Bayesian and Frequentist Inference for Synthetic Controls

Ignacio Martinez   Jaume Vives-i-Bastida

Google   MIT

February 14, 2023

[Preliminary Draft]

Abstract

The synthetic control method has become a widely popular tool to estimate causal effects with observational data. Despite this, inference for synthetic control methods remains challenging. Often, inferential results rely on linear factor model data generating processes. In this paper, we characterize the conditions on the factor model primitives (the factor loadings) for which the statistical risk minimizers are *synthetic controls* (in the simplex). Then, we propose a Bayesian alternative to the synthetic control method that preserves the main features of the standard method and provides a new way of doing valid inference. We explore a Bernstein-von Mises style result to link our Bayesian inference to the frequentist inference. For linear factor model frameworks we show that a maximum likelihood estimator (MLE) of the synthetic control weights can consistently estimate the predictive function of the potential outcomes for the treated unit and that our Bayes estimator is asymptotically close to the MLE in the total variation sense. Through simulations, we show that there is convergence between the Bayes and frequentist approach even in sparse settings. Finally, we apply the method to re-visit the study of the economic costs of the German re-unification and the Catalan secession movement. The Bayesian synthetic control method is available in the bsynth R-package.

1. Introduction

Synthetic control methods (Abadie and Gardeazabal (2003), Abadie et al. (2010)) are often used to estimate treatment effects of aggregate policy interventions. In fact, the method

has been described as "arguably the most important innovation in the policy evaluation literature in the last 15 years" (Athey et al. (2021)). Despite this, statistical inference for such methods is often non-trivial. Suppose that we observe data for a unit that is affected by an intervention of interest, as well as data on a donor pool of untreated units. Synthetic controls match the treated unit to the weighted average of the donor units that most closely resembles the characteristics of the treated unit before the intervention. Once a suitable synthetic control is selected, differences in outcomes between the treated unit and the synthetic control are taken as estimates of the treatment effect on the treated. Given that in this setting we only observe one treated unit, developing asymptotic guarantees for the synthetic control estimator is challenging.

Inference for synthetic controls has been approached from various angles. First, a body of literature has focused on permutation based inference relative to a benchmark assignment process, first proposed in Abadie et al. (2010) and more recently in Firpo and Possebom (2018) and Abadie and L'Hour (2021) among others. Second, large sample properties of synthetic control estimators have been studied in different settings. Starting with Abadie et al. (2010) that show that in linear factor models the synthetic control is biased if perfect pre-treatment fit can not be achieved. Asymptotic properties in linear factor models have also been derived, most notably in Ferman (2021) under the assumption that the treated unit factor loading is in the convex hull of the donor units factor loadings as the number of pre-treatment periods and donor units grows. Other research has focused on projection theory results for average treatment effects estimated by synthetic controls (Li (2020)) and panel data approaches that relax the simplex restriction (Hsiao et al. (2012)). Third, conformal inference procedures have also been proposed. For example, in Chernozhukov et al. (2021) under the assumption that the synthetic control estimator is as good as a true synthetic control that recovers the treated unit. Finally, Bayesian inference methods have also been proposed. For instance, Pang et al. (2022) and Pinkney (2021) for linear factor model structures, Bayesian regression methods in Kim et al. (2020), Bayesian structural time series in Brodersen et al. (2015) and Scott and Varian (2014), or empirical Bayes approaches in Amjad et al. (2018).

A common thread in the literature is that either the synthetic control restriction that the weights are in the simplex is dropped, or the assumption is made that there exist a true synthetic control that is able to perfectly recover the treated unit. Motivated by this, in this paper we ask: *in linear factor models when is the minimizer of the statistical risk a synthetic control?* In a simple model, we derive conditions on the primitives of the factor structure

(the factor loadings) that make the target parameter be in the simplex. Then, given this target parameter, we focus on frequentist and Bayesian inferential procedures. We show under which conditions the MLE can uniformly estimate the predictive part of the factor structure for the treated unit in post-treatment periods (without the error shock) and provide conditions for a Gaussian asymptotic approximation. Then, we propose a Bayesian synthetic control that preserves the main features of the standard model and derive a Bernstein-von Mises (BvM) style result to link the frequentist and Bayesian inference.

The BvM result may be of particular interest to applied researchers as it provides a new way to perform frequentist inference using synthetic controls that is computationally appealing. While the conditions for the result require a large number of pre-treatment periods relative to the number of donor units and the existence of a true synthetic control that distributes weight among many donor units, we show through simulations that BvM convergence can be achieved even in sparse settings.

This paper contributes to the synthetic control literature in two ways. First, it provides a new characterization result for synthetic controls under linear factor models. Given that these models are often used to motivate synthetic controls, we see our result as relevant to applied researchers; we provide conditions under which factor models are well suited to study synthetic controls. Our main results reinforce the rule of thumb in the literature that without good pre-treatment fit synthetic control estimates may be misleading (Abadie (2021), Abadie and Vives-i Bastida (2022)). However, we also expand the scope of the class of linear factor models that motivate synthetic controls. We show that the condition in Ferman (2021) that the treated unit factor loadings fall in the convex full of the donor unit factor loadings, while sufficient, might not be necessary; other factor structures may also motivate the use of synthetic controls.

Second, this paper contributes to the literature by providing a new inferential procedure. While other Bayesian synthetic control estimators have been proposed, implementations often drop the simplex assumption. We propose and justify theoretically a Bayesian synthetic control that preserves the simplex assumption. This feature is important for interpretability and to limit extrapolation. Importantly, it provides an easy way of evaluating if synthetic controls should be used to estimate a causal effect: by checking whether the synthetic treated unit can replicate the pre-treatment outcome of interest, the researcher is able to evaluate whether the synthetic control estimator is likely be biased, and whether their Bayesian model may be miss-specified. We implement our proposed Bayesian synthetic control in the *bsynth* R-package and provide additional features to help researchers understand the

posterior treatement effect distribution and implicit weight estimates.

Finally, we apply the Bayesian procedure to a classic synthetic control application studied by Abadie et al. (2015), the German re-unification, and a more recent intervention; the attempted unilateral declaration of independence (UDI) of Catalonia in 2017. The Bayesian synthetic control yields a similar estimate to the frequentist one for the effect of the German re-unification on the GDP of West Germany. We find that the German re-unification lead to a 7% decrease of the GDP per capita. The effect of the UDI on Catalan GDP is smaller, but nevertheless, we find that the UDI lead to a 1% decrease of GDP. We posit, as Esteller-Moré and Rizzo (2022) note, that this decrease is due to capital reallocation caused by to the increased political instability.

The paper proceeds as follows. Section 2 describes the standard frequentist synthetic control and conditions under which the statistical risk minimizer is a synthetic control and can be estimated consistently by MLE. Section 3 describes the Bayesian synthetic control, the Bayesian inference procedure, presents the Bernstein-von Mises result and the connection with frequentist inference through simulations. Section 4 describes the *bsynth* R-package. Finally, Section 5 discusses the empirical application to the German re-unification and the Catalan secession movement.

2. The Frequentist Synthetic Control

2.1. Standard Synthetic Control for a single unit

Consider a setting in which we observe $J + 1$ aggregate units for T periods. The outcome of interest is denoted by Y_{it} and only unit 1 is exposed to the intervention during periods $T_0 + 1, \ldots, T$. We are interested in estimating the treatment effect $\tau_{1t} = Y_{1t}^I - Y_{1t}^N$ for $t > T_0$, where Y_{1t}^I and Y_{1t}^N denote the outcomes under the intervention and in absence of the intervention respectively. Since we do not observe Y_{1t}^N for $t > T_0$ we estimate τ_{1t} by building a counterfactual $\hat{Y}_{1t}^N$ of the treated unit's outcome in absence of the intervention.

As in the standard synthetic control our counterfactual outcome will be given by a weighted average of the donor units' outcomes, that is $\hat{Y}_{1t}^N = \sum_{j=2}^{J+1} w_j Y_{jt}$ for a set of weights $\mathbf{w} = (w_2, \ldots, w_{J+1})'$. To choose the weight vector $\mathbf{w}$ we use observed characteristics of the units and pre-intervention measures of the outcome of interest. Formally, we let the $K \times 1$ design matrix for the treated unit be $\mathbf{X}_1 = (Z_1, \bar{Y}_1^{\mathbf{K}_1}, \ldots, \bar{Y}_1^{\mathbf{K}_M})'$, where $\{\bar{Y}_1^{\mathbf{K}_i}\}_1^M$ represent M linear combination of the outcome of interest for the pre-intervention period. Similarly, for the donor units, $\mathbf{X}_0$ is a $K \times J$ matrix constructed such that its jth column is given

by $(Z_j, \bar{Y}_j^{\mathbf{K}_1}, \ldots, \bar{Y}_j^{\mathbf{K}_M})'$. We call the K rows of the design matrices $\mathbf{X}_0$ and $\mathbf{X}_1$ the *predictors* of the outcome of interest. This can include, for example, lags of the outcome variable and important context dependent characteristics of the aggregate units averaged over the pre-treatment period.

Abadie et al. (2010) propose estimating the $\boldsymbol{w}$ by solving the following program:

$$\min_{\boldsymbol{w} \in \Delta^J} \|\mathbf{X}_1 - \mathbf{X}_0 \mathbf{w}\|_V = \left(\sum_{h=1}^{k} v_h (X_{h1} - W_2 X_{h2} - \cdots - W_{J+1} X_{hJ+1})^2 \right)^{1/2},$$

where Δ^J denotes the $J-$dimensional simplex and the researcher can choose the predictor weighting matrix $V = \mathrm{diag}(v_1, \ldots, v_k)$ using his domain knowledge or using a data-driven procedure to optimize pre-treatment fit.

Given our synthetic control $\hat{\mathbf{w}}$ we can estimate our treatment effect on the treated for $t > T_0$ by:

$$\hat{\tau}_{1t} = Y_{1t} - \sum_{j=2}^{J+1} \hat{w}_j Y_{jt}.$$

In the following section we motivate synthetic controls when the potential outcomes are given by linear factor models. We derive conditions under which the target parameter will be in the simplex and show how a MLE can estimate the treated factor structure are J and T_0 grow.

2.2. Linear factor model

In this section we explore the connection between the Bayesian synthetic control method and the frequentist synthetic control. We start by carefully examining identification and inference of the predictive part of the treated outcome in a simple factor model. Following Ferman (2021), Ferman and Pinto (2021) and Hsiao et al. (2012) we consider the following linear factor model for the potential outcomes:

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

The observed data y_{it} is given by

$$y_{it} = d_{it}Y_{it}(1) + (1 - d_{it})Y_{it}(0),$$

and only the first unit is treated, so $d_{it} = 1$ for $i = 1$ and $t > T_0$ and $d_{it} = 0$ otherwise. For ease of exposition, we make the following preliminary 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) $\epsilon_{it} \sim_{i.i.d} N(0, 1)$.

Under **A1-A2** it is shown in the appendix that the conditional distribution of Y_{1t} given a realization of $\boldsymbol{Y}_{Jt} = (Y_{2t}, \ldots, Y_{J+1t})$, which we denote by the lowercase $\boldsymbol{y}_{Jt}$, is

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

where

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

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

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

Hence, conditional on the realization of the outcomes for the donor units, the distribution of the treated unit depends only on the weights $w_j(\boldsymbol{\lambda}, \sigma)$. We denote the $J \times 1$ vector of such weights by $\tilde{\boldsymbol{w}}$. While the conditions **A1-A2** seem restrictive, the main results in the paper can be extended to include settings with multiple factor loadings and a non-trivial time series component.

2.3. Identification and characterization of synthetic controls

In the following proposition we show that $\tilde{\boldsymbol{w}}$ is a minimizer of the statistical risk for the square loss amongst predictors that are linear combinations of the donor units.

Theorem 1 (Linear Predictors) *Let $\boldsymbol{Y}_1(0)$ denote the $T_0 \times 1$ vector of outcomes for the treated unit and $\boldsymbol{y}_J$ the $T_0 \times J$ matrix of outcome realizations of the donor units for time periods $1, \ldots, T_0$. Under assumptions **A1-A2** it follows that*

$$\tilde{\boldsymbol{w}} \in \arg\min_{\boldsymbol{w}} \frac{1}{T_0} \mathbb{E}\left[(\boldsymbol{Y}_1(0) - \boldsymbol{y}'_J \boldsymbol{w})' V (\boldsymbol{Y}_1(0) - \boldsymbol{y}'_J \boldsymbol{w})\right],$$

for any positive semi-definite matrix V.

