California's success with Proposition 99, other states adopted similar policies.
  - As of April 20, 2009, 30 states, the DC, and 792 municipalities had laws requiring 100% smoke-free workplaces, bars, or restaurants (ANRF 2009).

Data and variables setting

- Proposition 99 went into effect in January 1989

$$\underbrace{1970, ..., 1988}_{\text{Pre-treatment}}; \underbrace{1989, ..., 2000}_{\text{Post-treatment}}$$

and so

  - we have 19 years of pre-intervention data.
  - we only take 10 year of post-intervention because at about this time anti-tobacco measures were implemented across many states, invalidating them as potential control units.

- Outcome is *cigsale*, annual per capita cigarette consumption at the state level, measured as per capita cigarette *sales* (in packs)
- Predictors of *cigsale* for California and the *j*s in DP take information *only on preintervention period* ◂ Matrices

1. $\overline{Y}^{\mathbf{K}_1}, ..., \overline{Y}^{\mathbf{K}_M}$: lagged (observed) smoking consumption in 1975, 1980, and 1988
2. $\boldsymbol{Z}$: averaged over the 1980–1988 period:
   - 2.1 average retail price of cigarettes
   - 2.2 per capita state personal income (logged)
   - 2.3 percentage of the population age 15–24
   - 2.4 per capita beer consumption

Choosing the Donor Pool

- The SC (California) is constructed as a *weighted average* of *j*s in the DP with weights chosen so that this best reproduces
  1. the values of a set of predictors of cigarette consumption in California in the pre-treatment period (before passage of Proposition 99)
  2. the outcome (*cigsale*) that would have been observed for California in absence of Proposition 99
- Hence, we need to *discard* from the DP those states that
  1. adopted some other large-scale anti tobacco program during any year: MA, AZ, OR, and FL
  2. raised cigarette taxes by 50c. or more over post-treatment period: AK, HI, MD, MI, NJ, NY, WA
  3. DC

  So, the DP includes the remaining 38 states $(J+1=38+1)$
- The treatment effect estimate obtained for California would be **attenuated** if *any* of the states in the DP that gets a weight in the SC increased unilaterally taxes that reduced smoking

The work plan

- Using the techniques described in Section 2, we
  1. Construct a SC for California that mirrors the values of the predictors of cigsale in California before the passage of Proposition 99.
  2. Estimate the effect of Proposition 99 on cigsale as the difference in cigarette consumption levels between California and its synthetic versions in the post-intervention years (after Proposition 99 was passed) following

$$\hat{\alpha}_{1t} = Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt}; t \in \{T_0 + 1, ..., T\} \text{ where } j = 2, ..., 38 \text{ and } 1 = \text{California}$$

  3. Perform **placebo studies** that confirm that our estimated effects for California are unusually large relative to the distribution of the estimate that we obtain when we apply the same analysis to the states in the DP

Outline

1. Motivation
2. Formal aspects
   - Potential outcomes, econometric model and its bias
   - Comparison to regression and estimation of the SC
   - Inference, falsification and robustness
   - Limitations and recommendations for practitioners
3. Empirical application
   - The intervention in Abadie et al. (2010)
   - The command synth and examples
   - Replication of the paper
4. Current developments
   - How reasonable is the Convex Hull condition?
   - The problem with the Cross Validation method for V^*
   - How to correctly specify the model and avoid cherry picking

`synth`: basic syntax and mandatory options

`synth` *depvar predictorvars*, ***trunit(#) trperiod(#)*** *[counit(numlist) xperiod(numlist) mspeperiod() resultsperiod() nested allopt unitnames(varname) figure keep(file) customV(numlist) optsettings]*

- Dataset must be a **balanced** panel after `xtset panelvar timevar`
  - *depvar* is the outcome variable.
  - *predictorvars* is the list of predictor variables.
- Mandatory options
  - *trunit(#)*: the unit *number* of the $j = 1$ as given in *panel id variable specified in* `xtset`
    - Only 1 number can be specified.
    - If the intervention affected several *js* we *combine these* and *then* treat them as 1
  - *trperiod(#)*: the time *period* when intervention occurred $(T_0 + 1)$ as given in *panel time variable* specified in `xtset`. Only 1 number can be specified.
- By default, *all predictor variables are averaged over the entire pre-intervention period*, which ranges
  - from the period earliest available in the panel time variable
  - to the period ***immediately prior*** to the beginning of the intervention (T_0)

synth: optional options for handling Xs

- *counit(numlist):* a list of *j*s in the DP the control units as given in the panel id variable
  - Should contain at least 2 *j*s in the DP
  - If *no* counit() is specified, the DP defaults to *all units available in the panel id variable* excluding the unit specified in *trunit()*
- *xperiod(numlist)*: a list of periods over which the predictor variables specified in predictorvars are averaged as given in the panel time variable
  - E.g. *xperiod(1980(1)1988)* indicates that predictor variables are averaged over all years from 1980, 1981,...,1988.
  - If *no* xperiod() is specified, *xperiod()* defaults to the entire pre-intervention period
  - Missing data points in the variables are ignored in the X matrix
- *figure*: produces a line plot with outcome trends for $j = 1$ and the SC for the years specified in *resultsperiod()*

`synth`: optional options for optimization

- *nested:*
  - `synth` uses a **data-driven regression based method** to obtain the variable weights contained in the V-matrix, which is *not* covered in Abadie and Gardeazabal [2003], Abadie et al. [2010] and Abadie et al. [2015].
    - This relies on a **constrained quadratic programming routine** that finds the best fitting *W*-weights conditional on the regression based *V*-matrix.
    - This procedure is fast and yields good results in terms of minimizing the MSPE
  - *nested* will lead to better performance using additional computing time
    - `synth` will embark on an optimization procedure that *searches among all (diagonal) positive semidefinite V-matrices and sets of W-weights for the best fitting convex combination of the control units*.
    - This takes as starting point the regression based V and produces convex combinations that achieve even lower MSPE
- *allopt*:
  - If nested is specified we can *also* specify *allopt* if we want to check that the minimum found is not a local one
  - This is done by running nested optimization 3 times using 3 different starting points (regression based V, equal V-weights, and using Stata's `ml search`) and returning the best of these 3.
  - This will take 3 times the amount of computing time compared to *nested*

