Bayesian and Frequentist Inference for Synthetic Controls

Jaume Vives-i-Bastida (MIT) and Ignacio Martinez (Google)

EEA-ESEM 2023

Overview

Goal: inference for SCs in linear factor model frameworks

This paper:

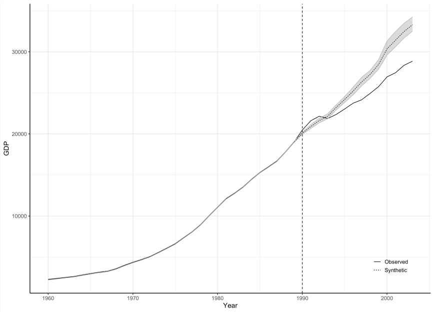
1. For a simple factor model, answer:
    - What **parameters** are we targeting?
    - What are we **identifying** as the number of donor units grows?
    - Under which conditions can we **estimate** them?
    - Under what conditions is the target parameter a "**synthetic control**"?
2. Provide a Bayesian alternative (**bsynth**) to SC and derive a **BvM** result

Today:

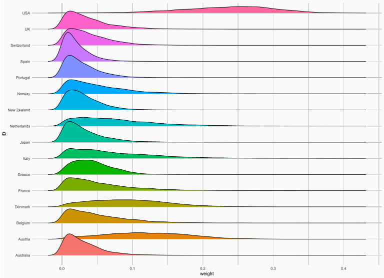
- Identification results
- (pseudo)-MLE with growing parameters
- **New!** Bayes estimator for SC $\implies$ Credible Intervals
- BvM (Bayes $\sim$ Frequentist)
- Unification and Secession

bsynth package for ADH 2015

```
synth <- bayesianSynth$new(data = germany,
                           time = year,
                           id = country,
                           treated = D,
                           outcome = gdp_pc)
```



(a) Treatment effect



(b) Implicit weight marginals

Why use bsynth?

1. **Simple** to implement in R!
2. Bayesian **inference** with few data points (**intervals!**)
3. Can easily incorporate **priors** on unit weights
4. Can approximate frequentist inference under some conditions (**BvM**)
5. Gives you the full posterior distribution, implicit weight distribution, correlations between weights etc.

Inference for Synthetic Controls

1. **Permutation Inference** (Abadie et al. 2010, Firpo and Possebom 2018, Abadie 2020)
2. **Projection Theory** for ATE (Li 2020, Hsiao et al. 2012)
3. **Conformal** Inference (Chernozhukov et al. 2021a)
4. **Large sample** properties in factor models (Ferman 2021, Ferman and Pinto 2019)
5. **Bayesian** inference (Pang, Liu, and Xu 2020, Arbour et al. 2021)

$\Longrightarrow$ Literature requires conditions on the weight vector w

- Good pre-treatment fit requirement.
- There exists a true w^*
- Sequence of w that gets diluted as $J \to \infty$

Today: re-write w in terms of the **factor model** and derive conditions on the **factor loadings**

Toy Factor Model I

Based on: Hsiao et al. 2012 and Ferman 2021

T_0 time periods, $J+1$ units

Potential outcomes are given by

$$Y_{it}(0) = \boldsymbol{\lambda}_i' \mathsf{F}_t + \epsilon_{it},$$
$$Y_{it}(1) = \tau_{it} + Y_{it}(0).$$

$\Longrightarrow$ Only unit 1 gets treated after T_0.

Toy Factor Model II

Target parameter (ATET):

$$\tau_{1T_0+1} = Y_{1T_0+1} - \underbrace{Y_{1T_0+1}(0)}_{\textit{unobserved}}$$

Estimators: based on observations $\boldsymbol{y}_{JT_0+1}$. Then, for $\boldsymbol{w} \in \mathbb{R}^J$

$$\hat{Y}_{1T_0+1}(0) = \boldsymbol{w}'\boldsymbol{y}_{JT_0+1}$$

Simplifying assumptions:

(A1) – *factors*

(a) we have only one factor such that $\lambda_i, F_t \in \mathbb{R}$

(b) $F_t \sim_{i.i.d} N(0, \sigma^2)$

(A2) – *idiosyncratic shocks*

(a) $\epsilon_{it} \sim_{i.i.d} N(0, 1)$

Conditional Distribution

Under **A1-A2** the conditional distribution of Y_{1t} given realizations $\boldsymbol{y}_{Jt}$ is

$$Y_{1t}|\boldsymbol{Y}_{Jt} = \boldsymbol{y}_{Jt} \sim N\left(\tilde{\mu}, \tilde{\Sigma}\right),$$

where

$$\tilde{\mu} = \sum_{j=2}^{J+1} \tilde{w}_j(\boldsymbol{\lambda}, \sigma) y_{jt},$$

$$\tilde{\Sigma} = 1 + \lambda_1 \sigma^2 (1 - \sum_{j=2}^{J+1} \tilde{w}_j(\boldsymbol{\lambda}, \sigma) \lambda_j), \text{ and}$$

$$\tilde{w}_j(\boldsymbol{\lambda}, \sigma) = \frac{\sigma^2 \lambda_1 \lambda_j}{1 + \sum_{j=2}^{J+1} \lambda_j^2 \sigma^2}$$

derivations

Target Parameter I

- The $\tilde{w}$ weights minimize the **statistical risk**.

Theorem (Linear Predictors)

*Under assumptions **A1-A2** it follows that*

$$\tilde{w} \in argmin_w \mathbb{E}\left[(Y_1(0) - y_J' w)' V (Y_1(0) - y_J' w)\right],$$

for any positive semi-definite matrix V.

Target Parameter II

What parameters of the factor model can we recover?

$$Y_{1T_0+1}(0) = \underbrace{\lambda_1 F_{T_0+1}}_{\text{predictive part}} + \underbrace{\epsilon_{iT_0+1}}_{\text{new shock}}$$

Theorem (Predictor convergence)

*Given **A1-A2**, if*

$$\frac{1}{\|\boldsymbol{\lambda}_J\|_2^2} \sum_j |\lambda_j| \to 0$$

as $J \to \infty$, then

$$\mathbf{y}'_{JT_0+1}\tilde{\mathbf{w}} \xrightarrow{p} \lambda_1 F_{T_0+1}$$

Target Parameter III

Convergence in probability requires a **density condition**:

$$\frac{1}{\|\boldsymbol{\lambda}_J\|_2^2} \sum_j |\lambda_j| \to 0$$

- Implies that $\|\tilde{\boldsymbol{w}}_J\|_2^2 \to 0$ as $J \to \infty$ (Ferman 2021).
- Implies that we recover the treated unit factor loading:

$$\sum_j \tilde{w}_j \lambda_j = \frac{\sigma^2 \lambda_1 \|\boldsymbol{\lambda}_J\|_2^2}{1 + \sigma^2 \|\boldsymbol{\lambda}_J\|_2^2} \to \lambda_1.$$

When is $\tilde{w}$ a synthetic control?

$$\tilde{w} \in \Delta^J = \{w | w \geq 0, \sum_j w_j = 1\}$$

Theorem (Synthetic Control Characterization I)

*For fixed J under **A1-A2**, $\tilde{w} \in \Delta^J$ iff the following conditions hold*

1. $sign(\lambda_1) = sign(\lambda_j)$ *for all j,*
2. $\sum_j \lambda_j^2 - \lambda_1 \sum_j \lambda_j + \frac{1}{\sigma^2} = 0.$

*Furthermore, for a fixed λ_1, as $J \to \infty$ if $\frac{1}{\|\boldsymbol{\lambda}_J\|_2^2} \sum_j |\lambda_j| \to 0$ then there exist **no sequences** $\{\lambda_j\}$ for which (2) and (1) hold simultaneously.*

Target Parameter V

- If λ_1 is fixed, at the limit the SC will be **biased**.
- If we let λ_1 be a function of the λ_j we can reconcile the condition.

Theorem (Synthetic Control Characterization II)

Given the previous theorem's assumptions, there exist conditions on λ_1 such that $\tilde{\boldsymbol{w}} \in \boldsymbol{\Delta}^J$. In particular, our conditions are implied by

$$\lambda_1 \in \boldsymbol{\Delta}(\boldsymbol{\lambda}_J)$$

So, we recover the sufficient conditions of Ferman 2021.