Theorem 1 can be generalized to cases in which the time series component of the common factors F_t matters by adding assumptions on the asymptotic behavior of the common factors and the error term. In these cases however, the statement will require the size of the pre-treatment set to go to infinity. While $\tilde{\boldsymbol{w}}$ is a minimizer of the statistical risk amongst linear predictors, it is not as clear whether it can recover the predictive part of treated outcome in future periods: $\lambda_1 F_t$ for $t > T_0$. The following proposition fleshes out the conditions under which $\boldsymbol{y}'_{JT_0+1} \tilde{\boldsymbol{w}}$ converges to $\lambda_1 F_{T_0+1}$.

Theorem 2 (Predictor convergence) *Given **A1-A2** the following hold:*

1. *There exist no values of λ_j that allow $\boldsymbol{y}'_{JT_0+1} \tilde{\boldsymbol{w}} \overset{m.s.}{\to} \lambda_1 F_{T_0+1}$ as $J \to \infty$.*

2. *If $\frac{1}{\|\boldsymbol{\lambda}_J\|_2^2} \sum_j |\lambda_j| \to 0$ as $J \to \infty$, then $\boldsymbol{y}'_{JT_0+1} \tilde{\boldsymbol{w}} \overset{p}{\to} \lambda_1 F_{T_0+1}$.*

Theorem 2 may look surprising as it seems to provide a negative result, but it is in line with the synthetic controls literature. Statement (1) speaks to the fact that identification and asymptotic results for synthetic control estimators often require conditioning on perfect (or good) pre-treatment fit (Abadie et al. 2010). Statement (2) provides conditions for convergence in probability. The condition implies that as $J \to \infty$, $\|\boldsymbol{\lambda}_J\|_2^2 \to \infty$ which implies the condition in Ferman 2021 that $\|\boldsymbol{w}_J\|_2^2 \to 0$ given the analytic form of $\tilde{\boldsymbol{w}}_J$. Furthermore, as $\|\boldsymbol{\lambda}_J\|_2^2 \to \infty$ for $J \to \infty$ we also recover the treated unit factor loading:

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

7

The proof technique may play a role in the conditions in Theorem 2. For example, a different proof technique may yield a different sufficient condition. It remains to show which is the weakest necessary and sufficient condition for convergence in probability. Overall, however, a necessary condition for convergence in probability is $\|\boldsymbol{\lambda}_J\|_2^2 \to \infty$. Intuitively, unless we can distribute the error terms over all units as the donor pool grows we are not able to get consistent inference. This is intuition is similar to the requirements in Ferman 2021 and can be thought as justifying the Ferman results in our conditional normal setting.

Next, we consider the question of whether $\tilde{\boldsymbol{w}}$ is a *synthetic control*. By this we mean whether the sum to one and non-negative constraints can be justified under our factor model. In general, if λ_1 is fixed and does not depend on other factor loadings then the result will be negative. However, if we allow λ_1 to depend on the $\boldsymbol{\lambda}_J$ then the consistency conditions and synthetic control constraints can be reconciled.

Theorem 3 (Synthetic Control Characterization) *For fixed J under* **A1-A2**, *$\tilde{\boldsymbol{w}} \in \Delta^J$ 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, the following statements follow

1. *For a fixed λ_1, a sufficient condition for the existence of sequences $\{\lambda_j\}$ such that (1) and (2) hold is that $\lambda_1^2 \geq \frac{4}{J\sigma^2}$.*

2. *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.*

3. *Suppose condition (1) holds, let $\lambda_1 = h(\boldsymbol{\lambda}_J)$ for a component-wise weakly increasing odd function $h : \mathbb{R}^J \to \mathbb{R}$, then if as $J \to \infty$, $\|\boldsymbol{\lambda}_J\|_2^2 \to \infty$, a sufficient condition for $\tilde{\boldsymbol{w}} \in \Delta^J$ is $|h(\boldsymbol{\lambda}_J)| \frac{\|\boldsymbol{\lambda}_J\|_1}{\|\boldsymbol{\lambda}_J\|_2^2} \to 1$.*

4. *For any $\boldsymbol{\lambda}_J$, if $\frac{\|\boldsymbol{\lambda}_J\|_1}{\|\boldsymbol{\lambda}_J\|_2^2} \in \Delta(\boldsymbol{\lambda}_J)$ then any function h such that $h(\boldsymbol{\lambda}_J) = \boldsymbol{\lambda}_J' \boldsymbol{w}$ for $\boldsymbol{w} \in \Delta_J$ satisfies the condition in (3).*

Theorem 3 clarifies the conditions on the factor loadings under which the target parameter $\tilde{\boldsymbol{w}}$ is a synthetic control. The main result is that a sufficient condition for $\tilde{\boldsymbol{w}}$ to be a synthetic control is

$$|h(\boldsymbol{\lambda}_J)| \frac{\|\boldsymbol{\lambda}_J\|_1}{\|\boldsymbol{\lambda}_J\|_2^2} \to 1,$$

when $\lambda_1 = h(\boldsymbol{\lambda}_J)$. Together with our sufficient conditions from Theorem 2, this implies that λ_1 has to grow with the factor loadings of the donor units. In particular, this rules out the possibility that λ_1 is a fixed constant. This however does not imply that synthetic controls are not possible, in fact, this condition is more general than the one required by Ferman 2021. When $\frac{\|\boldsymbol{\lambda}_J\|_1}{\|\boldsymbol{\lambda}_J\|_2^2} \in \Delta(\boldsymbol{\lambda}_J)$, it follows that there exists a $\boldsymbol{w}^*$ such that $\lambda_1 = \boldsymbol{\lambda}_J' \boldsymbol{w}^*$ that satisfies the condition and, therefore, implies that as $J \to \infty$, $\tilde{\boldsymbol{w}} \in \Delta^J$. Hence, in settings in which the treated unit factor loading is in the convex hull of the donor units factor loadings the target parameter *is* a synthetic control.

2.4. Inference

In this section we consider how to estimate $\tilde{\boldsymbol{w}}_J$ using a data set of pre-treatment outcomes $\{y_{1t}(0), \boldsymbol{y}_{Jt}(0)\}_{t=1}^{T_0}$. Given that we are interested in comparing frequentist and Bayesian procedures we focus on the maximum likelihood estimator. We do not directly observe the factor loadings, but we can estimate the $\tilde{\boldsymbol{w}}$ weights by maximizing the following *pseudo* log-likelihood for parameter $\boldsymbol{\theta} = (\boldsymbol{w}, \Sigma)$:

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

We derive our theoretical results for the MLE in two parts. First, in Theorem 4 we show that for fixed J, the size of the donor pool, the MLE can recover the predictive part of the treated unit factor model and has the standard Gaussian approximation as $T_0 \to \infty$. Recall, however, that our characterization and identification results in the previous section required $J \to \infty$. Therefore, we also derive conditions under which as J and T_0 go to ∞ the MLE uniformly converges to the predictive part of the treated unit factor model. The second set of results Theorem 5 and Corollary 5.1 extend results in the semi-parametric estimation literature to the synthetic control framework.

Theorem 4 (MLE for fixed 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**:*

1. *$\hat{\boldsymbol{w}}_{MLE} \xrightarrow{p} \tilde{\boldsymbol{w}}$ as $T_0 \to \infty$ for fixed J.*

2. *$\sqrt{T_0}(\hat{\boldsymbol{w}}_{MLE} - \tilde{\boldsymbol{w}}) \xrightarrow{a} N(0, V_{T_0})$ as $T_0 \to \infty$ for fixed J, for $V_{T_0} = \frac{1}{T_0}\mathbb{E}[(\nabla_{\boldsymbol{w}} l_{T_0}(\tilde{\boldsymbol{\theta}})\nabla_{\boldsymbol{w}} l_{T_0}(\tilde{\boldsymbol{\theta}})')^{-1}]$, where $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{w}}, \tilde{\Sigma})$.*

For fixed J, the MLE consistency and asymptotic normality result is straightforward. Theorem 4 shows that the target parameter $\tilde{w}$ can be consistently estimated by the MLE. However, as discussed, our identification result requires that the donor pool grows with the sample size, $J \to \infty$. This means that the parameter space is growing with the sample size T_0. In order to account for this, Theorem 5 provides conditions for uniform consistency and asymptotic normality to the target parameter as $J, T_0 \to \infty$.

Theorem 5 (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. *$\frac{1}{T_0} \sum_t \boldsymbol{y}_{Jt} \boldsymbol{y}'_{Jt} = D_{T_0}$ where $0 < \lim \inf_{T_0} \sigma_{min}(D_{T_0}) \leq \lim \sup_{T_0} \sigma_{max}(D_{T_0}) < \infty$,*

2. *$max_{t \leq T_0} \|\boldsymbol{y}_{Jt}\|_2^2 = O_p(J)$,*

3. *$\sup_{\boldsymbol{\beta}, \boldsymbol{\gamma} \in \mathcal{S}_J(1)} \sum_t |\boldsymbol{y}'_{Jt} \boldsymbol{\beta}|^2 |\boldsymbol{y}'_{Jt} \boldsymbol{\gamma}|^2 = O_p(T_0)$.*

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

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

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

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

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

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

Theorem 5 shows that L^2 norm convergence and uniform Gaussian approximation is possible when T_0 grows at rate faster J. In particular, we require that T_0 grows faster than J for the consistency result and that T_0 grows faster than J^2 for the asymptotic normality result. While these rates might be impractical in settings with small T_0, they speak to the discussion in Abadie et al. 2010 and Abadie and Vives-i-Bastida 2021 that large donor pools may increase the bias in synthetic control estimators. Intuitively, larger donor pools imply more parameters to estimate which might increase the finite sample bias of the estimator. Next, we apply this result to our specific setting when we use the data $\boldsymbol{y}_{JT_0+1}$ to predict the treated unit outcome in absence of the intervention.

Corollary 5.1 *Under the conditions of Theorem 5, 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

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

Corollary 5.1 provides conditions for valid frequentist inference to the predictive part of the treated unit factor model as $T_0, J \to \infty$. Similar semi-parametric results have also been derived in Ferman (2021). Our results are different in that they apply to a wider class of models under our characterization conditions and provide explicit rate conditions for J and T_0. Indeed, all conditions are imposed directly on the factor structure. It is important to note that the rates could potentially be improved to be $o(T_0) = J$, as we will show in Section 3.4 in simulations in well behaved cases.

3. The Bayesian Synthetic Control

We propose the Bayesian equivalent of the program described in section 2 to generate synthetic controls. In particular, our Bayesian formulation includes two key aspects of synthetic controls: (1) the synthetic treated unit is a convex combination of the donor units, that is, we do not want to extrapolate outside the convex hull of the donor units and (2) we construct the synthetic control by matching the predictors of treated unit and donor units for the outcome of interest. An advantage of the Bayesian approach is that we can directly quantify the uncertainty in our estimates. In this section, first we discuss a valid Bayesian inference procedure and then the Bayesian synthetic control model.

3.1. Bayesian Inference

We derive conditions for valid Bayesian inference in a general setting that includes our set up. Following Imbens and Rubin (1997) and Pang et al. (2022) we consider the Bayesian inference problem as a missing data problem. Let the adoption process of the intervention for one treated unit be encoded by a treatment random variable $\mathbf{d}_1 = (d_{11}, \dots, d_{1T})$ that is fully determined by an adoption time random variable a_1. Similarly, we can define these variables for the donor pool as $\mathbf{D}$ and $\mathbf{a}$.

The potential outcome function of our outcome of interest Y_{it} under SUTVA is given by

$Y_{it}(\mathbf{d}_i(a_i)) = Y_{it}(a_i)$, where no anticipation implies that $Y_{it}(a_i) = Y_{it}(c)$ for $t < a_i$ and c denoting the counterfactual state in which the unit is never treated.

Assumptions:

1. **(B1) - (B2)**: SUTVA and no anticipation.

2. **(B3) Latent Ignorability**: there exist latent variables $\mathbf{U}_i = (u_{i1}, \ldots, u_{iT})$ such that:

$$\mathbf{y}_i(0) \perp \mathbf{D}_i | \mathbf{U}_i.$$

3. **(B4) Exchangeability**: given $\mathbf{U}$, permutations of the indices it of the sequence $\{(Y_{it}(0)\}_{i\in[N],t\in[T}$ do not alter the joint distribution.