Example 1

```
synth cigsale cigsale(1988) cigsale(1980) cigsale(1975)
beer(1984(1)1988) lnincome retprice age15to24, trunit(3)
trperiod(1989)
```

- Parameters
  - *trunit(3)* : the unit affected by the intervention ($j = 1$) is unit 3 (California). Since no *counit()* is specified, the DP defaults to the other 38 states in the dataset: $j = 1, 2, 4, ..., 39$
  - *trperiod(1989)* : the *first* year of the treatment ($T_0 + 1 = 1989$)
- Model: what we include in X ▸ X matrices
  - *cigsale(1988) cigsale(1980) cigsale(1975)* : the observed values of cigsale in those 3 years (Y)
  - *beer(1984(1)1988)*: the average valueof beer taking *only* 1984, 1985, 1986, 1987 and 1988
  - *lnincome retprice age15to24*: the average value of lnincome, retprice and age15to24 for the whole pre-treatment period (1970 to 1988) since no *xperiod()* is provided

Example 2

```
synth cigsale cigsale(1988 1980 1975) lnincome(1980 1985)
beer retprice age15to24, trunit(3) trperiod(1989) fig
```

- Parameters
  - *trunit(3)* : the unit affected by the intervention ($j = 1$) is unit 3 (California). Since no *counit()* is specified, the DP defaults to the other 38 states in the dataset: $j = 1, 2, 4, ..., 39$
  - *trperiod(1989)* : the *first* year of the treatment ($T_0 + 1 = 1989$)
- Model: what we include in X ▸ X matrices
  - *cigsale(1988 1980 1975)*: the *average value* of *cigsale* taking only 1998, 1980 and 1975. This is DIFFERENT from *cigsale(1988) cigsale(1980) cigsale(1975)*
  - *lnincome(1980 1985)*: the *average value* of *lnincome* taking only 1980 and 1985. This is DIFFERENT from *lnincome(1980) lnincome(1985)*
  - *beer*: the *average value* of *beer* for the whole pre-treatment period (1970 to 1988). But since *beer* has missings pre-1984, synth will inform about this for this variable and these missings are ignored in the averaging
  - *retprice age15to24*: the *average value* of *retprice* and *age15to24* for the whole pre-treatment period (1970 to 1988) since no *xperiod()* is provided

Example 3

```
synth cigsale cigsale(1970 1979) retprice age15to24,
trunit(33) counit(1(1)20) trperiod(1980)
resultsperiod(1970(1)1990) fig
```

- Parameters
  - *trunit(33)*: the unit affected by the intervention ($j = 1$) is unit no 33
  - *counit(1(1)20)*: since this is specified, the DP is restricted to *only* states no 1,2,...,20. Note how the treated state does not appear here
  - *trperiod(1989)*: the first year of the treatment
  - *resultsperiod(1970(1)1990)*: results are obtained for 1970,1971,...,1990
- Model: what we include in X ▸ X matrices
  - *cigsale(1970 1979)*: the *average value* of cigsale taking only 1970 and 1979. This is DIFFERENT from *cigsale(1970) cigsale(1979)*
  - *retprice age15to24*: the *average value* of retprice and age15to24 for the whole pre-treatment period (1970 to 1988) since no *xperiod()* is provided
- Note the equivalence of this with

```
keep if inrange(state, 1, 20) | state==33
synth cigsale cigsale(1970 1979) retprice age15to24, trunit(33) trperiod(1980)
resultsperiod(1970(1)1990) fig
```

Example 4

```
ssynth cigsale cigsale(1975 1988 1980) beer lnincome
retprice age15to24, trunit(3) trperiod(1989)
xperiod(1980(1)1988) nested
```

- Parameters
  - *trunit(3)* : the unit affected by the intervention ($j = 1$) is unit 3 (California). Since no *counit()* is specified, the DP defaults to the other 38 states in the dataset: $j = 1, 2, 4, ..., 39$
  - *trperiod(1989)* : the year of the treatment ($T_0 + 1 = 1989$)
- Model: what we include in X
  - *cigsale(1975 1988 1980)*: the *average value* of *cigsale* taking only 1998, 1980 and 1975. This is DIFFERENT from *cigsale(1988) cigsale(1980) cigsale(1975)*
  - *beer lnincome retprice age15to24*: average of beer (obviating its missings) lnincome, retprice and age15to24 for the whole pre-treatment period (1970 to 1988) since NO *xperiod()* is provided.
- By specifying *nested*, synth will do an optimize searching *among all* (diagonal) positive semidefinite V-matrices and sets of W-weights for the best fitting convex combination of the control units.

Outline

1. Motivation
2. Formal aspects
   - Potential outcomes, econometric model and its bias
   - Comparison to regression and estimation of the SC
   - Inference, falsification and robustness
   - Limitations and recommendations for practitioners
3. Empirical application
   - The intervention in Abadie et al. (2010)
   - The command synth and examples
   - Replication of the paper
4. Current developments
   - How reasonable is the Convex Hull condition?
   - The problem with the Cross Validation method for V^*
   - How to correctly specify the model and avoid cherry picking

Figure 1

- *Figure 1. Trends in per-capita cigarette sales: California vs. the rest of the United States.*
  - This shows how *bad* a simple average of cigsales for the 38 *j*s in the DP approximates cigsales for California
- So we need
  1. for *every year*, collapse into 1 the cigsales for the 38 *j*s in the DP
  2. plot it along the observed series for California

