Munich Personal RePEc Archive

Synthetic Controls with Imperfect Pre-Treatment Fit

Ferman, Bruno and Pinto, Cristine

Sao Paulo School of Economics - FGV

23 November 2016

Online at https://mpra.ub.uni-muenchen.de/95524/
MPRA Paper No. 95524, posted 13 Aug 2019 05:13 UTC

Synthetic Controls with Imperfect Pre-Treatment Fit*

Bruno Ferman† Cristine Pinto‡

Sao Paulo School of Economics - FGV

First Draft: June, 2016

This Draft: August, 2019

Please click here for the most recent version

Abstract

We analyze the properties of the Synthetic Control (SC) and related estimators when the pre-treatment fit is imperfect. In this framework, we show that these estimators are generally biased if treatment assignment is correlated with unobserved confounders, even when the number of pre-treatment periods goes to infinity. Still, we also show that a modified version of the SC method can substantially improve in terms of bias and variance relative to standard methods. We also consider the properties of these estimators in settings with non-stationary common factors.

Keywords: synthetic control; difference-in-differences; policy evaluation; linear factor model
JEL Codes: C13; C21; C23

*We would like to thank Alberto Abadie, Josh Angrist, Ivan Canay, Marcelo Fernandes, Jin Hahn, Guido Imbens, Aureo de Paula, Tong Li, Victor Filipe Martins-da-Rocha, Ricardo Masini, Pedro Sant'Anna, Rodrigo Soares and conference and seminar participants at UC Berkeley, UC Santa Barbara, USC, UDELAR, UCL, University of Bristol, USP, George Washington University, University of Miami, PUC-Chile, IPEA, PUC-Rio, California Econometrics Conference 2016, Bristol Econometric Study Group 2016, IAAE conference 2017, EEA conference 2017, LACEA 2017, and the MIT Conference on Synthetic Controls and Related Methods for comments and suggestions. We also thank Deivis Angeli and Luis Alvarez for excellent research assistance. Bruno Ferman gratefully acknowledges financial support from FAPESP. An earlier version of this paper circulated under the title "Revisiting the Synthetic Control Estimator".

†E-mail: bruno.ferman@fgv.br. Address: Sao Paulo School of Economics, FGV, Rua Itapeva no. 474, Sao Paulo - Brazil, 01332-000.

‡E-mail: cristine.pinto@fgv.br. Address: Sao Paulo School of Economics, FGV, Rua Itapeva no. 474, Sao Paulo - Brazil, 01332-000.

1 Introduction

In a series of influential papers, Abadie and Gardeazabal (2003), Abadie et al. (2010), and Abadie et al. (2015) proposed the Synthetic Control (SC) method as an alternative to estimate treatment effects in comparative case studies when there is only one treated unit. The main idea of the SC method is to use the pre-treatment periods to estimate weights such that a weighted average of the control units reconstructs the pre-treatment outcomes of the treated unit, and then use these weights to compute the counterfactual of the treated unit in case it were not treated. According to Athey and Imbens (2017), *"the simplicity of the idea, and the obvious improvement over the standard methods, have made this a widely used method in the short period of time since its inception"*, making it "*arguably the most important innovation in the policy evaluation literature in the last 15 years*". As one of the main advantages that helped popularize the method, Abadie et al. (2010) derive conditions under which the SC estimator would allow confounding unobserved characteristics with time-varying effects, as long as there exist weights such that a weighted average of the control units fits the outcomes of the treated unit for a long set of pre-intervention periods.

In this paper, we analyze, in a linear factor model setting, the properties of the SC and other related estimators when the pre-treatment fit is imperfect.[1] In a model with "stationary" common factors and a fixed number of control units (J), we show that the SC weights converge in probability to weights that do *not*, in general, reconstruct the factor loadings of the treated unit when the number of pre-treatment periods (T_0) goes to infinity.[2,3] This happens because, in this setting, the SC weights converge to weights that simultaneously attempt to match the factor loadings of the treated unit *and* to minimize the variance of a linear combination of the transitory shocks. Therefore, weights that reconstruct the factor loadings of the treated unit are not generally the solution to this problem, even if such weights exist. While in many applications T_0 may not be large enough to justify large-T_0 asymptotics (e.g. Doudchenko and Imbens (2016)), our results can also be interpreted as the SC weights not converging to weights that reconstruct the factor loadings of the treated unit, when the pre-treatment fit is imperfect, *even when* T_0 *is large*.

As a consequence, the SC estimator is, in this setting with an imperfect pre-treatment fit, biased if treatment assignment is correlated with the unobserved heterogeneity, even when the number of pre-treatment periods goes to infinity. The intuition is the following: if treatment assignment is correlated with common factors in the post-treatment periods, then we would need a SC unit that

[1]We refer to "imperfect pre-treatment fit" when there may be no set of weights such that a weighted average of the outcomes of the control unit perfectly fits the outcome of the treated unit for all pre-treatment periods. The perfect pre-treatment fit condition is presented in equation 2 of Abadie et al. (2010).