The above assumptions allow us to get an expression for the posterior on the "missing" data, the outcome counterfactual for treated units after treatment assignment if they had not been treated. The predictive inference consists in sampling from the posterior distribution of the missing data conditional on the observed data and a model indexed by parameter vector $\boldsymbol{\theta}$ with prior distribution $\pi(\theta)$. Let the missing and observed data be denoted by $\mathbf{Y}^{mis} = \{Y_{1t}\}_{t>T_0}$ and $\mathbf{Y}^{obs} = (\{Y_{it}\}_{i\in[J+1],t\leq T_0}, \{Y_{jt}\}_{j\in[J],t>T_0}, \mathbf{D})$. Then, under **B1-B3** we can factor out the assignment:

$$
\begin{aligned}
P(\mathbf{Y}^{mis}|\mathbf{Y}(0)^{obs}, \mathbf{D}, \theta) &\propto P(\mathbf{Y}(0)^{mis}, \mathbf{Y}(0)^{obs}, \mathbf{D}, \theta) \\
&\propto P(\mathbf{Y}(0), \mathbf{U})P(\mathbf{D}|\mathbf{Y}(0), \theta) \\
&\propto P(\mathbf{Y}(0), \theta)P(\mathbf{D}|\theta) \\
&\propto P(\mathbf{Y}(0), \theta).
\end{aligned}
$$

Then under **B4** by de Finetti's Theorem we can separate out the posterior predictive distribution and the likelihood

$$
\begin{aligned}
P(\mathbf{Y}(0)^{mis}|\mathbf{Y}(0)^{obs}, \mathbf{D}, \theta) &\propto P(\mathbf{Y}(0), \theta) \\
&\propto \int \Pi_{it} f(Y_{it}(0)|\theta)\pi(\theta)d\theta \\
&\propto \int \underbrace{\left(\Pi_{it\in \text{ mis}} f(Y_{it}(0)^{mis}|\theta)\right)}_{\text{Posterior Predictive Distribution}} \underbrace{\left(\Pi_{it\in \text{ obs}} f(Y_{it}(0)^{obs}|\theta)\right)}_{Likelihood} \pi(\theta)d\theta
\end{aligned}
$$

where f is the marginal density of the potentially observable data.

The assumptions imposed by our sampling model **A1**-**A2** in which the data is i.i.d and the potential outcomes are given by a latent factor model implicitly satisfy **B1**-**B4**. Therefore, under our sampling we will be able to sample from the Bayes posterior distribution to recover the distribution of the missing data and, therefore, to recover the distribution of the target treatment effect. In order to estimate the model, we use a HMC sampler (NUTS) to get M draws $\{\theta^m\}$ and then our predictive posterior is given by the realizations of

$$\tau_{1t}^m = Y_{1t}(1) - y_{1t}^m(0),$$

In Section 4 we describe the Bayesian estimation procedure we implement in the *bsynth* package in more detail. In the next section, we consider under which conditions on the prior model can be use a Bayes estimator to approximate the frequentist inference.

3.2. Bayesian Model

A Bayesian model imposes a functional form for the prior distribution and data density f, by modelling explicitly the potential outcome in absence of treatment. In the previous section we showed under which conditions can we perform valid Bayesian inference. In this section, we impose restrictions on the prior structure. We start by considering a Gaussian prior model

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

Given that the Gaussian conjugate prior is Gaussian, it can be shown that the **Bayes estimator** for the implicit weights is given by

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

Furthermore, in this case the predictive posterior distribution is normally distributed and

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

Motivated by the characterization of synthetic controls in Theorem 3 we consider Bayesian models in which the parameters μ_j are in the simplex. This suggests adding the following restriction:

$$\mu_j \sim Dir(1),$$

where $Dir(1)$ denotes the Dirichlet distribution with scale one. Intuitively, this restriction forces the means of the target weights to be in the simplex.

In Section 4 we describe alternative Bayesian models implemented in the *bsynth* package that allow for additional covariates and a Gaussian process term. The Bayesian synthetic control with additional covariates, which has not been implemented in the literature yet, and the use of Gaussian processes, also suggested by Arbour et al. (2021). Our theoretical results could potentially be extended to apply to both cases with additional regularity conditions. We leave this for a future iteration of this paper. In the case of Gaussian process note that in a general setting a BvM result for Gaussian processes was derived in Ray and van der Vaart (2020). In the next section, we link the Bayesian and frequentist inference for the baseline Bayesian model.

3.3. Bernstein-von Mises Result

In this section we consider how the Bayesian inference can be used to approximate the frequentist inference. We derive a Bernstein-von Mises style result in which under the correct prior specification the Bayesian posterior predictive distribution converges in the total variation sense to the MLE sample distribution as $T_0, J \to \infty$. Intuitively, our result states that if we assume that the factor loading of the treated unit can be recovered by a convex combination of the treated units then under the same assumptions that yield a valid MLE estimator, the Bayes estimator is able to consistently estimate the predictive term $\lambda_1 F_{T_0+1}$. If additionally, the uncertainty in our predictions converges to the frequentist sampling variance, then the two estimators distributions are close in the total variation sense.

Theorem 6 (BvM) *Under **A1-A2**, the assumptions of Corollary 5.1 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

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

14

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

Theorem 6 imposes strong conditions on the class of priors necessary to approximate the MLE distribution. In particular, it is key that we choose the μ_j in the simplex and in a way that recovers the treated unit factor loading. The requirement that such a sequence of priors exists is motivated by our characterization conditions in Theorem 3. In the same spirit as in the frequentist synthetic control, if the Bayesian synthetic control is unable to find implicit weights and a prior that replicate the treated unit outcomes in the pre-treatment period, then it is likely the Bayesian synthetic control will be biased in the same as the frequentist synthetic control.

The other requirement in the proof of Theorem 6 is that the Bayesian posterior is Gaussian. While this requirement is important for the proof method, which relies on the analytical form of the KL divergence between Gaussian distributions, it is not a necessary requirement. A more general result could be derived by imposing weaker restrictions on the functional form and the second moments of the posterior distribution.

The BvM result in Theorem 6 is important because it gives conditions under which researchers might be able to interpret their Bayesian credible intervals as valid confidence intervals. As $T_0, J \to \infty$ under the conditions of Theorem 6 the $1 - \alpha$ credible interval defined by the limit Bayesian posterior predictive distribution and the $1 - \alpha$ confidence interval defined by the MLE estimator will coincide. Hence, using Bayesian synthetic controls offers a new way of performing valid asymptotic inference for synthetic controls without the need of exact inference or permutation tests.

3.4. Simulation Evidence

In this section we compare the standard synthetic control and the Bayesian synthetic control for a grouped linear factor model data generating process. Similar data generating processes have been used to study the properties of synthetic controls estimators, for example in Firpo and Possebom (2018) or in Abadie and Vives-i Bastida (2022). In particular, we let the potential outcome in absence of intervention be given by

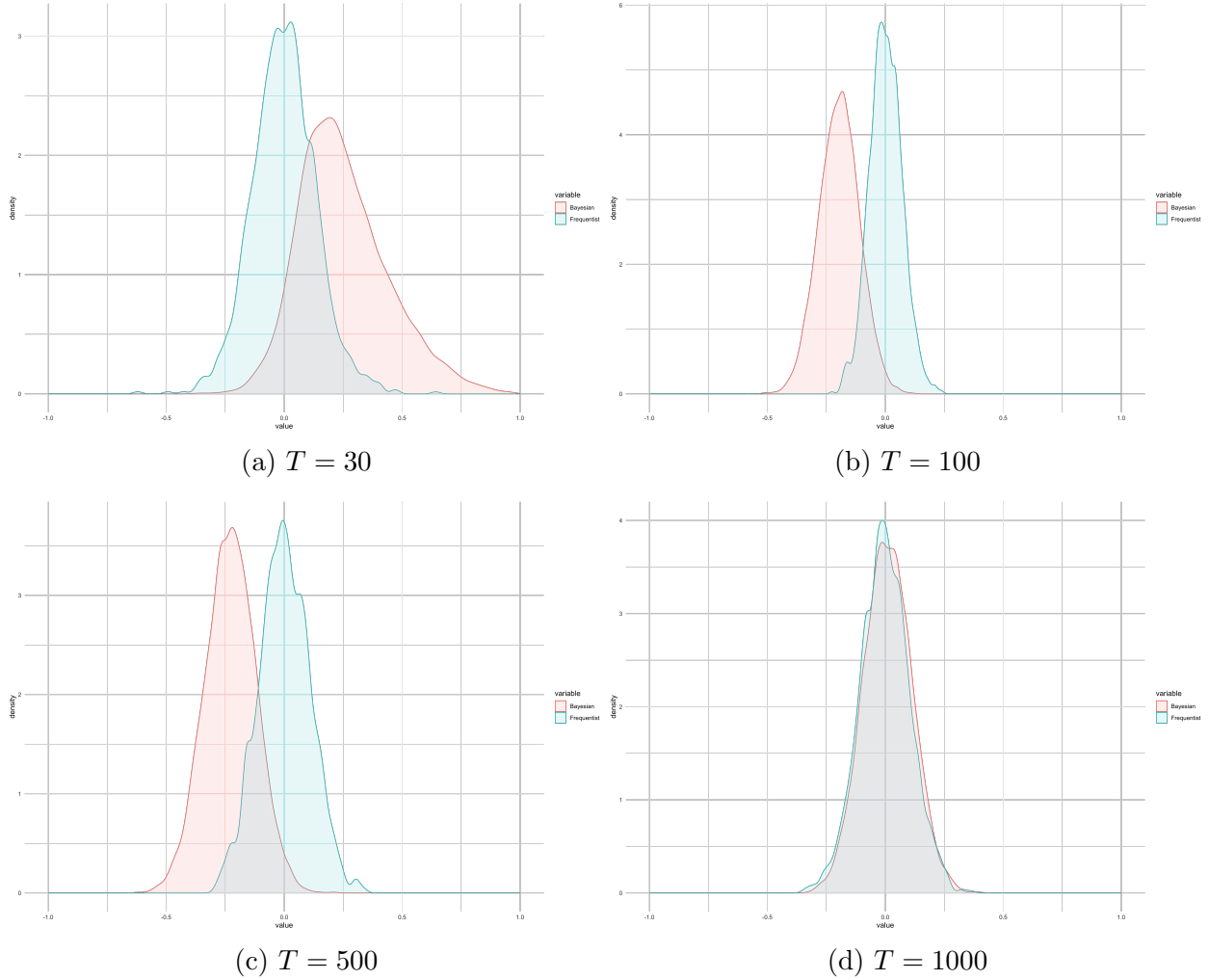$$Y_{it}(0) = \lambda_{f(i)t} + \epsilon_{it}. \tag{1}$$

where the λ_{ft} follow an $AR(1)$ with $\rho = 0.5$ and standard Gaussian innovations and $f(1) = f(2)$ so that a synthetic control that puts all the weight in unit 2 is unbiased. The noise is given by $\epsilon_{it} \sim N(0, \sigma^2)$ and we assume without loss of generality that only unit 1 is treated and the treatment effect is 0.

This setting satisfies the convex recovery assumptions of Theorem 3 and Theorem 6 as there exists a sequence of weights that perfectly recreates the factor loading of the treated unit. In this

case $w_2 = 1$ and $w_j = 0$ for all $j > 2$. However, it does not necessarily satisfy the *density* condition or the prior condition such that the target parameter $\|\tilde{\boldsymbol{w}}\| \to 0$ as $J \to \infty$. In fact, we consider a fixed $J = 20$ design. We will see however, that despite this, there will still be BvM convergence, indicating that our density assumption may not be a necessary condition.

For our simulation analysis, we let $T_0 \to \infty$ and for each T_0 we estimate the standard synthetic control and report our estimated treatment effect on the treated for 10000 draws. We compute the treatment effect over 10 post-treatment periods $\hat{\tau}_1 = \frac{1}{T-T_0} \sum_t \hat{\tau}_{1t}$, this allows us average over the additional noise terms ϵ_{it} for $t > T_0$. We compare this empirical distribution to the posterior predictive distribution of our Bayesian model for one draw. Figure 1 shows the two distributions for different values of T_0.

Figure 1: Convergence of frequentist and Bayesian coverage as $T \to \infty$.