Recap

- The **target weights** are the linear CEF.
- Under some conditions we can recover the **predictive part** as $J \to \infty$.
- The set of distributions s.t. we can do so with **SC** is **small** but **non-empty**.

Next $\implies$ Inference!

Goal: estimate $\tilde{\boldsymbol{w}}_J$ using a data set of pre-treatment outcomes $\{y_{1t}(0), \boldsymbol{y}_{Jt}(0)\}_{t=1}^{T_0}$.

Log-likelihood for parameter $\boldsymbol{\theta} = (\boldsymbol{w}, \Sigma)$:

$$l_{T_0}(\boldsymbol{\theta}) = -\frac{1}{2}\log(2\pi\Sigma) - \frac{1}{T_0}\sum_{t=1}^{T_0}\frac{1}{2\Sigma}\left(y_{1t} - \sum_{j=2}^{J+1} w_j y_{jt}\right)^2.$$

- For fixed J, like standard MLE.
- But we need $J \to \infty$!

Theorem (MLE with growing J)

*Let $\hat{\boldsymbol{\theta}}_{MLE} \in argmax_{\boldsymbol{\theta}} l_{T_0}(\boldsymbol{\theta} \in \Theta)$ for a compact parameter space Θ, then under **A1-A2** and λ_j are uniformly bounded:*

1. $\frac{1}{T_0}\sum_t \mathbf{y}_{Jt}\mathbf{y}_{Jt}' = D_{T_0}$ *where* $0 < lim\ inf_{T_0}\sigma_{min}(D_{T_0}) \leq lim\ sup_{T_0}\sigma_{max}(D_{T_0}) < \infty$,
2. $max_{t \leq T_0}\|\mathbf{y}_{Jt}\|_2^2 = O_p(J)$,
3. $\sup_{\boldsymbol{\beta},\boldsymbol{\gamma} \in \mathcal{S}_J(1)} \sum_t |\mathbf{y}_{Jt}'\boldsymbol{\beta}|^2 |\mathbf{y}_{Jt}'\boldsymbol{\gamma}|^2 = O_p(T_0)$.

Then, it follows that if $o(T_0) = J(\log J)^3$

$$\|\hat{\mathbf{w}}_{MLE} - \tilde{\mathbf{w}}\|_2^2 = O_p(J/T_0).$$

If $o(T_0) = J^2 \log(J)$ then

$$\sqrt{T_0}\boldsymbol{\alpha}'(\hat{\mathbf{w}}_{MLE} - \tilde{\mathbf{w}})/\sigma_\alpha \xrightarrow{d} N(0,1),$$

for any $\boldsymbol{\alpha} \in \mathbb{R}^J$ and

$$\sigma_\alpha^2 = (\mathbb{E}[\epsilon_{Jt}^2])^{-1}\boldsymbol{\alpha}' D_{T_0}^{-1}\boldsymbol{\alpha}.$$

Inference: MLE II

Corollary

Under the conditions of the previous theorem, as $J, T_0 \to \infty$:

1. *If $o(T_0) = J(\log J)^3$ and $\frac{1}{\|\boldsymbol{\lambda}_j\|_2^2} \sum_j |\lambda_j| \to 0$, then*

$$\boldsymbol{y}'_{JT_0+1} \hat{\boldsymbol{w}}_{MLE} \xrightarrow{p} \lambda_1 F_{T_0+1}.$$

2. *If $o(T_0) = J^2 \log(J)$ and $\frac{1}{\|\boldsymbol{\lambda}_j\|_2^2} \sum_j |\lambda_j| \to 0$, then*

$$\sqrt{T_0}(\boldsymbol{y}'_{JT_0+1} \hat{\boldsymbol{w}}_{MLE} - \lambda_1 F_{T_0+1})/\sigma_{y_{JT_0+1}} \xrightarrow{d} N(0,1).$$

- Intuition: T_0 has to grow faster than J (quite faster).
- Based on methods by He and Shao 1996, 2000 and Bai and Wu 1994 (JMA).