[2]We refer to "stationary" in quotation marks because we only need the assumption that pre-treatment averages of the first and second moments of the common factors converge when the number of pre-treatment periods goes to infinity for this result.

[3]We focus on the SC specification that uses the outcomes of all pre-treatment periods as predictors. Specifications that use the average of the pre-treatment periods outcomes and other covariates as predictors are also considered in Appendix A.5.

is affected in exactly the same way by these common factors as the treated unit, but did not receive the treatment. This would be attained with weights that reconstruct the factor loadings of the treated unit. However, since the SC weights do not converge to weights that satisfy this condition when the pre-treatment fit is imperfect, the distribution of the SC estimator will still depend on the common factors, implying in a biased estimator when selection depends on the unobserved heterogeneity.[4] Our results are not as conflicting with the results from Abadie et al. (2010) as it might appear at first glance. The asymptotic bias of the SC estimator, in our framework, goes to zero when the variance of the transitory shocks is small. This is the case in which one should expect to have a close-to-perfect pre-treatment match, which is the setting the SC estimator was originally designed for. Our theory complements the theory developed by Abadie et al. (2010), by considering the properties of the SC estimator when the pre-treatment fit is imperfect.

The asymptotic bias we derive for the SC estimator when the number of control units is fixed does not rely on the fact that the SC unit is constrained to convex combinations of control units, so it also applies to other related panel data approaches that have been studied in the context of an imperfect pre-treatment fit, such as Hsiao et al. (2012), Li and Bell (2017), Carvalho et al. (2018), Carvalho et al. (2016), and Masini and Medeiros (2016). We show that these papers implicitly rely on assumptions that exclude the possibility of selection on unobservables.[5] Therefore, an important contribution of our paper is to clarify what selection on unobservables means in this setting, and to show that these estimators are generally biased if treatment assignment is correlated with the unobserved heterogeneity.

One important implication of the SC restriction to convex combinations of the control units is that the SC estimator, in this setting with an imperfect pre-treatment fit, may be biased even if treatment assignment is only correlated with time-invariant unobserved variables, which is essentially the identification assumption of the difference-in-differences (DID) estimator. We therefore consider a modified SC estimator, where we demean the data using information from the pre-intervention period, and then construct the SC estimator using the demeaned data.[6] An advantage of demeaning is that it is possible to, under some conditions, show that the SC estimator dominates

[4] Ando and Sävje (2013) point out that the SC estimator can be biased if there is no set of weights that reconstructs the factor loadings of the treated unit. However, they do not analyze in detail the minimization problem that is used to estimate the SC weights. In contrast, we show that this minimization problem inherently leads to weights that do not reconstruct the factor loadings of the treated unit, *even if such weights exist*. Moreover, we show that this potential problem persists even when the number of pre-treatment periods is large.

[5] Chernozhukov et al. (2018) suggest an alternative estimator and analyze its properties in a setting with both large J and T. As we explain in more detail in Section 5, they also rely on an assumption that essentially excludes the possibility of selection on unobservables. Since they consider a setting with both large J and T, however, it is possible that their estimator is consistent when there is selection on unobservables under conditions similar to the ones considered by Ferman (2019).

[6] Demeaning the data before applying the SC estimator is equivalent to relaxing the non-intercept constraint, as suggested, in parallel to our paper, by Doudchenko and Imbens (2016). We formally analyze the implication of this modification to the bias of the SC estimator. The estimator proposed by Hsiao et al. (2012) relaxes not only the the non-intercept but also the adding-up and non-negativity constraints.

the DID estimator in terms of variance and bias.[7]

Finally, we consider the properties of the SC and related estimators in a model with a combination of $I(1)$ common factors and/or deterministic polynomial trends, in addition to $I(0)$ common factors. We show that, in this setting, the demeaned SC weights converges to weights that reconstruct the factor loadings associated to the non-stationary common trends of the treated unit, but that generally fails to reconstruct the factor loadings associated with the $I(0)$ common factors.[8] Therefore, non-stationary common trends will not generate asymptotic bias in the demeaned SC estimator, but we need that treatment assignment is uncorrelated with the $I(0)$ common factors to guarantee asymptotic unbiasedness. Given that, we recommend that researchers applying the SC method should *also* assess the pre-treatment fit of the SC estimator after de-trending the data.