(a) $T = 30$

(b) $T = 100$
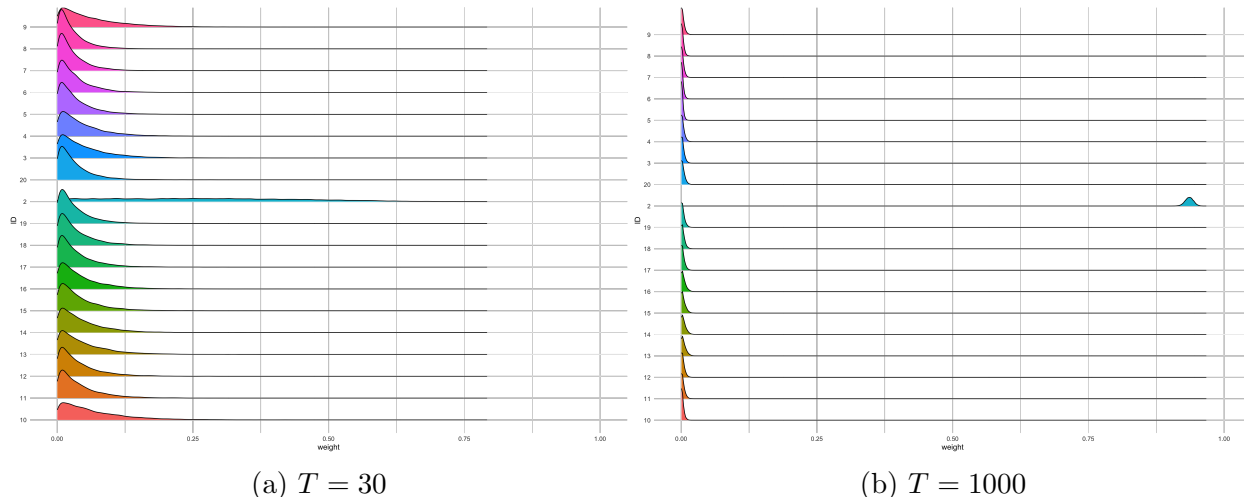
(c) $T = 500$

(d) $T = 1000$

Notes: Kernel densities of the frequentist empirical distribution of the estimated treatment effect over 10000 draws and the Bayesian posterior distribution for one draw for different values of T_0. The potential outcomes are generated by the grouped factor model (1) with $\sigma = 0.25$.

As can be seen in Figure 1 the convergence is slow, but as $T_0 \to \infty$ the Bayesian posterior and the frequentist empirical distribution coincide. As expected, when T_0 is small the Bayesian synthetic control is biased, but as T_0 increases and the prior becomes dominated, the bias disappears and both methods converge. Faster convergence is possible in less sparse settings in which the true w_j weights are more densely distributed. In the appendix we show a larger array of simulations for different designs. These findings suggest that the scope of our BvM result might be wider than anticipated by the theory. For applied researchers with medium size T_0 and J it may be possible to use Bayesian synthetic control methods and interpret their findings from the frequentist lens. An

17

important consequence of this is that researchers could potentially use their predictive posterior distribution to create valid confidence intervals for synthetic control estimators.

Figure 2: Implicit weights as $T \to \infty$.

(a) $T = 30$ (b) $T = 1000$

Notes: Kernel densities of the Bayesian implicit weights for different values of T_0. The potential outcomes are generated by the grouped factor model (1) with $\sigma = 0.25$.

Figure 2 shows that the implicit weights of the Bayesian estimator converge to the true factor loading representation. This convergence is stronger that the L^2 convergence that we show theoretically in Theorem 5. This adds to the previous discussion that we may be able to use the Bayesian synthetic control to study the frequentist properties of synthetic control estimators.

4. The *bsynth* package

We have implemented the Bayesian synthetic control model we propose in this paper in the publicly available *bsynth* R-package.[1] The *bsynth* R-package extends the Bayesian model we propose to include more complex models. In particular, it allows for additional features such as modelling the time series component with a Gaussian process and adding additional covariates. In particular, the *bsynth* package allows for Bayesian models with the following form:

$$X_1|w, \sigma \sim N(X_0 w + f_{1t}, \sigma^2 \text{diag}(\Gamma)^{-2}),$$
$$w \sim Dir(1),$$
$$f_{1t} \sim \mathcal{GP},$$
$$\sigma \sim N(0,1)^+,$$

---

[1]The package can be accessed at https://github.com/google/bsynth.

$$\Gamma \sim Dir(v_1, \ldots, v_K), \quad v_k \in \Delta^k,$$

where here X_1 and X_0 denote the design matrices for the treated and donor units which may include the outcome variables as well as additional predictors. The f_{1t} term is modelled through a Gaussian process and the weight of the predictors are modelled by Γ. To preserve the main features of synthetic controls both the w and the v weights are assumed to be in the simplex.

The *bsynth* package offers the possibility to compute different statistics of the posterior distribution. Of special interest is an upper bound on the frequentist bias given by the Bayesian model. This bound is motivated by the finite sample bound first developed in Abadie et al. (2010) and expanded in Vives-i Bastida (2022) to include additional predictors. Given the BvM style result from Theorem 6 the bound can be used to check the likelihood that the Bayesian synthetic control is badly biased due to model misspecification. Intuitively, if the Bayesian synthetic control can not replicate in the pre-treatement period the outcomes of the treated unit then our frequentist interpretation of the method will be biased in the same way standard synthetic controls are biased when perfect pre-treatment fit can not be achieved. In the following section we use the Bayesian synthetic control to study two political economy questions.
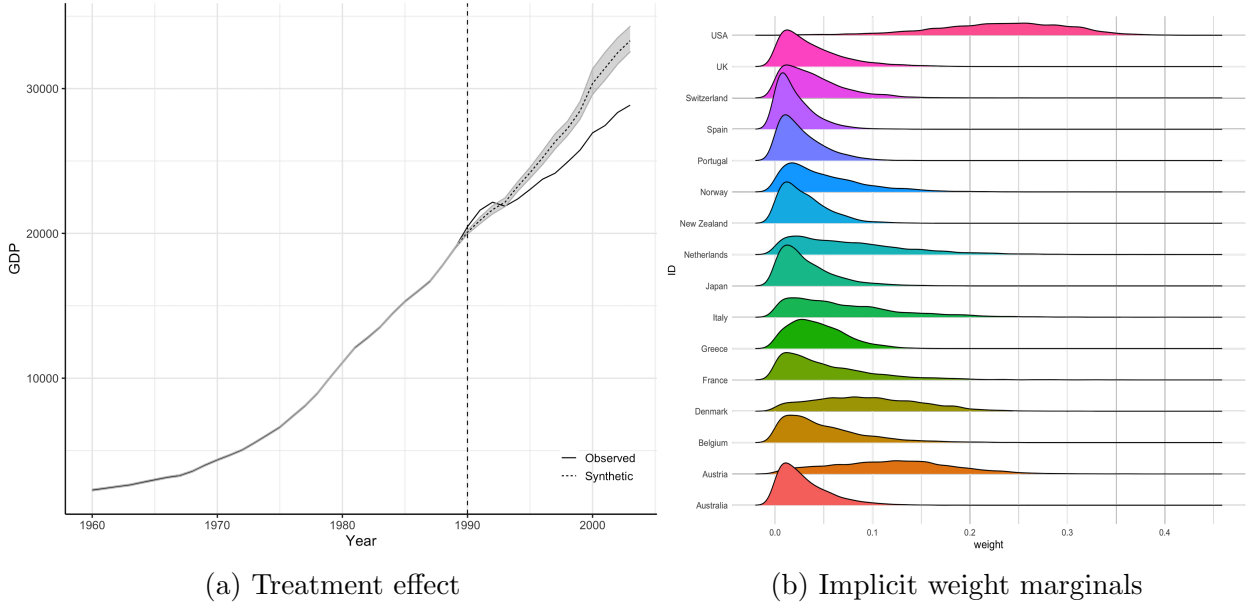
5. Empirical application

One of the most salient applications of synthetic controls is the study of the impact of the German re-unification in 1990 to the GDP of West Germany. In this paper, we replicate this finding using the Bayesian synthetic control and we highlight the usefulness of the Bayesian inference procedure.

5.1. Re-visiting the German re-unification

In 1989, after the fall of the Berlin wall, the process of re-unifying West Germany and East Germany started. Abadie et al. (2015) found that in absence of the re-unification, West Germany's GDP would have been 8% higher in 2003, 13 years later. Using the same data and specification as Abadie et al. (2015) in Figure 3 we display the Bayesian synthetic control for West Germany over and the marginal distribution of implicit weights of the donor units.

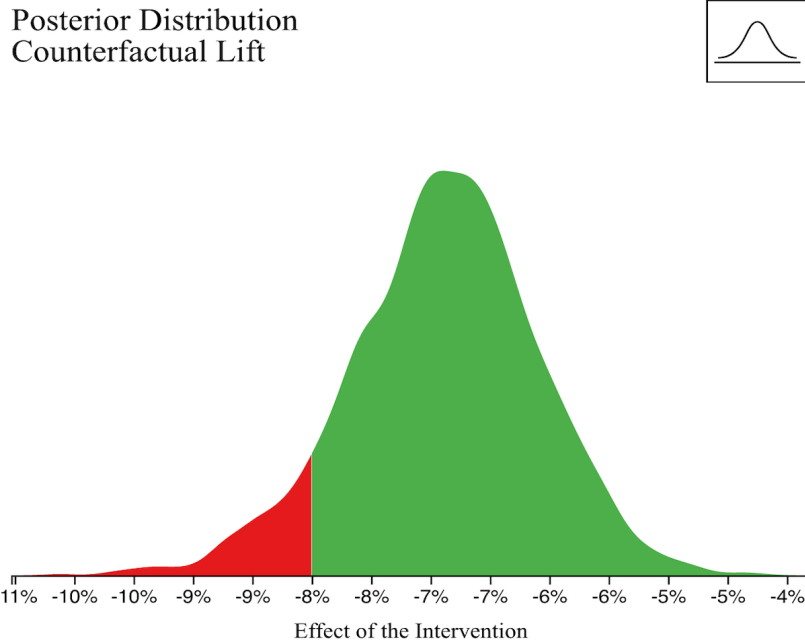Figure 3: Bayesian synthetic control for West Germany.

(a) Treatment effect

(b) Implicit weight marginals

Notes: Panel (a) shows the West Germany GDP per capita and Bayesian synthetic control estimates from 1960 to 2003 with credible intervals shaded in grey. Panel (b) shows the marginal distributions of the implicit weights of the Bayesian synthetic control.

The Bayesian synthetic control in Figure 3 is similar to the standard synthetic control in Abadie et al. (2015). It shows an overall increasing trend in absence of the intervention with a slight reversal in the first few years after the intervention. Similarly, in panel (b) the implicit weights of the Bayesian synthetic control also indicate that West Germany is best replicated by a combination the United States and Austria. In the appendix, we also report the correlations between the implicit weights, a statistic that could be of interest to applied researchers seeking to understand the robustness of their synthetic control estimate. The correlation matrix shows that the implicit weight distributions of the United Sates and Austria are positively correlated, as expected, but it also shows that an alternative synthetic control could have included Denmark, Spain, Portugal and Australia which are also correlated with each other.

An advantage of the Bayesian synthetic control is that we estimate a full posterior distribution. Figure 4 shows the treatment posterior distribution for the post-treatment period relative to the baseline year. The mean of the posterior distribution is around 7% which is slightly lower than the frequentist estimate of the ATET of 8%. In the appendix, we show the posterior distribution of the frequentist bias bound that depends on the MAD in the pre-treatment period. The bound indicates that the likelihood that the sign of the result could be overturned due to the bias induced by model misspecification (bad pre-treatment fit) is small.

Figure 4: Treatment effect posterior distribution.