Consider the following **Bayesian** model:

$$y_{1t}|\boldsymbol{y}_{Jt}, \boldsymbol{w}, \sigma_y \sim N(\boldsymbol{y}'_{Jt}\boldsymbol{w}, \sigma_y^2),$$
$$w_j|\boldsymbol{y}_{Jt} \sim N(\mu_j, \tau_j^2).$$

Bayes estimator

$$\hat{w}_j^B = \mathbb{E}_B[w_j|\boldsymbol{y}_t] = \int w_j p(w_j|\boldsymbol{y}_t) dw_j.$$

Then, the predictive posterior distribution is normal with

1. Mean:
$$\hat{Y}_{1t}^B = \boldsymbol{y}'_{Jt}\mathbb{E}_B[w_j|\boldsymbol{y}_t] = \frac{\sigma_y^2}{\sigma_y^2 + \sum_j \tau_j^2}\boldsymbol{y}'_{Jt}\mu_J + \frac{\sum_j \tau_j^2}{\sigma_y^2 + \sum_j \tau_j^2} y_{1t}.$$
2. Variance:
$$\mathbb{V}_B(\boldsymbol{y}'_{Jt}\boldsymbol{w}|\boldsymbol{y}_t) = \frac{\sigma_y^2 \sum_j \tau_j^2}{\sigma_y^2 + \sum_j \tau_j^2}.$$

Dirichlet prior: $\mu_j \sim Dir(1)$

Theorem (BvM)

*Under **A1-A2**, the assumptions the Corollary and*

1. ***Prior conditions**: $\|\mu_J\|_2^2 \to 0$, $\{\tau_j\}$ such that $\sum_j \tau_j^2 = O(J^\alpha)$, for $0 < \alpha < 1$, as $J \to \infty$, and $\sigma_y \to 1$.*
2. ***Convex recovery**: $\|\lambda_1 - \boldsymbol{\lambda}_J' \boldsymbol{\mu}_J\|_2 \to 0$ as $J \to \infty$.*

Then, as $T, J \to \infty$ at rate $o(T_0) = J^2 \log(J)$,

$$\boldsymbol{y}'_{JT_0+1} \mathbb{E}_B[\boldsymbol{w}|\boldsymbol{y}_{T_0}] \xrightarrow{p} \lambda_1 F_{T_0+1},$$

and

$$\|\Phi^{MLE}_{T_0,J} - Q_{T_0,J}\|_{TV} \to 0,$$

where $\Phi^{MLE}_{T_0+1,J}$ denotes the MLE finite sample distribution and $Q_{T_0+1,J}$ the Bayes posterior predictive distribution.

1. We derived conditions on the **factor loadings** such that SC recovers the target parameter.
2. In general, the set of such DGPs may be small, but **intuitive sufficient conditions exist**.
3. Inference through **pseudo-MLE**.
4. Conditions exist for Bayesian SC to converge to frequentist in TV (**BvM**).

Grouped Linear Factor Model (as in Ferman and Pinto 2018)

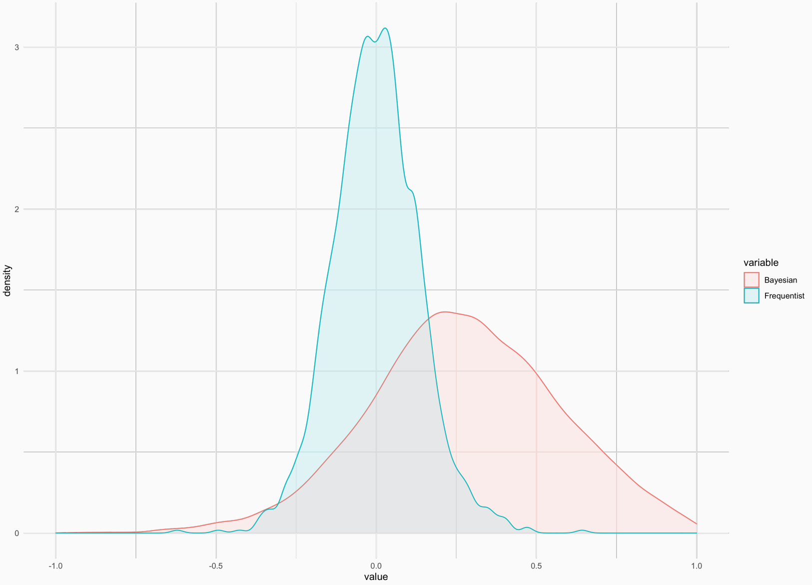
$$y_{it}(0) = \lambda_{f(i)t} + \epsilon_{it}.$$