The command synth for this study

- To do what follows, we need to run the synth command to find $\boldsymbol{W}^*$

```
synth cigsale lnincome age15to24 retprice beer(1984(1)1988)
cigsale(1988) cigsale(1980) cigsale(1975), trunit(3)
trperiod(1989) xperiod(1980(1)1988)
resultsperiod(1970(1)2000) nested
```

where

- **Z**s
  - *beer(1984(1)1988)*: takes the average value of beer only between those years because this *has missings* before
  - *lnincome age15to24 retprice*: takes the average value of these variables over *only* 1980-1988 because we specified *xperiod(1980(1)1988)* which means that we no longer take the *whole pre-treatment period* (1970-1988).
- $\overline{Y}^{\mathbf{K}_1}, ..., \overline{Y}^{\mathbf{K}_M}$ *: cigsale(1975) cigsale(1980) cigsale(1988)*: takes the values of cigsale for onlu those years
- *resultsperiod(1970(1)2000)*: the latest year is because by 2001 many states had adopted similar policies

Table 2

- *Table 2. State weights in the synthetic California*
  - This is the main result of the SCM estimator: the SC weights
  - This is conveniently stored as a return matrix after the estimation of the command
    `matrix list e(W_weights)`
- Note how there are 38 entities in the DP
  - Yet, there are only 5 entities (states) who receive (positive) weights
  - This is a consequence of the 2 restrictions imposed on **W**, which yields **sparse synthetic controls**
  - In fact, this is a powerful feature of SCM since, precisely, this sparsity allows for easiness of interpretation and analysis

Table 1

- *Table 1. Cigarette sales predictor means*
  - This shows how *good* the weighted average of the $\boldsymbol{X}$*s* using the $w_1^*, w_2^*, w_4^*..., w_{39}^*$ for the 38 *j*s in the DP approximates $\boldsymbol{X}$*s* for California
  - This is important to do before we calculate the SC (using $\sum_{j=2}^{J+1} w_j^* Y_{jt}$)
- So
  1. for column 1: for California only (so only 1 entity)
     - 1.1 simple average over 1980-1988 for *lnincome*, *age15to24*, *retprice*
     - 1.2 simple average over 1984-1988 for *beer*
     - 1.3 observed values over 1975, 1980 and 1988 for *cigsales*
  2. for column 2: for the DP only (so 38 entities)
     - 2.1 weighted average (using the $\boldsymbol{W}^*$) over 1980-1988 for *lnincome*, *age15to24*, *retprice*
     - 2.2 weighted average (using the $\boldsymbol{W}^*$) over 1984-1988 for *beer*
     - 2.3 weighted average (using the $\boldsymbol{W}^*$) of observed values over 1975, 1980 and 1988 for *cigsales*
  3. for column 3: same as column 2 (for the DP only so 38 entities) but taking simple averages

Figure 2

- *Figure 2. Trends in per-capita cigarette sales: California vs. synthetic California*
  - This is where we plot the observed series for California (between 1970 and 2000) and the calculated SC using, for any $t = 1970, ..., 2000$:

$$\underbrace{\sum_{j=2}^{J+1} w_j^* Y_{jt}}_{\text{SC for California}} = \begin{pmatrix} 0 & 0 & \underbrace{0.147}_{\text{CO}} & \underbrace{0.059}_{\text{CT}} & 0 & \dots & \underbrace{0.205}_{\text{MT}} & 0 & \underbrace{0.242}_{\text{NV}} & \dots & 0 & \underbrace{0.346}_{\text{UT}} & \dots & 0 \end{pmatrix}_{1\times 38}$$

$$\times \begin{pmatrix} Y_{1t} & Y_{2t} & Y_{4t} & Y_{5t} & Y_{6t} & \dots & Y_{19t} & Y_{20t} & Y_{21t} & \dots & Y_{33t} & Y_{34t} & \dots & Y_{38t} \end{pmatrix}'_{38\times 1}$$

  - These are stored in

```
matrix list e(Y_treated)
matrix list e(Y_synthetic)
```

- We can calculate `e(Y_synthetic)` manually
  - Let's keep Y_{j1970} for all the js in the DP

```
keep if state!=3 & year==1970
br state year cigsale
```

  - Now let's generate a variable *weights*, with 0 for all the states other than those 5 which have positive weights in $\mathbf{W}^*$
  - Multiply this variable *weights* with *cigsale*
  - Sum the values: this is the value to e(Y_synthetic) for 1970!

Figure 3

- *Figure 3. Per-capita cigarette sales gap between California and synthetic California*
  - This is where we use our TE estimator $\hat{\alpha}_{1t}$ defined above, which equals the observed series for California minus the calculated SC using, for any $t = 1970, ..., 2000$:

$$\hat{\alpha}_{1t} = \underbrace{Y_{1t}}_{\text{observed California}} - \underbrace{\sum_{j=2}^{J+1} w_j^* Y_{jt}}_{\text{SC for California}} \; ; t \in \{T_0 + 1, ..., T\}$$

    where $1 =$ *California*
  - We have to create this manually (no big deal)
- We can calculate the TE manually

  `gen te=cigsale_cal - cigsale_scm`

  and then we plot this

Figure 4

- *Figure x. Per-capita cigarette sales gaps in California and placebo gaps in all 38 control states*
  - This is where we take the permutation tests (in-space placebos) to see the exact significance of our estimates
  - Figure 4 plots *all* the 38 permutations + the estimate for California (so there are 38 + 1 lines). This provides the "raw material" for the next 3 graphs
- This requires looping. For a given j (to simplify the coding we also include here California even though we already found its SC estimate)
  1. re-run the `synth` command we ran at the beginning but specifying `trunit(`j`)`
  2. save separately `e(Y_treated)`, `e(Y_synthetic)` and `e(RMSPE)` for this j
  3. accumulate these along with those of the previous j

  We will end up with a matrix, we save into stata via `svmat` and then we graph it

Figure 5, 6 and 7

- *Figure x. Per-capita cigarette sales gaps in California and placebo gaps in Z control states (discards states with pre-Proposition 99 MSPE X times higher than California's).*
  - Figure 5, 6 and 7 drop lines (states) with increasingly large MSPEs relative to that of California
  - This is nothing but figure 4 without some lines, those above a given threshold
- So,