Notes: Posterior distribution of the average treatment effect in the post-treatment period, the change of color indicates the frequentist estimated effect of 8%. This figure is generated directly from the output of the *bsynth* R-package.
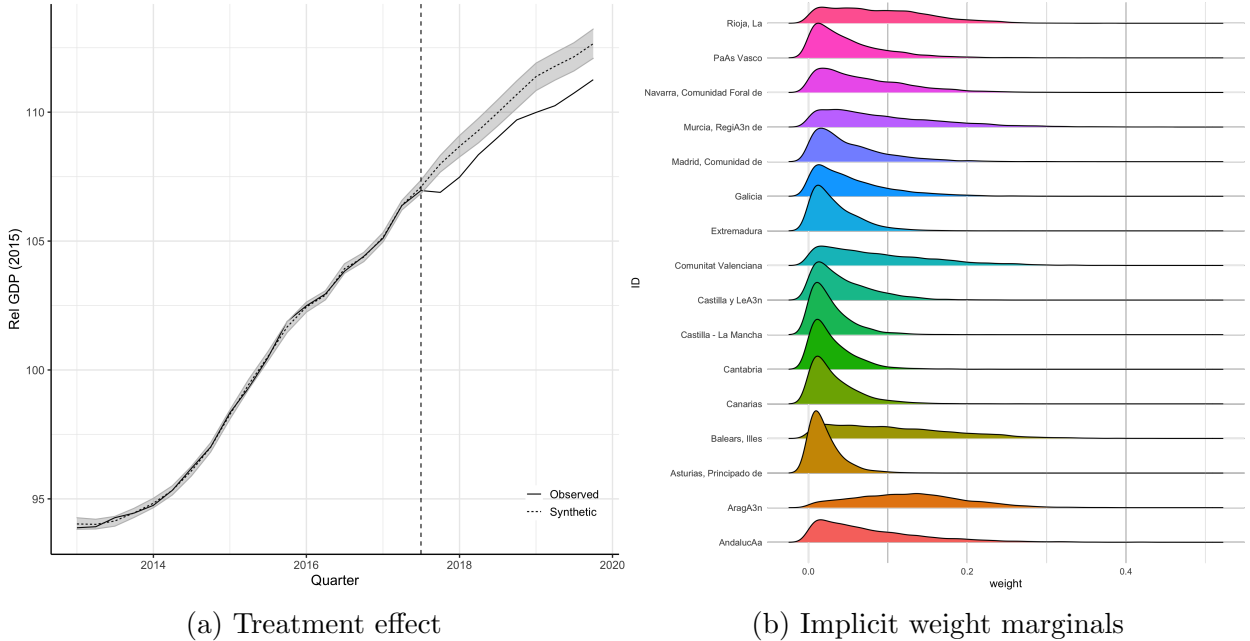
5.2. The impact of the Catalan secession movement

The pro-independence demands of the Catalan electorate, commonly referred as the Catalan secession process, have been the focus of Catalan politics, and Spanish politics, since 2012. While the movement started with a plea for higher fiscal autonomy (the ability of the Catalan government to collect and administer its own taxes), quickly it shifted towards independence through a series of plebiscitary elections that were won by pro-independence parties. This transition towards a pro-independence mandate culminated in 2017 with the celebration of a non-binding independence referendum on the 1st of October which received wide international media coverage. Following the referendum, on the 27th of October a unilateral declaration of independence (UDI) was voted and approved by the Catalan parliament. However, the UDI was deemed unconstitutional by the Spanish constitutional court and the Spanish government suspended the Catalan government (activating the *article 155*). More so, the members of the Catalan parliament and the government behind the UDI were arrested while several high ranking members of the government, including the president Carles Puigdemont, sought asylum in other countries.

Overall the UDI and the events that ensued after the Spanish government suspended the Catalan

government had a great impact on the Catalan economy. An extensive analysis of different outcomes affected by the UDI has been done by Esteller-Moré and Rizzo (2022), also using a synthetic control methodology. The main finding in Esteller-Moré and Rizzo (2022) is that the UDI lead to a significant drop in short term bank deposits (capital out-flow), accounting to up to 2.5bn euros. This result is in line with reported evidence in the media[2] that the UDI and political instability lead to many firms and individuals to reallocate outside of Catalonia. In this paper, we focus on quarterly GDP as our main outcome of interest. To estimate the Bayesian synthetic control we consider all quarters between 2012 and the UDI in 2017 as the pre-treatment period and all quarters from the UDI to 2020 as the post-treatment period. We choose this time frame to avoid the effects of the 2008 financial crisis and the 2010 European debt crisis. Given that Catalonia has a different economic structure than other regions in Spain the effect of the two crises was different and therefore matching outcomes in that time period is potentially misleading. We also restrict the post period to avoid the covid-19 shock in 2020.

Figure 5: Bayesian synthetic control for Catalonia.

(a) Treatment effect

(b) Implicit weight marginals

Notes: Panel (a) shows the Catalan quarterly GDP (relative to 2015) and Bayesian synthetic control estimates from 2013 to 2020 with 75% credible intervals shaded in grey. Panel (b) shows the marginal distributions of the implicit weights of the Bayesian synthetic control.

Figure 5 shows the Bayesian synthetic control and implicit weight marginal distributions for

---

[2]See for example the following Financial Times article: https://www.ft.com/content/18f0ca8c-607b-4633-a83f-e99f71a046e8.

Catalonia. The main finding is that the UDI lead to a quarterly GDP decrease of 1% over 2017-2020 relative to the third quarter of 2017. The Bayesian synthetic control allows us to quantify uncertainty with the 75% credible interval being $[-0.6\%, -1.3\%]$ and the 95% credible interval slightly larger at $[-0.3\%, -1.6\%]$. Overall this is indicative evidence that the capital flight described in the press and Esteller-Moré and Rizzo (2022) may have had a significant impact on the Catalan GDP. In panel (b) of Figure 5 we can see that the Bayesian synthetic Catalonia is mostly composed of neighboring regions: Aragon, the Balearic islands and Valencia. A potential concern is that the UDI also affected positively the regions that received the out-flowing capital, biasing the estimated treatment effect upwards. In particular, as detailed in media coverage, many Catalan firms migrated their fiscal head quarters to Madrid or the neighboring Valencia.[3] In light of this, in the appendix, we re-do the exercise removing Madrid and Valencia from the donor pool and find similar treatment effects.

6. Conclusion

This paper contributes to the synthetic control literature in two ways. First, we characterize the conditions on the primitives of factor models (the factor loadings) that generate target parameters (minimizers of the statistical risk) that are *synthetic controls*. This result complements the existing literature on the asymptotic properties of synthetic controls under linear factor models by providing guidance on the set of data generating processes for which synthetic control estimators are best suited. We show that the target parameters can be estimated by MLE and derive rate conditions for uniform consistency as the number of time periods and donor units grow.

Second, we propose the Bayesian synthetic control as an alternative to perform inference for synthetic control methods. We derive a Bernstein-von Mises style result that states conditions under which the Bayesian synthetic control and the MLE estimator converge in the total variation sense. This result can be used to approximate the frequentist inference using the Bayesian synthetic control in large samples. Through simulations we show that this result might also be useful in settings in which the theoretical assumptions don't hold, highlighting the potential for theoretical extensions in future work.

Finally we apply the Bayesian synthetic control to study two important political economy interventions: that of the re-unification of a country and that of secession. We replicate using the Bayesian synthetic control the findings in Abadie et al. (2015) that the German re-unification lead to a significant decrease in West Germany's GDP and show that the Catalan unilateral declaration of independence of 2017 may have lead to a 1% decrease in GDP. The Bayesian synthetic control allows us to quantify uncertainty and in both applications we can conclude that the effects are unlikely to be zero.

---

[3]See the article in footnote 2 for reference.

In future work we aim to generalize the results in the paper to a larger class of models and extend the functionality of the *bsynth* R-package.

References

Abadie, A. (2021). Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, 59(2):391–425.

Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505.

Abadie, A., Diamond, A., and Hainmueller, J. (2015). Comparative politics and the synthetic control method. *American Journal of Political Science*, 59(2):495–510.

Abadie, A. and Gardeazabal, J. (2003). The economic costs of conflict: A case study of the basque country. *American Economic Review*, 93(1):113–132.

Abadie, A. and L'Hour, J. (2021). A penalized synthetic control estimator for disaggregated data. *Journal of the American Statistical Association*. Forthcoming.

Abadie, A. and Vives-i Bastida, J. (2022). Synthetic controls in action.

Amjad, M., Shah, D., and Shen, D. (2018). Robust synthetic control. *Journal of Machine Learning Research*, 19(22):1–51.

Arbour, D., Ben-Michael, E., Feller, A., Franks, A., and Raphael, S. (2021). Using multitask gaussian processes to estimate the effect of a targeted effort to remove firearms.

Athey, S., Bayati, M., Doudchenko, N., Imbens, G., and Khosravi, K. (2021). Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*. Forthcoming.

Barron, A. R. (1986). Entropy and the Central Limit Theorem. *The Annals of Probability*, 14(1):336 – 342.

Belloni, A., Chernozhukov, V., Chetverikov, D., and Kato, K. (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186(2):345–366.

Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N., and Scott, S. L. (2015). Inferring causal impact using Bayesian structural time-series models. *The Annals of Applied Statistics*, 9(1):247 – 274.

Chernozhukov, V., Wüthrich, K., and Zhu, Y. (2021). An exact and robust conformal inference method for counterfactual and synthetic controls. *Journal of the American Statistical Association*. Forthcoming.

Esteller-Moré, A. and Rizzo, L. (2022). The economic costs of a secessionist conflict: The case of catalonia. *Defence and Peace Economics*, 33(6):655–688.

Ferman, B. (2021). On the properties of the synthetic control estimator with many periods and many controls. *Journal of the American Statistical Association*. Forthcoming.

Ferman, B. and Pinto, C. (2021). Synthetic controls with imperfect pre-treatment fit. *Quantitative Economics*. Forthcoming.

Firpo, S. and Possebom, V. (2018). Synthetic Control Method: Inference, Sensitivity Analysis and Confidence Sets. *Journal of Causal Inference*, 6(2):1–26.

He, X. and Shao, Q.-M. (1996). A general Bahadur representation of M-estimators and its application to linear regression with nonstochastic designs. *The Annals of Statistics*, 24(6):2608 – 2630.

He, X. and Shao, Q.-M. (2000). On parameters of increasing dimensions. *Journal of Multivariate Analysis*, 73(1):120–135.

Hsiao, C., Steve Ching, H., and Ki Wan, S. (2012). A panel data approach for program evaluation: Measuring the benefits of political and economic integration of Hong Kong with mainland China. *Journal of Applied Econometrics*, 27(5):705–740.

Imbens, G. W. and Rubin, D. B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics*, 25(1):305 – 327.

Kim, S., Lee, C., and Gupta, S. (2020). Bayesian synthetic control methods. *Journal of Marketing Research*, 57(5):831–852.

Li, K. T. (2020). Statistical inference for average treatment effects estimated by synthetic control methods. *Journal of the American Statistical Association*, 115(532):2068–2083.

Pang, X., Liu, L., and Xu, Y. (2022). A bayesian alternative to synthetic control for comparative case studies. *Political Analysis*, 30(2):269–288.

Pinkney, S. (2021). An improved and extended bayesian synthetic control.

Ray, K. and van der Vaart, A. (2020). Semiparametric bayesian causal inference. *The Annals of Statistics*, 48(5).

Scott, S. L. and Varian, H. R. (2014). Predicting the present with bayesian structural time series. *International Journal of Mathematical Modelling and Numerical Optimisation*, 5(1-2):4–23.

Vaart, A. W. v. d. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Vives-i Bastida, J. (2022). Predictor selection for synthetic controls.