- λ_{ft} follow an $AR(1)$ with $\rho = 0.5$ and standard Gaussian innovations.
- $\epsilon_{it} \sim N(0, \sigma^2)$ with $\sigma = 0.25$.
- Only unit 1 is treated, but treatment effect is 0.
- $f(1) = f(2)$ so unit 2 is the unbiased synthetic control.
- Fix $T_0 = T - 10$ and take $T \to \infty$.
- $J = 20$.

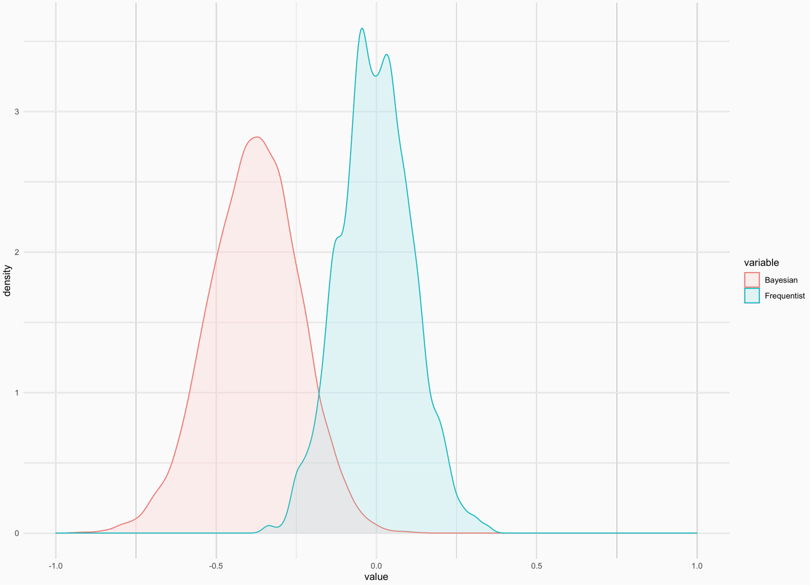
Simulation of $\hat{\tau}_1 = \frac{1}{T-T_0} \sum_t \hat{\tau}_{1t}$