1. We need the `e(RMSPE)` $= RMSPE_j^{\boxed{pre}}$ stored before to create

$$ratio = \frac{RMSPE_j^{pre}}{RMSPE_{\boxed{1}}^{pre}} = \frac{\sqrt{\left(\frac{1}{T_0}\sum_{t=1}^{T_0}\left(Y_{jt} - \sum_{j=1}^{J+1} w_k^* Y_{kt}\right)^2\right)}}{\sqrt{\left(\frac{1}{T_0}\sum_{t=1}^{T_0}\left(Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt}\right)^2\right)}} \quad \texttt{with } k = 1, ..., J+1 \texttt{ without the current } j$$

   where $1 =$ *California*. This will provide with 1 unique value of the ratio for the series of every state (38+1 unique values), equal to 1 *only* for California
2. Based on this we filter the dataset: we drop those series with a value of *ratio* larger than a threshold
   - 2.1 $ratio > 20$
   - 2.2 $ratio > 5$
   - 2.3 $ratio > 2$

Figure 8

- *Figure 8. Ratio of post-Proposition 99 MSPE and pre-Proposition 99 MSPE: California and 38 control states*
  - A final way to evaluate the California TE relative to the TEs of the placebo runs is looking at the distribution of *ratios* of post/pre- RMSPE.
  - The main advantage of looking at **ratios** is that it obviates choosing a cutoff for the exclusion of ill-fitting placebo runs.
- So,
  1. We use e(RMSPE)$=RMSPE_j^{pre}$ and *also* to calculate $RMSPE_j^{post}$

$$\frac{RMSPE_j^{post}}{RMSPE_j^{pre}} = \frac{\sqrt{\left(\frac{1}{T-T_0}\sum_{t=T_0+t}^{T}\left(Y_{jt}-\sum_{k=1}^{J+1}w_k^* Y_{kt}\right)^2\right)}}{\sqrt{\left(\frac{1}{T_0}\sum_{t=1}^{T_0}\left(Y_{jt}-\sum_{k=1}^{J+1}w_k^* Y_{kt}\right)^2\right)}} \quad \text{with } k=1,\ldots,J+1 \text{ without the current } j$$

     This will provide with 1 unique value of the ratio for the series of every state (39 unique values), equal to 1 only for California
  2. Based on this we graph the distribution of the post/pre-Proposition 99 ratios of the MSPE for California and all 38 control states.
  3. If we were to assign the intervention at random in the data, the probability of obtaining a post/pre-Proposition 99 MSPE ratio as large as California's is

$$\frac{1}{39} = 0.026 < 0.05$$

Outline

1. Motivation
2. Formal aspects
   - Potential outcomes, econometric model and its bias
   - Comparison to regression and estimation of the SC
   - Inference, falsification and robustness
   - Limitations and recommendations for practitioners
3. Empirical application
   - The intervention in Abadie et al. (2010)
   - The command synth and examples
   - Replication of the paper
4. Current developments
   - How reasonable is the Convex Hull condition?
   - The problem with the Cross Validation method for V^*
   - How to correctly specify the model and avoid cherry picking

Ferman & Pinto (2019): only stationary common factors

- Analyze the properties of the SC when the pre-treatment fit is imperfect, i.e. when *the CH assumption is violated* ◂ CH assump.
- When $T_0 \to \infty$, when the pre-treatment fit is imperfect in a model with $I(0)$ $\boldsymbol{\lambda}_t$, SC weights converge in probability to weights that *do not match the factor loadings* $\boldsymbol{\mu}$ *of* $j = 1$
- Intuition: if treatment assignment is correlated with **common factors** $\boldsymbol{\lambda}_t$ *in the post-treatment periods*, then *we would need a SC unit which*
  - *did not receive the treatment*
  - *is affected in exactly the same way by these* **common factors** $\boldsymbol{\lambda}_t$ *as the* $j = 1$

  this would be attained with *weights that reconstruct the factor loadings* $(\boldsymbol{\mu}_j)$ *of* $j = 1$. However, when the pre-treatment fit is imperfect
  - SC weights do *not* converge to weights that satisfy this condition
  - the distribution of the SC estimator will still depend on the $\boldsymbol{\lambda}_t$, *implying a biased estimator when selection depends on these* $\boldsymbol{\lambda}_t$
- They propose a **modified SC estimator** which demeans the data using information from the preintervention period

$$\hat{\alpha}_{0t}^{SC'} = y_{0t} - \mathbf{y}_t' \hat{\mathbf{w}}^{SC'} - \left(\bar{y}_0 - \bar{\mathbf{y}}' \hat{\mathbf{w}}^{SC'} \right)$$

under $E(\boldsymbol{\lambda}_t | D(0,0) = 1) = \boldsymbol{\omega}_0$ for $t > 0$ and stability conditions in the pre- and

Ferman & Pinto (2019): $I(0)$ and $I(1)$ common factors

- So far, λ_t is stationary; but what if we also have $\gamma_t \sim I(1)$ such that treatment assignment can be correlated with $\boldsymbol{\lambda}_t$ *and* γ_t in the post-intervention?
- Under the *additional* assumption on the **existence of weights** that reconstruct the $\boldsymbol{\mu}$ of unit 1 associated with the $\gamma_t \sim I(1)$**,** then *the asymptotic distribution $T_0 \to \infty$ of the **demeaned** SC estimator does not depend on the $I(1)$ common trends*
- Intuition: demeaned SC weights will converge to weights that reconstruct the $\boldsymbol{\mu}$s of $j = 1$ associated with the $I(1)$ $\boldsymbol{\lambda}_t$s. Then
  - $I(1)$ common factors will *not* lead to asymptotic bias
  - We only need caring about correlation between treatment and the $\boldsymbol{\lambda}_t$ which are $I(0)$
- We need checking **detrended** pre-treatment fit (eliminating $I(1)$ trends)
  - This implies subtracting from the outcome of the treated *and* control units the average of the *control units* only at t, $a_t = \frac{1}{J}\sum_{j \neq 1} y_{jt}$
  - This is indicative of potential bias from possible correlation between treatment assignment and $I(0)$ common factor
  - If pre-treatment fit with these series is *not as good as the original,* for a finite T_0, *a mismatch in $\boldsymbol{\mu}$s associated with $I(0)$ common factors might be relevant*