Appendix

A.1. Conditional Distribution Model Derivations

Suppose we have the following simple model independent of time:

$$Y_i = \lambda_i F_t + \epsilon_i,$$

where $F_t \sim N(0, \sigma^2)$ and $\epsilon_i \sim N(0, 1)$. Then it follows that $\mathbb{E}[Y_1] = 0$, $\mathbb{E}[\boldsymbol{Y}_J] = \boldsymbol{0}$, $cov(Y_i, Y_j) = \lambda_i \lambda_j \sigma^2$ and $var(Y_j) = 1 + \lambda_i^2 \sigma^2$. Then the joint distribution of ys is normal. We are interested in the conditional distribution of Y_1 given $\boldsymbol{y}_J = (y_2, \ldots, y_{J+1})$:

$$Y_1 | \boldsymbol{y}_J \sim N\left(\tilde{\mu}, \tilde{\Sigma}\right),$$

where

$$\tilde{\mu} = cov(Y_1, \boldsymbol{Y}_J) \Sigma_{(2, J+1)}^{-1} \boldsymbol{y}_J = \sum_{j=2}^{J+1} w_j(\boldsymbol{\lambda}, \sigma) y_j \tag{A.1}$$

$$\tilde{\Sigma} = var(Y_1) - cov(Y_1, \boldsymbol{Y}_J) \Sigma_{(2, J+1)}^{-1} cov(\boldsymbol{Y}_J, Y_1). \tag{A.2}$$

In this set up the data are repeated observations over time. For a pre-treatment period of length T_0 then we will be interested in estimators $\hat{w}_j$ such that our estimated counterfactual is given by:

$$\hat{Y}_{1t}(0) = \sum_{j=2}^{J+1} \hat{w}_j y_{jt}.$$

Start by noting that in this setting $\Sigma_{(2, J+1)}$ is positive definite and invertible. Hence, by the spectral theorem we can express it as a linear combination of eigenvalues and eigenvectors:

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

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 $s_{J+1} = 1 + \sum_{j=2}^{J+1} \lambda_j^2 \sigma^2$. Therefore, we can express $\tilde{\mu}$ as a linear combination of the data, $\boldsymbol{\lambda}$ and σ:

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

$$= \sum_{j=2}^{J+1} \sigma^2 \lambda_1 \sum_{i=2}^{J+1} \lambda_i \sum_{k=2}^{J+1} \frac{1}{s_k} [\boldsymbol{u}_k \boldsymbol{u}_k^T]_{ji} y_j.$$

By substituting the eigenvectors and eigenvalues we can find the closed form for the weights:

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

$$= \frac{\sigma^2 \lambda_1 \lambda_j}{1 + \sum_{j=2}^{J+1} \lambda_j^2 \sigma^2}.$$

Similarly, we can derive a closed form for the variance:

$$\tilde{\Sigma} = 1 + \lambda_1 \sigma^2 - \frac{\sigma^4 \lambda_1^2 \sum_{j=2}^{J+1} \lambda_j^2}{1 + \sigma^2 \sum_{j=2}^{J+1} \lambda_j^2}$$

$$= 1 + \lambda_1 \sigma^2 (1 - \sum_{j=2}^{J+1} w_j \lambda_j)$$

$$= 1 + \lambda_1 \sigma^2 - (1 + \sigma^2 \sum_{j=2}^{J+1} \lambda_j^2) \sum_{j+2}^{J+1} w_j^2.$$

A.2. Proof of Theorem 1

Observe that without loss of generality we can consider $V = I$. For notational convenience we drop the potential outcome subscript in what follows such that $\boldsymbol{Y}_1 \equiv \boldsymbol{Y}_1(0)$. Consider the objective function:

$$
\begin{aligned}
\mathbb{E}\left[\boldsymbol{Y}_1 - \boldsymbol{Y}_J' \boldsymbol{w})'(\boldsymbol{Y}_1 - \boldsymbol{y}_J' \boldsymbol{w})\right] &= \mathbb{E}\left[(\boldsymbol{Y}_1' \boldsymbol{Y}_1 - 2\boldsymbol{w}' \boldsymbol{y}_J \boldsymbol{Y}_1 + \boldsymbol{w}' \boldsymbol{y}_J \boldsymbol{y}_J' \boldsymbol{w}\right] \\
&= \mathbb{E}\left[\mathbb{E}[\boldsymbol{Y}_1' \boldsymbol{Y}_1 | \boldsymbol{y}_J] - 2\boldsymbol{w}' \boldsymbol{y}_J \mathbb{E}[\boldsymbol{Y}_1 | \boldsymbol{y}_J] + \boldsymbol{w}' \boldsymbol{y}_J \boldsymbol{y}_J' \boldsymbol{w}\right] \\
&= \mathbb{E}\left[\tilde{\boldsymbol{w}}' \boldsymbol{y}_J \boldsymbol{y}_J' \tilde{\boldsymbol{w}} + T_0 \tilde{\Sigma}^2 - 2\boldsymbol{w}' \boldsymbol{y}_J \boldsymbol{y}_J' \tilde{\boldsymbol{w}} + \boldsymbol{w}' \boldsymbol{y}_J \boldsymbol{y}_J' \boldsymbol{w}\right],
\end{aligned}
$$

where the steps follow from the derivations for the conditional multivariate normal and law of iterated expectations. The $\tilde{\Sigma}^2$ is a variance term that occurs due the sampling process. Given that the function is convex and expectations preserve linearity, conditional on the data, it follows that

$\tilde{\boldsymbol{w}}$ is a global minimizer of the expression.

A.3.   Proof of Theorem 2

Statement (1): we want to show that as $J \to \infty$

$$\mathbb{E}\left[(\boldsymbol{y}'_{JT_0+1}\tilde{\boldsymbol{w}} - \lambda_1 F_{T_0+1})^2\right] \to 0.$$

Expanding the expression conditional on F_{T_0+1} and given the assumptions and derivations for $\tilde{\boldsymbol{w}}$:

$$\mathbb{E}\left[(\boldsymbol{y}'_{JT_0+1}\tilde{\boldsymbol{w}} - \lambda_1 F_{T_0+1})^2 | F_{T_0+1}\right] = \frac{\lambda_1^2 \sigma^4}{(1 + \|\boldsymbol{\lambda}_J\|_2^2 \sigma^2)^2} \mathbb{E}\left[\left(F_{T_0+1}\|\boldsymbol{\lambda}_J\|_2^2 + \sum_j \lambda_j^2 \epsilon_{jT_0+1}\right)^2 | F_{T_0+1}\right]$$

$$+ \lambda_1^2 F_{T_0+1}^2 \left(\frac{1 - \sigma^2 \|\boldsymbol{\lambda}_J\|_2^2}{1 + \sigma^2 \|\boldsymbol{\lambda}_J\|_2^2}\right)$$

$$- \frac{2\sigma^2 \lambda_1^2}{1 + \sigma^2 \|\boldsymbol{\lambda}_J\|_2^2} \mathbb{E}\left[\sum_j \lambda_j \epsilon_{jT_0+1} | F_{T_0+1}\right]$$

$$= \frac{\lambda_1^2 \sigma^4}{(1 + \|\boldsymbol{\lambda}_J\|_2^2 \sigma^2)^2} \left(F_{T_0+1}^2 \|\boldsymbol{\lambda}_J\|_2^4 + \|\boldsymbol{\lambda}_J^2\|_2^2\right)$$

$$+ \lambda_1^2 F_{T_0+1}^2 \left(\frac{1 - \sigma^2 \|\boldsymbol{\lambda}_J\|_2^2}{1 + \sigma^2 \|\boldsymbol{\lambda}_J\|_2^2}\right)$$

Observe that there exists no value of $\|\boldsymbol{\lambda}_J\|_2^2$ for which the above expression is zero. By the law of iterated expectations this implies that convergence in mean squared can not be achieved unless $\sigma^2 \to 0$ or we condition on perfect pre-treatment fit.

Statement (2): By Markov's inequality consider:

$$\mathbb{P}(|\boldsymbol{y}'_{JT_0+1}\tilde{\boldsymbol{w}} - \lambda_1 F_{T_0+1}| \geq \epsilon) \leq \frac{\mathbb{E}[|\boldsymbol{y}'_{JT_0+1}\tilde{\boldsymbol{w}} - \lambda_1 F_{T_0+1}|]}{\epsilon}.$$

Under the assumption that independently $\epsilon_{jt} \sim N(0,1)$:

$$\mathbb{E}\left[|\boldsymbol{y}'_{JT_0+1}\tilde{\boldsymbol{w}} - \lambda_1 F_{T_0+1}| \, | F_{T_0+1}\right] = \mathbb{E}\left[\left|\frac{\sigma^2 \lambda_1}{1 + \|\boldsymbol{\lambda}_J\|_2^2 \sigma^2}\left(\|\boldsymbol{\lambda}_J\|_2^2 F_{T_0+1} + \sum_j \lambda_j \epsilon_{jT_0+1}\right) - \lambda_1 F_{T_0+1}\right| \, | F_{T_0+1}\right]$$

$$= \mathbb{E}\left[\left|\frac{-F_{T_0+1}\sigma^2 \lambda_1}{1 + \|\boldsymbol{\lambda}_J\|_2^2 \sigma^2} + \frac{\sigma^2 \lambda_1}{1 + \|\boldsymbol{\lambda}_J\|_2^2 \sigma^2} \sum_j \lambda_j \epsilon_{jT_0+1}\right| \, | F_{T_0+1}\right]$$

$$\leq \left| \frac{\sigma^2 \lambda_1 F_{T_0+1}}{1 + \|\boldsymbol{\lambda}_J\|_2^2 \sigma^2} \right| + \left| \frac{\sigma^2 \lambda_1}{1 + \|\boldsymbol{\lambda}_J\|_2^2 \sigma^2} \right| \mathbb{E} \left[\left| \sum_j \lambda_j \epsilon_{jT_0+1} \right| \, \middle| F_{T_0+1} \right]$$

$$\leq \left| \frac{\sigma^2 \lambda_1 F_{T_0+1}}{1 + \|\boldsymbol{\lambda}_J\|_2^2 \sigma^2} \right| + \left| \frac{\sigma^2 \lambda_1}{1 + \|\boldsymbol{\lambda}_J\|_2^2 \sigma^2} \right| \sum_j |\lambda_j| \frac{\sqrt{2}}{\sqrt{\pi}}.$$

The last step uses the mean of the half-normal distribution. Hence, by the law of iterated expectations convergence in probability holds if as $J \to \infty$, $\frac{1}{\|\boldsymbol{\lambda}_J\|_2^2} \sum_j |\lambda_j| \to 0$. Observe, that by Statement 1 we know that this proof would not work with the application of Chebyshev's inequality.

A.4. Proof of Theorem 3

The first part of the theorem follows directly from the derivation for $\tilde{\boldsymbol{w}}$ given A1-A2, setting

$$\frac{\sigma^2 \lambda_1 \lambda_j}{1 + \sum_{j=2}^{J+1} \lambda_j^2 \sigma^2} = 1,$$

and noting that the denominator is non-negative yields the result.

For statement 1, Condition (2) implies that given a λ_1 the λ_j lie in a sphere in $\mathbb{R}^J$. A sufficient condition for existence of such λ_j is that the discriminant for the second degree equation for each λ_j is positive. We exemplify this argument for $J = 2$. Condition (2) then requires

$$\lambda_2^2 - \lambda_1 \lambda_2 + 1/(2\sigma^2) + \lambda_3^2 - \lambda_1 \lambda_3 + 1/(2\sigma^2) = 0,$$

which has real roots when the discriminant of each second degree regression is non-negative. In both cases, the condition required is $\lambda_1^2 - 4/(2\sigma^2) \geq 0$. For an arbitrary J it follows that a sufficient condition for real roots is $\lambda_1^2 \geq 4/(J\sigma^2)$.

For statement 2, observe that given that $\|\boldsymbol{\lambda}_J\|_2^2 \geq 0$, we can re-write condition (2) as

$$1 - \lambda_1 \frac{\sum_j \lambda_j}{\|\boldsymbol{\lambda}_J\|_2^2} + \frac{1}{\|\boldsymbol{\lambda}_J\|_2^2 \sigma^2} = 0.$$

Given condition (1) it is without loss to write the second term as

$$\frac{\sum_j |\lambda_1 \lambda_j|}{\|\boldsymbol{\lambda}_J\|_2^2} \leq |\lambda_1| \frac{\sum_j |\lambda_j|}{\|\boldsymbol{\lambda}_J\|_2^2} \to 0$$

by the assumption $\frac{1}{\|\boldsymbol{\lambda}_J\|_2^2} \sum_j |\lambda_j| \to 0$. Given that this assumption implies $\|\boldsymbol{\lambda}_J\|_2^2 \to \infty$, as $J \to \infty$ condition (2) is violated as $1 \neq 0$.

For statement 3, recall that an odd function h satisfies that $h(-x) = -h(x)$ which we extend for multidimensional functions to mean this definition simultaneously holds for all components λ_j: $h(-\lambda_2, \ldots, -\lambda_{J+1}) = -h(\lambda_2, \ldots, \lambda_{J+1})$. When h is component-wise weakly increasing, this ensures that for all j, $sign(\lambda_j) = sign(h(\boldsymbol{\lambda}_J)) = sign(\lambda_1)$, so condition (1) is satisfied. As before, if $\|\boldsymbol{\lambda}_J\|_2^2 \to \infty$ condition (2) is satisfied when as $J \to \infty$

$$|\lambda_1| \frac{\sum_j |\lambda_j|}{\|\boldsymbol{\lambda}_J\|_2^2} \to 1,$$

where the λ_1 can be substituted in to get the desired result.

For statement 4, note that we can write the system of equations

$$\boldsymbol{\lambda}_J' \boldsymbol{w} \|\boldsymbol{\lambda}_J\|_1 = \|\boldsymbol{\lambda}_J\|_2^2, \quad \boldsymbol{w} \in \Delta^J,$$

as a linear programming problem. A sufficient condition for this program to have a solution is for $\frac{\|\boldsymbol{\lambda}_J\|_2^2}{\|\boldsymbol{\lambda}_J\|_1}$ to be in the convex hull of the $\boldsymbol{\lambda}_J$.