1. Distribution over 10000 draws of the frequentist SC.
2. Bayesian posterior distribution (MCMC).

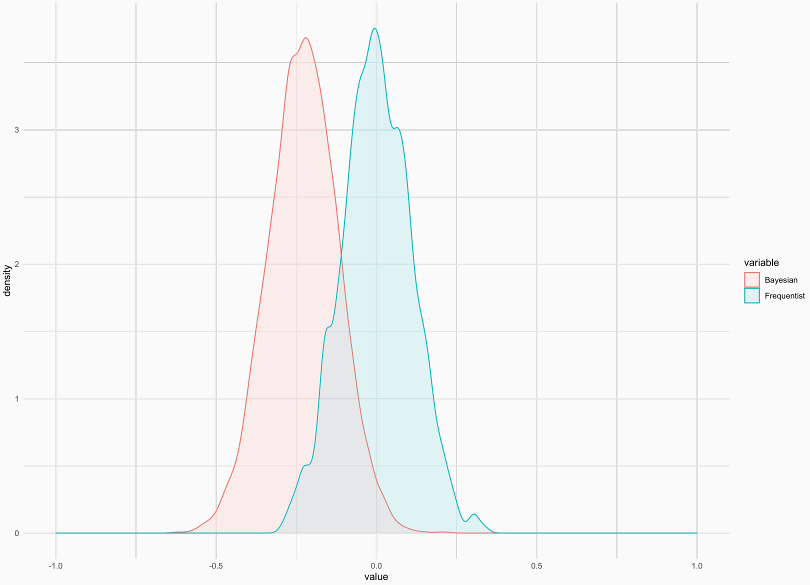
Simulation Evidence as $T \to \infty$



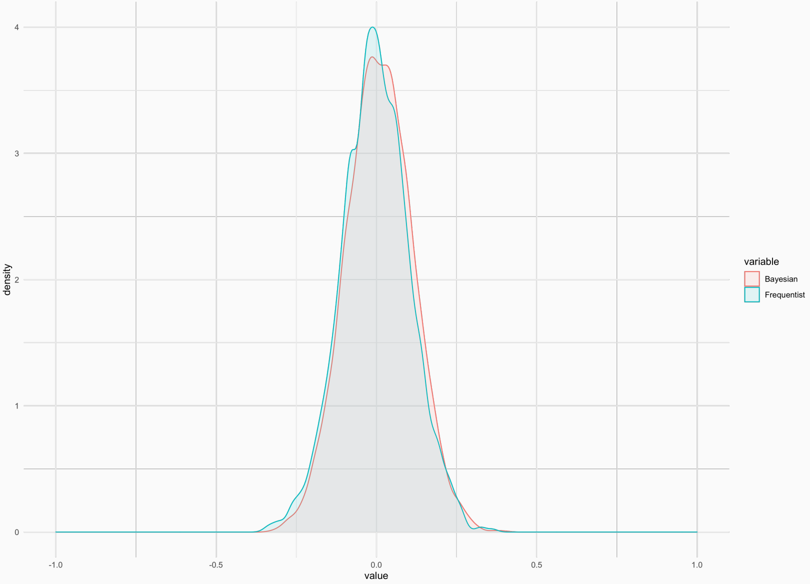
Simulation Evidence $T \to \infty$



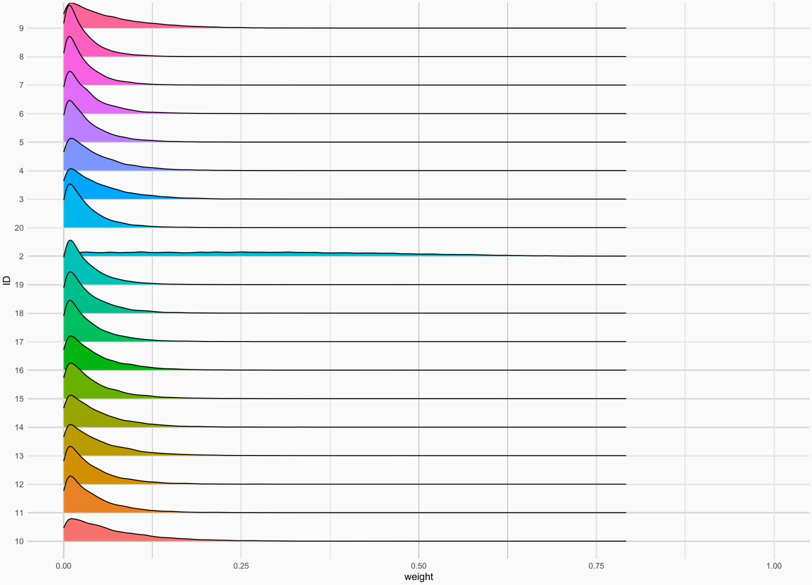
Simulation Evidence $T \to \infty$



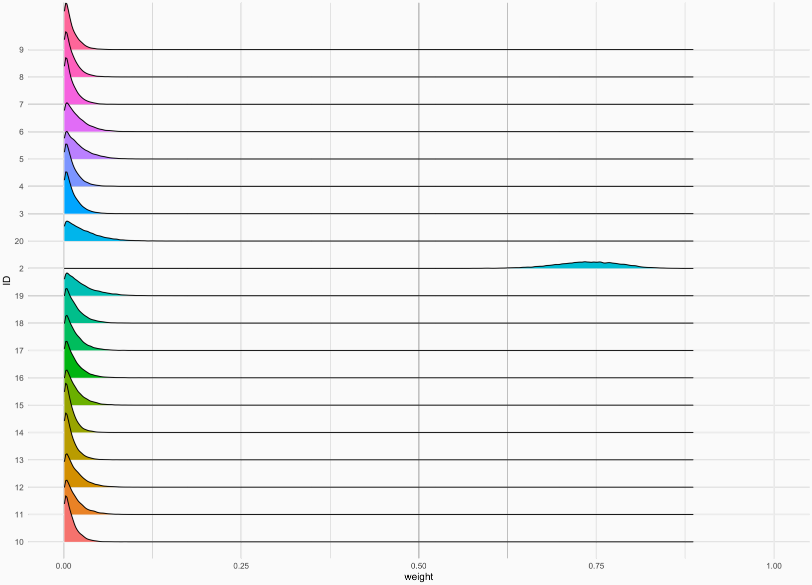
Simulation Evidence $T \to \infty$



Implicit Weights $T \to \infty$



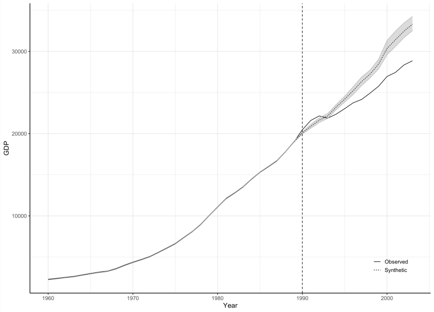
Implicit Weights $T \to \infty$



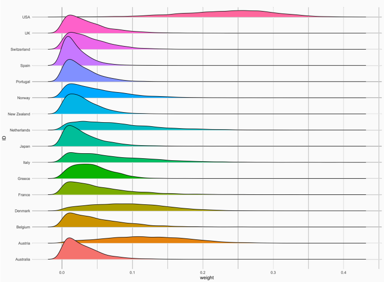
Implicit Weights $T \to \infty$

Implicit Weights $T \to \infty$

- Implementation of Bayesian model in BSYNTH R-package
- Results for German re-unification very similar to standard SC



(a) Treatment effect

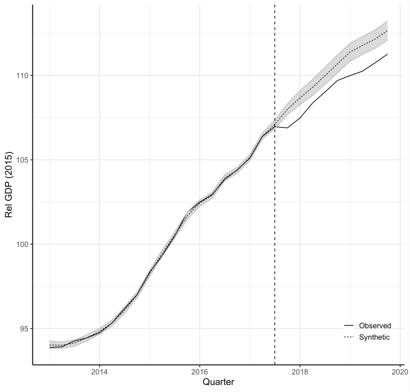


(b) Implicit weight marginals

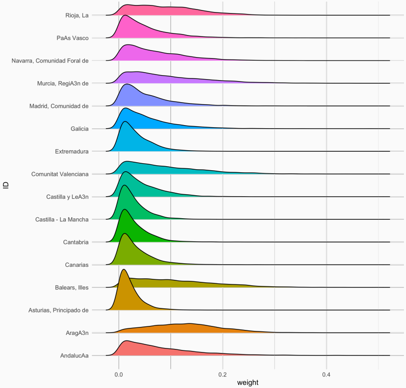
Figure 2: Bayesian synthetic control for West Germany

Catalan UDI

- We find that the UDI lead to a 0.3%-1.6% decrease in GDP.



(a) Treatment effect



(b) Implicit weight marginals

Figure 3: Bayesian synthetic control for Catalonia