Outline

1. Motivation
2. Formal aspects
   - Potential outcomes, econometric model and its bias
   - Comparison to regression and estimation of the SC
   - Inference, falsification and robustness
   - Limitations and recommendations for practitioners
3. Empirical application
   - The intervention in Abadie et al. (2010)
   - The command synth and examples
   - Replication of the paper
4. Current developments
   - How reasonable is the Convex Hull condition?
   - The problem with the Cross Validation method for V^*
   - How to correctly specify the model and avoid cherry picking

Klößner, S. et al. (2015)

- Abadie et al. 2015 state that a way to find $\mathbf{V}^*$ consists in a Cross Validation approach.
  - We separate the pre-treatment period in a training and validation period
  - Using data from the training period we take any given V and find a preliminary $\check{W}$
  - Using data from the validation period we use this $\check{W}$ to find the V^*
  - Using data from the validation period we use this last V^* to find the W^*
- Yet, Klößner, S. et al. (2015) show that this method is not recommended for empirical practice
  - The reason for this is the non injectivity of the $\check{W}$ in the training period : the same $\check{W}$ which leads to very different $\mathbf{V}^*$s
  - This implies that we end up with different W^* depending on factors such as the ordering of the data!!
- Hence, when finding the $\mathbf{V}$ we need to rely on the regression based method or the nested approach

Outline

1. Motivation
2. Formal aspects
   - Potential outcomes, econometric model and its bias
   - Comparison to regression and estimation of the SC
   - Inference, falsification and robustness
   - Limitations and recommendations for practitioners
3. Empirical application
   - The intervention in Abadie et al. (2010)
   - The command synth and examples
   - Replication of the paper
4. Current developments
   - How reasonable is the Convex Hull condition?
   - The problem with the Cross Validation method for V^*
   - How to correctly specify the model and avoid cherry picking

Kaul et al. (2018)

1. If in $\boldsymbol{X}$ we use *all pre-treatment* Y along with $\mathbf{Z}$s, then these $\mathbf{Z}$s become irrelevant and the W^*will be calculated as if optimizes considering only Y
   - I.e. we get a SC based on
     - using economically meaningless covariates, or
     - not using covariates
   - This hold true regardless of using the RM or the nested approach to find V^*
2. Theoretically, doing this introduces a trade off
   2.1 we get an additional **small-sample bias** to the estimation*:* this is likely to be significant, especially when effect of the $\boldsymbol{Z}$s is large
   2.2 we take more care of unobserved confounders by fitting with respect to lagged outcomes alone
3. In MC experiment, they find that using all lags of the outcome in $\boldsymbol{X}$
   3.1 effectively ignores the covariates which leads to small-sample bias when estimating the outcome's counterfactual development
   3.2 makes estimates less precise in terms of RMSE compared to those which effectively use the covariates

Ferman et al. (2019)

- Still, theory rarely tells us about what covariates to include in $\mathbf{X}$
  - If different models result in different choices of the SC unit, *then a researcher would have relevant opportunities to select "statistically significant" specifications even when there is no effect.*
  - This flexibility may undermine the main advantage of the SC method, as it *implies some discretionary power for the researcher to construct the counterfactual for the treated unit*
- Ferman et al. show that the SC method *and* results using p-values in Abadie et al. (2010) are robust to specification searching only if
  1. T_0 is large
  2. we use models whose number of pre-treatment outcome lags go to ∞ with T_0

  otherwise, specification searching is a problem.
- They recommend
  1. Focusing on the specification that uses **all** the pre-treatment outcome lags unless there is a strong belief that it is *crucial* to *also* balance on specific $\mathbf{Z}$s
  2. Presenting results for different specifications; e.g. $\mathbf{x}_j = \left(Y_{j,1}, ..., Y_{j,\frac{3T_0}{4}}\right)'$ or $\mathbf{x}_j = \left(Y_{j,1}, ..., Y_{j,\frac{T_0}{2}}\right)'$

But remember that their analysis is **asymptotic**: as $T_0 \to \infty$!

END

- Thanks for your time :)

What is cluster sampling? (Wooldridge 2010, Chap 20)

- Most of the quasi experiments can be thought (see Abadie et al. 2017) in terms of **Cluster sampling**: *clusters or groups, rather than individuals, randomly drawn from a large population of clusters*
- E.g. in evaluating the impact an immigration shock on individual wages in Peru, one might sample departments from the entire country (as opposed to randomly drawing individuals from the population of Peru).
  - these departments *constitute the clusters*
  - the workers *within* the departments are the individual units.
- The cluster sampling scheme generally implies that
  1. the *outcomes of units within a cluster are correlated through unobserved* "**cluster effects**."
  2. some covariates, such as GDP, will be *perfectly correlated* because workers in the same region have the same GDP. Other covariates, such as education level, are likely to have substantial correlation but will vary within a region.
- So
  1. We have many clusters that can be assumed to be independent of each other
  2. Observations within a cluster are correlated
  3. Cluster samples are naturally unbalanced even without sample selection

Back

PTA (Lee 2016)

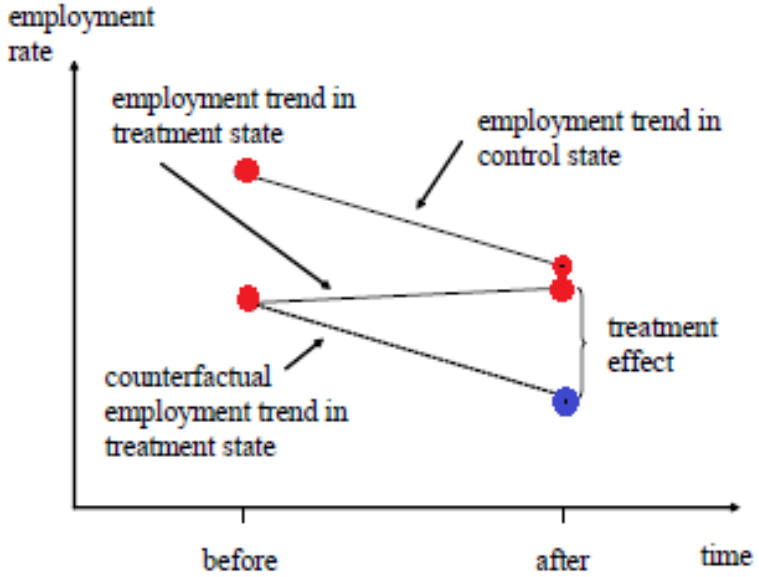


Figure 5.2.1: Causal effects in the differences-in-differences model

Source: A&P (2008)