A.5. Proof of Theorem 4

Recall that under assumptions **A1-A2** by the weak law of large numbers we have that as $T_0 \to \infty$, $\frac{1}{T_0} \sum_t F_t^2 \xrightarrow{p} \sigma^2$, $\frac{1}{T_0} \sum_t F_t \xrightarrow{p} 0$, $\frac{1}{T_0} \sum_t \epsilon_{it}^2 \xrightarrow{p} 1$ and $\frac{1}{T_0} \sum_t \epsilon_{it} \xrightarrow{p} 0$. For notational convenience drop in what follows the MLE subscript in the weight estimator. By an application of l'hopital rule for the limit, observe that the sum of squares in the log-likelihood l_{T_0} is proportional to

$$
\begin{aligned}
\frac{1}{T_0} \sum_{t=1}^{T_0} \left(Y_{1t} - \sum_{j=2}^{J+1} \hat{w}_j Y_{jt} \right)^2 &= \frac{1}{T_0} \sum_{t=1}^{T_0} (\lambda_1 - \sum_{j=2}^{J+1} \hat{w}_j \lambda_j)^2 F_t^2 \\
&\quad + 2(\lambda_1 - \sum_{j=2}^{J+1} \hat{w}_j \lambda_j) F_t (\epsilon_{1t} - \sum_{j=2}^{J+1} \hat{w}_j \epsilon_{jt}) \\
&\quad + (\epsilon_{1t} - \sum_{j=2}^{J+1} \hat{w}_j \epsilon_{jt})^2 \\
&= \frac{1}{T_0} \sum_{t=1}^{T_0} (\lambda_1 - \sum_{j=2}^{J+1} \hat{w}_j \lambda_j)^2 \sigma^2 + 1 + \sum_{j=2}^{J+1} \hat{w}_j^2 + o_p(1).
\end{aligned}
$$

Minimizing the last expression is equivalent to minimizing the Ridge regression problem and given

that the problem is convex a global minimizer is:

$$\hat{w}_j = \frac{\sigma^2 \lambda_1 \lambda_j}{1 + \|\boldsymbol{\lambda}_J\|_2^2 \sigma^2} + o_p(1).$$

Hence, as $T_0 \to \infty$ we have that $\hat{\boldsymbol{w}}_{MLE} \overset{p}{\to} \tilde{\boldsymbol{w}}$. The second statement follows from the conditions given in Theorem 2 and the continuous mapping theorem.

[Sketch] For the third statement, note that under the model assumptions, the standard M-estimator regularity conditions (Vaart (1998)) are satisfied. By part (1) we have the estimator is consistent. Furthermore, by the CLT given that $\mathbb{E}[\nabla_{\boldsymbol{w}} l_{T_0}(\boldsymbol{\theta})|_{\tilde{\boldsymbol{w}}} = 0$ it follows that:

$$\sqrt{T_0} \nabla_{\boldsymbol{w}} l_{T_0}(\boldsymbol{\theta})|_{\tilde{\boldsymbol{w}}} \overset{d}{\to} N(0, V),$$

where $V = V_{T_0} T_0$. The proof follows from an application of the mean value theorem.

A.6. Proof of Theorem 5

This proof follows results in He and Shao (2000) and He and Shao (1996). A more modern treatment of similar results can also be found in Belloni et al. (2015). The proof considers the more general problem for $M-$estimators in linear models. Start by defining the objective function

$$\rho(\boldsymbol{w}) = \sum_t (y_{1t} - \boldsymbol{y}'_{Jt} \boldsymbol{w})^2,$$

and the score function

$$\phi(\boldsymbol{w}) = 2 \sum_t (y_{1t} - \boldsymbol{y}'_{Jt} \boldsymbol{w}) \boldsymbol{y}_{Jt}.$$

Focus on the first assumption:

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

where $0 < \lim \inf_{T_0} \sigma_{min}(D_{T_0}) \le \lim \sup_{T_0} \sigma_{max}(D_{T_0}) < \infty$. First note that under A1-A2 the matrix $\boldsymbol{y}_{Jt} \boldsymbol{y}'_{Jt}$ is invertible and has eigenvalues bounded away from zero, so this assumption is satisfied in our setting. Then, under the other assumptions,

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

$$\leq c \sum_t |\boldsymbol{y}'_{Jt}\boldsymbol{\alpha}||\boldsymbol{y}'_{Jt}(\hat{\boldsymbol{w}}_{MLE} - \tilde{\boldsymbol{w}})|^2$$

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

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

The above expression means that condition (C3) in He and Shao (2000) is satisfied. Next, we check that condition (C2) is satisfied. Observe that

$$\|\phi(\tilde{\boldsymbol{w}})\| \leq 2 \sum_t \|y_{1t} - \boldsymbol{y}'_{Jt}\tilde{\boldsymbol{w}}\|\|\boldsymbol{y}_{Jt}\| = O((T_0 J)^{1/2}),$$

given assumptions (2) and (3) in the theorem. By a similar argument (C1) is satisfied and (C3) implies that (C4) or (C5) are satisfied. Hence, we can apply Corollary 2.1 in He and Shao (2000) to get the desired result.

A.7. Proof of Corollary 5.1

To derive the first statement in Corollary 5.1 observe that

$$\boldsymbol{y}'_{Jt}(\tilde{\boldsymbol{w}} - \boldsymbol{w}) \leq \|\boldsymbol{y}_{Jt}\|^2 \|\tilde{\boldsymbol{w}} - \boldsymbol{w}\|^2 = O_p(J^2/T_0),$$

by condition 2 and result 1 in Theorem 5. A similar, approach to He and Shao (2000) can be followed to improve the rate to J/T_0 up to log terms.

To derive the second statement, we show that Theorem 5 applies when α is itself a stochastic sequence. Observe, that in contrast to most uniform consistency results, $\alpha \in \mathbb{R}^J$ instead of being in a J-dimensional ball. We will show that in our setting the boundedness assumptions on the DGP allows us to preserve stochastic equi-continuity, and so condition (C2) in He and Shao (2000) and in the proof of Theorem 5 applies. Consider the following quantity

$$\gamma(\boldsymbol{\alpha}) = T_0\boldsymbol{\alpha}'D_{T_0}(\hat{\boldsymbol{w}}_{MLE} - \tilde{\boldsymbol{w}}).$$

Observe that when $\boldsymbol{\alpha} \in S^J$, a ball of dimension J, by a standard maximal inequality for simplex random variables we get that this object can be bounded by $\sqrt{J T_0}$. It is without loss to substite

34

$\boldsymbol{\alpha}$ with $\boldsymbol{y}_{Jt}$ given assumption 2 in Theorem 5. Indeed, in that case we can also upper bound the quantity by the same rate. Overall, this implies that condition (C2) in He and Shao (2000) is satisfied and so the result follows.

A.8. Proof of Theorem 6

Recall that our Bayes model given an $i.i.d$ sample of size T_0 with $J+1$ units means that the posterior predictive distribution is normal with the following parameters:

1. Mean:
$$\mu_{T_0,J}^B = \frac{\sigma_y^2}{\sigma_y^2 + T_0 \sum_j \tau_j^2} \sum_t \boldsymbol{y}_{Jt}' \mu_J + \frac{\sum_j \tau_j^2}{\sigma_y^2 + T_0 \sum_j \tau_j^2} \sum_t y_{1t}.$$

2. Variance:
$$\Sigma_{T_0,J}^B = \frac{\sigma_y^2 \sum_j \tau_j^2}{\sigma_y^2 + T_0 \sum_j \tau_j^2}.$$

Suppose, as in the assumptions, that $\|\boldsymbol{\mu}_J\|^2 \to 0$ and $\|\lambda_1 - \boldsymbol{\lambda}_J'\boldsymbol{\mu}_J\| \to 0$ as $J \to \infty$. Under the DGP given by **A1**-**A2** we have that $y_{jt} = \lambda_j F_t + \epsilon_{jt}$, for $\epsilon_{jt} \sim_{i.i.d} N(0,1)$. It follows that

$$\boldsymbol{y}_{JT_0+1}'\boldsymbol{\mu}_J = \sum_j (\lambda_j F_{T_0+1} + \epsilon_{jT_0+1})\mu_j$$
$$= F_{T_0+1} \sum_j \lambda_j \mu_j + \sum_j \epsilon_{jT_0+1}\mu_j.$$

Next, consider the following expectation

$$\mathbb{E}[(\boldsymbol{y}_{JT_0+1}'\boldsymbol{\mu}_J - \lambda_1 F_{T_0+1})^2 | F_{T_0+1}] = \mathbb{E}[(F_{T_0+1}(\sum_j \lambda_j \mu_j - \lambda_1) + \sum_j \epsilon_{jT_0+1}\mu_j)^2 | F_{T_0+1}]$$
$$= \mathbb{E}[(F_{T_0+1}^2(\sum_j \lambda_j \mu_j - \lambda_1)^2 + (\sum_j \epsilon_{jT_0+1}\mu_j)^2 | F_{T_0+1}]$$
$$\leq \mathbb{E}[(F_{T_0+1}^2(\sum_j \lambda_j \mu_j - \lambda_1)^2 | F_{T_0+1}] + \sigma_\epsilon^2 \sum_j \mu_j^2$$
$$\to 0.$$

where the second and third inequalities follow from ϵ_{jt} being mean zero and $i.i.d.$ Under the assumptions on $\boldsymbol{\mu}_J$ it follows that this inequality goes to zero as $J \to \infty$. Therefore, the convergence in probability follows by Chebyshev's inequality.

Next, we show that the mean of the posterior predictive distribution $\mu_{T_0,J}^B$ converges to the same mean as the MLE estimator. First, note that under **A1**-**A2** and $\sum_j \tau_j^2 = O(T_0^{-\alpha})$ the second term

in $\mu_{T_0,J}^B$ is $o_p(1)$ as $\frac{1}{T_0} \sum_t y_{1t} \xrightarrow{p} 0$. The first term then converges to the treated unit factor loading by a similar argument as the first part of this proof given our convex recovery assumption.

To derive the Bernstein-von Mises result we start by noting that due to Pinsker's inequality

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

Hence, we procced in bounding the KL divergence. A useful result is Barron (1986), which provide conditions under which the KL divergence and CLT can be related. The following Lemma summarizes the result.

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

The conditions in Lemma A.1 are satisfied given our assumptions, the previous derivations and the results in Theorem 5. Therefore, we need to bound $D_{KL}(\Phi^*||Q^*)$. Given our Gaussian assumption we can bound this quantity by comparing the variances of the two distributions. The following Lemma gives an exact derivation of the KL divergence for Gaussian distributions.

Lemma A.2 (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,*

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

Observe that in our case the distributions are one dimensional so it is sufficient to show that as $T_0, J \to \infty$

$$\frac{\mathbb{V}(\Phi_{T_0,J}^{MLE})}{\Sigma_{T_0,J}^B} \to 1.$$

Starting with the Bayesian model, recall that $\sum_j \tau_j^2 = O(J^\alpha)$, so for the prediction period $T_0 + 1$,

$$\Sigma_{T_0+1,J}^B = \frac{\sigma_y^2}{\sigma_y^2(\sum_j \tau_j^2)^{-1} + 1} \to \sigma_y^2.$$