1. When are the target parameters synthetic controls under simple factor model settings? **density conditions**
2. How can we do inference as $J, T_0 \to \infty$? **pseudo-MLE**
3. Can we use a Bayesian procedure to approximate the frequentist SC? **yes**

Method:

1. *bsynth* R-package can estimate different models (GP) and offers post-estimation functions
2. Application to the German re-unification and the Catalan UDI

Paper available at: **https://arxiv.org/abs/2206.01779**

Thanks!

$$\Sigma_{(2,J+1)} = \begin{pmatrix} 1+\lambda_2^2\sigma^2 & \lambda_2\lambda_3\sigma^2 & \cdots & \lambda_2\lambda_{J+1}\sigma^2 \\ \lambda_2\lambda_3\sigma^2 & 1+\lambda_3^2\sigma^2 & \cdots & \lambda_3\lambda_{J+1}\sigma^2 \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_2\lambda_{J+1}\sigma^2 & \lambda_3\lambda_{J+1}\sigma^2 & \cdots & 1+\lambda_{J+1}^2\sigma^2 \end{pmatrix} = \sum_{j=2}^{J+1} s_j \boldsymbol{u}_j \boldsymbol{u}_j^T,$$

where s_j is the eigenvalue associated with the $\boldsymbol{u}_j$ eigenvector. Observe that the eigenvalues are given by $s_2 = \cdots = s_J = 1$ and $s_{J+1} = 1 + \sum_{j=2}^{J+1} \lambda_j^2\sigma^2$.