- Remember, in the RCS case with 2 areas and 2 periods, where
  $S_i = 1$[individual i sampled at $t = 1$] is the treatment period and
  $Q_i = 1$[individual i is in the treatment group] is the treatment qualification
  - $Y^0_{t=0} = \beta_1 + \beta_q Q + \mathbf{W}_2' \beta_w + u_2$
  - $Y^0_{t=1} = Y^0_{t=0} + \underline{\beta_\tau} = \beta_1 + \underline{\beta_\tau} + \beta_q Q + \mathbf{W}_3' \beta_w + u_3$ (time effect β_τ added)
  - $Y^1_{t=1} = Y^0_{t=1} + \underline{\beta_d} = \beta_1 + \underline{\beta_\tau} + \underline{\beta_d} + \beta_q Q + \mathbf{W}_3' \beta_w + u_3$ (treatment effect β_d added)

PTA (Lee 2016)

- Then

$$Y_i = (1 - S_i)\, Y_{i,t=0} + S_i\, Y_{i,t=1} = (1 - S_i) \underbrace{Y^0_{i,t=0}}_{=Y_{i,t=0}} + S_i \underbrace{\left[(1 - Q_i)\, Y^0_{i,t=1} + Q_i\, Y^1_{i,t=1}\right]}_{=Y_{i,t=1}}$$

  after plugging and operating

$$Y_i = \beta_1 + \beta_\tau S_i + \beta_q Q_i + \boxed{\beta_d} \underbrace{S_i \times Q_i}_{=D_i} + \mathbf{W}_i \beta_w + u_i$$

  where $D_i = \begin{cases} 1 & \text{if } S_i = 1 \text{ and } Q_i = 1 \\ 0 & \text{otherwise} \end{cases}$, $(1 - S_i)\, W_2' + S_i\, W_3' \equiv W_i$ and $(1 - S_i)\, U_2 + S_i\, U_3 \equiv U_i$.

- In the case with more areas and periods, e.g. if treatment occurs in $t = 1$ and $t = 2$, we can add a dummy for each of these and
  - 1 interaction for the treatment $S_i \times Q_i$ (as seen) only
  - 1 interactions of the treatment qualification with a particular year
    - 1[individual i sampled at $t = 1$] $\times\, Q_i = 1$[individual i is in ANY of the treatment groups]
    - 1[individual i sampled at $t = 2$] $\times\, Q_i = 1$[individual i is in ANY of the treatment groups]

Back

Example of good covariate pre-treat. balance

- it is important to calculate the magnitude of $Y_{1t} - \sum_{j=2}^{J+1} w_j^* \boldsymbol{Z}_j = \boldsymbol{Z}_1$ between $j = 1$ and the SC to have an idea of the size of bias of the SC we could face

Table 1. Cigarette sales predictor means

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

NOTE: All variables except lagged cigarette sales are averaged for the 1980–1988 period (beer consumption is averaged 1984–1988). GDP per capita is measured in 1997 dollars, retail prices are measured in cents, beer consumption is measured in gallons, and cigarette sales are measured in packs.

Source: Abadie et al. (2010)

▸ Back

Example of good pre-treat. fit

- The ability of SC to reproduce the trajectory of the Y_1 over an extended period of time provides an indication of low bias
- An example of a good fit, with $T_0 = 1988$, is

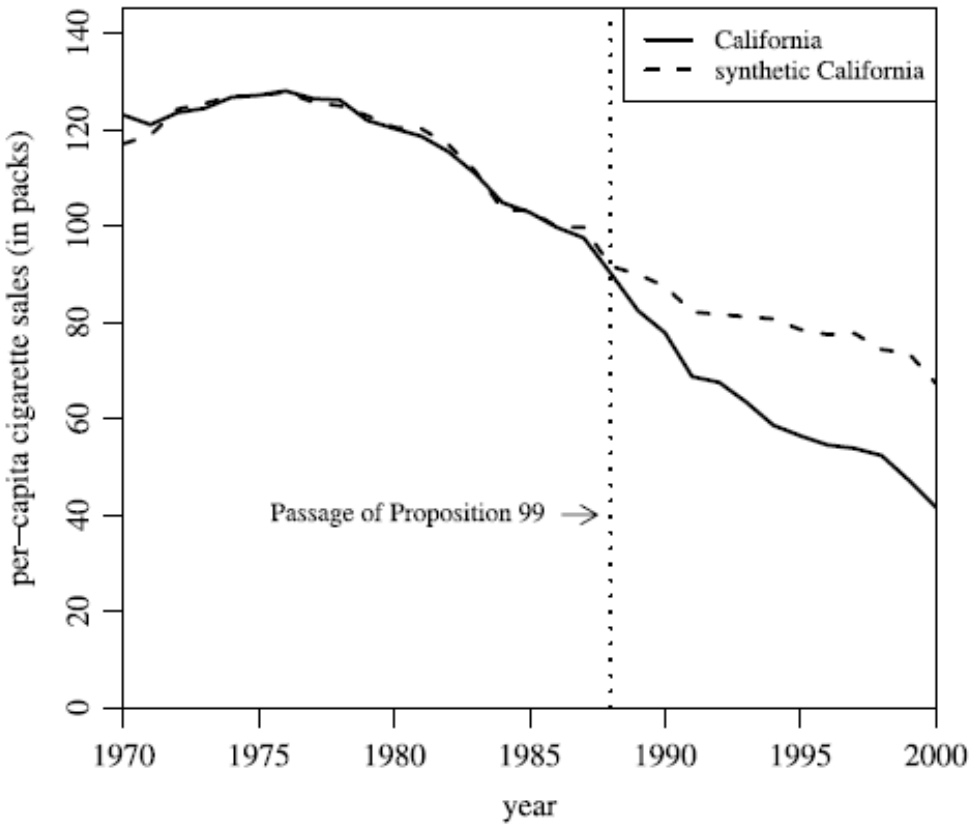


Figure 2. Trends in per-capita cigarette sales: California vs. synthetic California.