From Theorem 5 recall that $\sigma_\alpha^2 = (\mathbb{E}[\epsilon_{Jt}^2])^{-1}\boldsymbol{\alpha}' D_{T_0}^{-1}\boldsymbol{\alpha}$. Following, Corollary 5.1 when $\boldsymbol{\alpha}$ is given by $\boldsymbol{y}_{Jt}$ and $\frac{1}{T_0}\sum_t \boldsymbol{y}_{Jt}\boldsymbol{y}_{Jt}' = D_{T_0}$ observe that

$$
\begin{aligned}
\sigma_\alpha^2 &= (\sigma_\epsilon^2)\boldsymbol{\alpha}' D_{T_0}^{-1}\boldsymbol{\alpha} \\
&= (\sigma_\epsilon^2)T_0 \boldsymbol{y}_{JT_0+1}'(\boldsymbol{y}_{Jt}\boldsymbol{y}_{Jt}')^{-1}\boldsymbol{y}_{JT_0+1} \\
&= (\sigma_\epsilon^2)T_0 \mathrm{tr}((\boldsymbol{y}_{Jt}\boldsymbol{y}_{Jt}')^{-1}\boldsymbol{y}_{JT_0+1}\boldsymbol{y}_{JT_0+1}') \\
&= (\sigma_\epsilon^2)O(T_0),
\end{aligned}
$$

where the last step follows from Theorem 5's assumption 2 regarding the l_2 norm of $\boldsymbol{y}_J t$. This implies that the sample MLE variance is given by $1/T_0\sigma_\alpha^2 \to \sigma_\epsilon^2$. If the time series structure is such that $\mathrm{tr}((\boldsymbol{y}_{Jt}\boldsymbol{y}_{Jt}')^{-1}\boldsymbol{y}_{JT_0+1}\boldsymbol{y}_{JT_0+1}') = 1$ as $T_0, J \to \infty$, then it follows that a sufficient condition for
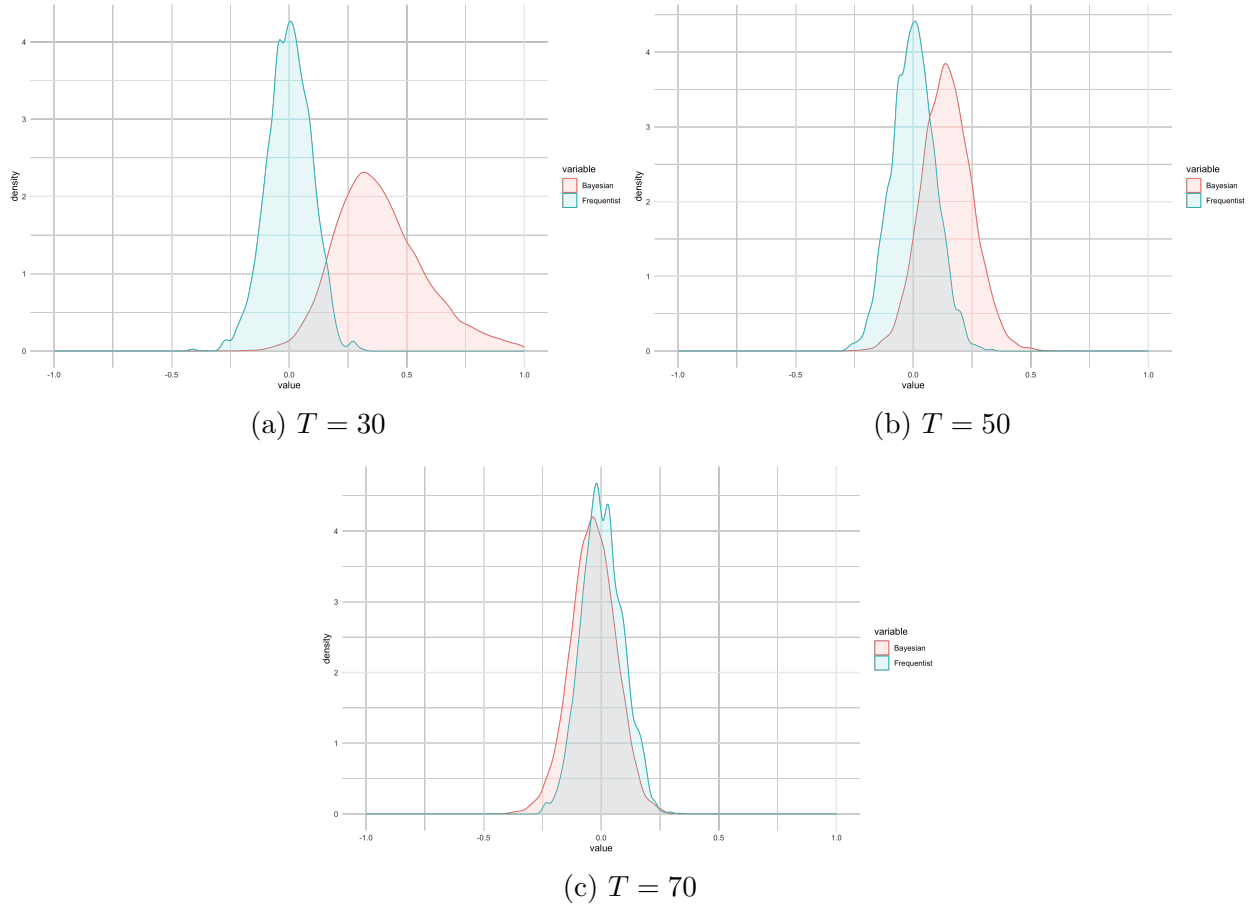
$$D_{KL}(\Phi^* \| Q^*) \to 0$$

is that $\sigma_y/\sigma_\epsilon \to 1$ as $J, T \to \infty$. Given that under **A1** - **A2** $\sigma_\epsilon = 1$, we can write $\sigma_y \to 1$. The trace condition is satisfied under assumption 2 of Theorem 5 and **A1**-**A2**, but it would also be satisfied under weaker conditions such as mixing time series regimes.

A.9.   Simulations for dense setting

Figure 1 highlights the BvM convergence in a sparse setting given by the grouped factor model 1 in which unit 2 is the only unit to have the same factor loading as unit 1. Figure 6 replicates Figure 1 for a denser setting in which we have four groups of five units with equal factor loadings. Unit 1 has the same factor loading as units 2 to 5. The Figure shows that convergence is achieved earlier than in the sparse case, when $T = 70$ instead that when $T = 1000$. This is in line with the theory that suggests that a requirement for convergence is that the weights are evenly distributed amongst many units. However, observe that the convergence rate is faster than expected as for 20 units the theory would suggest that at least $20^2 = 400$ time periods are necessary.

Figure 6: Convergence of frequentist and Bayesian coverage as $T \to \infty$.
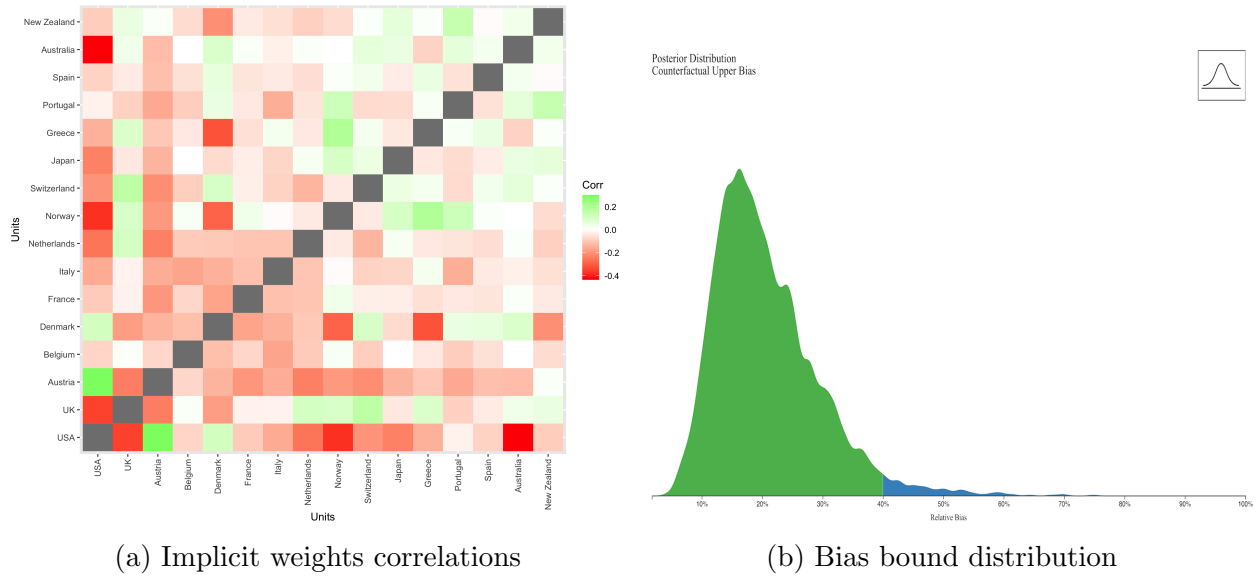

(a) $T = 30$

(b) $T = 50$

(c) $T = 70$

Notes: Kernel densities of the frequentist empirical distribution of the estimated treatment effect over 10000 draws and the Bayesian posterior distribution for one draw for different values of T_0. The potential outcomes are generated by the grouped factor model with 4 groups of 5 units with the same factor loadings and $\sigma = 0.25$.

A.10. German re-unification additional plots

In this section we provide additional materials for the study of the German re-unification. In Figure 8 we show the correlation between the implicit weight distributions and the distribution of an upper bound on the bias (as proposed by Abadie et al. (2010)) for each Bayesian draw.
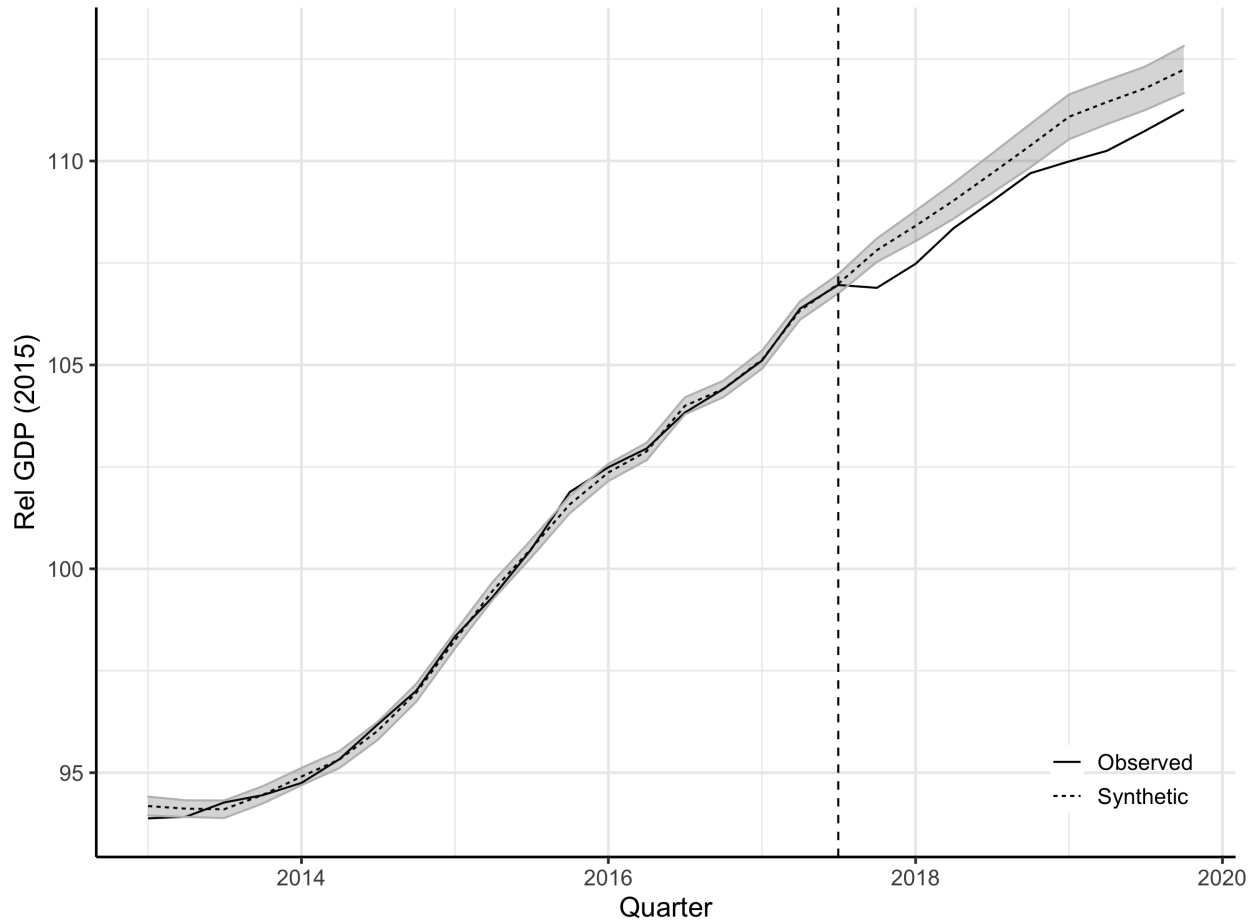
Figure 7: Additional Plots

(a) Implicit weights correlations

(b) Bias bound distribution

Notes: Panel (a) shows the correlations between implicit weights of the donor countries. Panel (b) shows the distribution of the bias bound term (computed as the sum of MAD of the pre-treatment outcomes) relative to the size of the mean treatment effect for each Bayesian draw.

A.11. Catalan secession movement additional plots

Figure 8: Additional Plot Synthetic Catalonia

Notes: Bayesian synthetic control for Catalonia excluding Madrid and Valencia from the donor pool. The average treatment effect relative to the third quarter of 2017 is -0.85%.