$$\tilde{\mu} = \sigma^2\lambda_1 \sum_{j=2}^{J+1} \sum_{i=2}^{J+1} \lambda_i [\Sigma_{(2,J+1)}^{-1}]_{ji} y_j$$

$$\begin{aligned} w_j(\boldsymbol{\lambda}, \sigma) &= \sigma^2\lambda_1 \sum_{i=2}^{J+1} \lambda_i \sum_{k=2}^{J+1} \frac{1}{s_k} [\boldsymbol{u}_k \boldsymbol{u}_k^T]_{ji} \\ &= \frac{\sigma^2\lambda_1\lambda_j}{1 + \sum_{j=2}^{J+1} \lambda_j^2\sigma^2}. \end{aligned}$$

Focus on the first assumption:

$$\frac{1}{T_0}\sum_t \boldsymbol{y}_{Jt}\boldsymbol{y}_{Jt}' = D_{T_0},$$

where $0 < \lim\inf_{T_0}\sigma_{min}(D_{T_0}) \leq \lim\sup_{T_0}\sigma_{max}(D_{T_0}) < \infty$. Then, under the other assumptions:

$$\left|\boldsymbol{\alpha}'\left(\sum_t \mathbb{E}((y_{1t} - \boldsymbol{y}_{Jt}'\hat{\boldsymbol{w}}_{MLE}) - (y_{1t} - \boldsymbol{y}_{Jt}'\tilde{\boldsymbol{w}}))\boldsymbol{y}_{Jt}\right) - T_0\boldsymbol{\alpha}'D_{T_0}(\hat{\boldsymbol{w}}_{MLE} - \tilde{\boldsymbol{w}})\right| \leq c\sum_t |\boldsymbol{y}_{Jt}'\boldsymbol{\alpha}||\boldsymbol{y}_{Jt}'(\hat{\boldsymbol{w}}_{MLE} - \tilde{\boldsymbol{w}})|^2$$

Then, there exist a sequence of $J \times J$ matrices D_{T_0} with bounded eigenvalues such that for any $\delta > 0$, uniformly in $\boldsymbol{\alpha} \in \mathcal{S}_J(1)$,

$$\sup_{||w-\tilde{w}||\leq\delta(J/T_0)^{1/2}}\left|\boldsymbol{\alpha}'\left(\sum_t \mathbb{E}((y_{1t} - \boldsymbol{y}_{Jt}'\boldsymbol{w}) - (y_{1t} - \boldsymbol{y}_{Jt}'\tilde{\boldsymbol{w}}))\boldsymbol{y}_{Jt}\right) - T_0\boldsymbol{\alpha}'D_{T_0}(\hat{\boldsymbol{w}}_{MLE} - \tilde{\boldsymbol{w}})\right| = o((T_0J)^{1/2})$$

BvM II

Proof idea:

$$\|\Phi_{MLE} - Q\|_{TV} \leq \sqrt{\frac{1}{2} D_{KL}(\Phi_{MLE}||Q)}.$$

Lemma (KL Convergence (Barron 1986))

Let $\Phi_{J,T}$ be the MLE estimator distribution and $Q_{T,J}$ be the smooth, bounded Bayes posterior predictive distribution for fixed J and T_0. Suppose that as $J, T \to \infty$,

1. $\Phi_{J,T} \to P^*$,
2. $Q_{T,J} \to Q^*$,
3. *Q^* and P^* have the same mean and have bounded fourth moments.*

Then, it follows that

$$D_{KL}(\Phi_{J,T}||Q_{T,J}) = D_{KL}(\Phi^*||Q^*) + O(1/(TJ)).$$

BvM III

Then, we just need to compare the variances.

Lemma (Gaussian KL)

Suppose that Q and P are normal random variables with equal means and $k \times k$ covariance matrices Σ_Q and Σ_P. Then,

$$D_{KL}(P||Q) = \frac{1}{2}\left(\log \frac{|\Sigma_P|}{|\Sigma_Q|} - k + tr(\Sigma_Q^{-1}\Sigma_P)\right).$$