Source: Abadie et al. (2010)

Example of X matrices

- For example, if Y is hourly wage and we have yearly data 2001-2010 abd the treatment happens from 2007 $(= T_0 + 1)$ onwards in Lyc only
  - $\mathbf{X}_1$ has $r = 3$ and $M = 3$ rows and 1 column

	LyC	
informality rate	0.72	Average over 2001-2006
low skill workers	0.43	Average over 2001-2006
% agric. sector	0.2	Average over 2001-2006
hourly wage 2005	8.3	Observed
hourly wage 2003	7.5	Observed
hourly wage 2001	7.1	Observed

  - $\mathbf{X}_0$ has $r = 3$ and $M = 3$ rows and 23 columns

	Arequipa	Ayacucho	...	Tacna	
informality rate	0.6	0.58	...	0.8	Average over 2001-2006
low skill workers	0.3	0.75		0.65	Average over 2001-2006
% agric. sector	0.4	0.55		0.25	Average over 2001-2006
hourly wage 2005	8.9	5.3		6.2	Observed
hourly wage 2003	7.6	4.5		5.1	Observed
hourly wage 2001	6.5	4.4		4.8	Observed

▸ Back

Example of permutation test

- The solid line is the estimated TE for California and the gray line is the estimated TE for the other states in each placebo run
- If our TE is significative, we expect that after $T_0 + 1$ the black line is more extreme compared to the rest

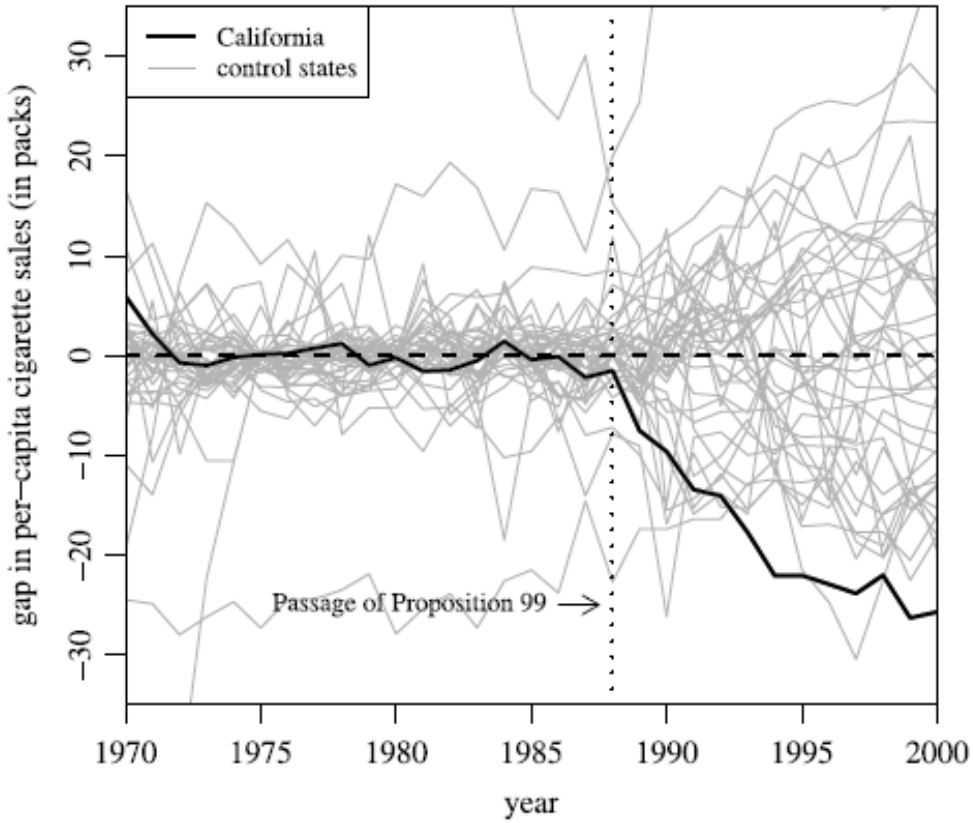
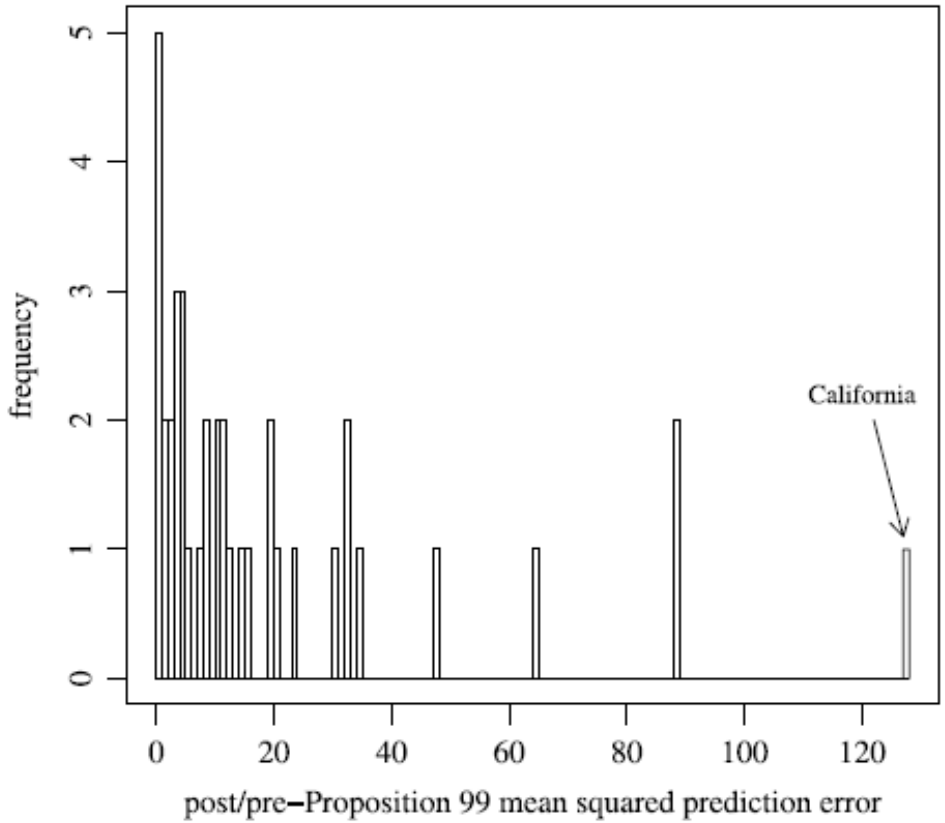


Figure 4. Per-capita cigarette sales gaps in California and placebo

Example of p-value in SCM

- Each bar is the ratio of MSPE for Califonia and for each placebo run
- If our TE is significative, we expect that the ratio for California very extreme



Source: Abadie et al. (2010)

Back

Example 1 matrices

- The matrix for $(X_1)_{(r+M=)7\times1}$ has both $(\mathbf{Z}_1)_{(r=)4\times1}$ and $\left(\overline{\mathbf{Y}}_1^{K_1},...,\overline{\mathbf{Y}}_1^{K_m}\right)_{(M=)3\times1}$

	3	
lnincome	10.03	Average over 1970-88
age15to24	0.18	Average over 1970-88
retprice	66.64	Average over 1970-88
beer(1984(1)1988)	24.28	Average over 1984-88
cigsale(1975)	127.10	Observed
cigsale(1980)	120.20	Observed
cigsale(1988)	90.10	Observed

$$\boldsymbol{X}_1 = \begin{pmatrix} \dfrac{Z_1}{\overline{Y}_1^{K_1}} \\ \cdots \\ \overline{Y}_1^{K_M} \end{pmatrix}_{k\times1=(r+M)\times1}$$

- The matrix for $(X_0)_{(r+M=)7\times38}$ has both $(\mathbf{Z}_0)_{(r=)4\times38}$ and $\left(\overline{\mathbf{Y}}_0^{K_1},...,\overline{\mathbf{Y}}_0^{K_m}\right)_{(M=)3\times38}$

	1	2	4		38	39	
lnincome	9.63	9.61	9.93	...	9.85	9.90	Average over 1970-88
age15to24	0.18	0.17	0.18	...	0.18	0.18	Average over 1970-88
retprice	66.99	67.69	60.39	...	69.88	59.38	Average over 1970-88
beer(1984(1)1988)	18.96	18.52	25.08	...	32.04	24.98	Average over 1984-88
cigsale(1975)	111.70	114.80	131.00	...	113.50	160.70	Observed
cigsale(1980)	123.20	131.80	131.00	...	117.60	158.10	Observed
cigsale(1988)	112.10	121.50	94.60	...	102.60	114.30	Observed

$$\boldsymbol{X}_0 = \begin{pmatrix} \dfrac{Z_2}{\overline{Y}_2^{K_1}} & \dfrac{Z_3}{\overline{Y}_3^{K_1}} & \cdots & \dfrac{Z_{J+1}}{\overline{Y}_{J+1}^{K_1}} \\ \cdots & \cdots & \cdots & \cdots \\ \overline{Y}_2^{K_M} & \overline{Y}_3^{K_M} & \cdots & \overline{Y}_{J+1}^{K_M} \end{pmatrix}_{k\times J=(r+M}$$

▸ Back

Example 2 matrices

- The matrix for $(X_1)_{(r+M=)5\times 1}$ has *only* $(\mathbf{Z}_1)_{(r=)5\times 1}$

	3	
cigsale(1988 1980 1975)	112.27	Average over 1988, 1980 and 1975
lnincome	10.03	Average over 1970-88
age15to24	0.18	Average over 1970-88
retprice	66.64	Average over 1970-88
beer(1984(1)1988)	24.28	Average over 1984-88

$$\boldsymbol{X}_1 = (\boldsymbol{Z}_1)_{k\times 1=r\times 1}$$

- The matrix for $(X_0)_{(r+M=)5\times 38}$ has *only* $(\mathbf{Z}_0)_{(r=)5\times 38}$

	1	2	4		38	39	
cigsale(1988 1980 1975)	115.67	122.70	118.87		111.23	144.37	Average over 1988, 1980 and 1975
lnincome	9.63	9.61	9.93	...	9.85	9.90	Average over 1970-88
age15to24	0.18	0.17	0.18	...	0.18	0.18	Average over 1970-88
retprice	66.99	67.69	60.39	...	69.88	59.38	Average over 1970-88
beer(1984(1)1988)	18.96	18.52	25.08	...	32.04	24.98	Average over 1984-88

$$\boldsymbol{X}_0 = (\boldsymbol{Z}_0)_{k\times 1=r\times 1}$$

▸ Back

Example 3 matrices

- The matrix for $(X_1)_{(r+M=)3\times1}$ has *only* $(\mathbf{Z}_1)_{(r=)3\times1}$

	33	
cigsale(1970 1979)	129.10	Average over 1970 and 1979
age15to24	0.18	Average over 1970-88
retprice	66.64	Average over 1970-88

$$\boldsymbol{X}_1 = (\boldsymbol{Z}_1)_{k\times1=r\times1}$$

- The matrix for $(X_0)_{(r+M=)3\times20}$ has *only* $(\mathbf{Z}_0)_{(r=)3\times20}$

	1	**2**	**3**		**19**	**20**	
cigsale(1970 1979)	113.80	111.55	135.81	...	131.50	125.90	Average over 1970 and 1979 and 1975
age15to24	0.18	0.17	0.18	...	0.17	0.17	Average over 1970-88
retprice	66.99	67.69	60.39	...	63.36	67.92	Average over 1970-88

$$\boldsymbol{X}_0 = (\boldsymbol{Z}_0)_{k\times1=r\times1}$$

Back