Our paper is related to a recent literature that analyzes the properties of the SC estimator and of generalizations of the method. Ben-Michael et al. (2018) derive finite-sample bounds on the bias of the SC estimator, and show that the bounds they derive do not converge to zero when J is fixed and $T_0 \to \infty$. This is consistent with our results, but does not directly imply that the SC estimator is asymptotically biased when J is fixed and $T_0 \to \infty$. In contrast, our result on the asymptotic bias of the SC estimator imply that it would be impossible to derive bounds that converge to zero in this case. Moreover, we show the conditions under which the estimator is asymptotically biased. Our results are also valid for other related methods, as the ones considered by Hsiao et al. (2012), Li and Bell (2017), Carvalho et al. (2018) and Carvalho et al. (2016). Amjad et al. (2017) suggest an interesting de-noising algorithm that leads to a consistent estimator even when the number of control units is fixed. Their method, however, relies on the assumption that transitory shocks are independent across units and time. Under this assumption, an IV-like SC estimator we present in Appendix A.5.4 would also be valid. We do not focus on this strategy because it relies on the assumption that the time-series correlation of the outcome variable can only be driven by serial correlation in the common factors. Powell (2017) proposes a 2-step estimation in which the SC unit is constructed based on the fitted values of the outcomes on unit-specific time trends. However, we show that the demeaned SC method is already very efficient in controlling for polynomial time trends, so the possibility of asymptotic bias in the SC estimator would come from correlation between treatment assignment and common factors beyond such time trends, which would not generally be captured in this strategy. When both J and T_0 diverge, Gobillon and Magnac (2016), Xu (2017), Athey et al. (2017), and Arkhangelsky et al. (2018) provide alternative estimation methods that are asymptotically valid when the number of both pre-treatment periods

[7] We also provide in Appendix A.5.4 an instrumental variables estimator for the SC weights that generates an asymptotically unbiased SC estimator under additional assumptions on the error structure, which would be valid if, for example, the idiosyncratic error is serially uncorrelated *and* all the common factors are serially correlated. The idea behind this strategy is similar to the strategy outlined by Heckman and Scheinkman (1987).

[8] We assume existence of weights that perfectly reconstructs the factor loadings of the treated unit associated with the non-stationary trends. In a setting with $\mathcal{I}(1)$ common factors, this is equivalent to assume that the vector of outcomes is cointegrated. If there were no set of weights that satisfies this condition, then the asymptotic distribution of the SC estimator would depend on the non-stationary common trends.

and controls increase.[9] Finally, Ferman (2019) provides conditions under which the original SC estimator is also asymptotically unbiased in this setting with a large number of pre-treatment periods and a large number of control units.

The remainder of this paper proceeds as follows. We start Section 2 with a brief review of the SC estimator. We highlight in this section that we rely on different assumptions relative to Abadie et al. (2010). In Section 3, we show that, in a model such that pre-treatment averages of the first and second moments of the common factors converge, the SC estimator is, in our framework, generally asymptotically biased if treatment assignment is correlated with the unobserved heterogeneity. In Section 4, we contrast the SC estimator with the DID estimator, and consider the demeaned SC estimator. In Section 5, we show that our main results also apply to other related panel data approaches that have been considered in the literature. In Section 6, we consider a setting in which pre-treatment averages of the common factor diverge. In Section 7, we present a particular class of linear factor models in which we consider the asymptotic properties of the SC estimator, and MC simulations with finite T_0. We conclude in Section 8.

2 Base Model

Suppose we have a balanced panel of $J+1$ units indexed by $j = 0, ..., J$ observed on a total of T periods. We want to estimate the treatment effect of a policy change that affected only unit $j = 0$, and we have information before and after the policy change. Let T_0 be the number of pre-intervention periods. Since we want to consider the asymptotic behavior of the SC estimator when $T_0 \to \infty$, we label the periods as $t \in \{-T_0+1, ..., -1, 0, 1, ..., T_1\}$, where $T_1 = T - T_0$ is the total number of post-treatment periods. Let $\mathcal{T}_0$ ($\mathcal{T}_1$) be the set of time indices in the pre-treatment (post-treatment) periods. The potential outcomes are given by

$$\begin{cases} y_{jt}^N = \delta_t + \lambda_t \mu_j + \varepsilon_{jt} \\ y_{jt}^I = \alpha_{jt} + y_{jt}^N, \end{cases} \tag{1}$$

where δ_t is an unknown common factor with constant factor loadings across units, λ_t is a $(1 \times F)$ vector of common factors, μ_j is a $(F \times 1)$ vector of unknown factor loadings, and the error terms ε_{jt} are unobserved transitory shocks. We only observe $y_{jt} = d_{jt} y_{jt}^I + (1 - d_{jt}) y_{jt}^N$, where $d_{it} = 1$ if unit i is treated at time t. Since for most results we hold the number of units $(J+1)$ fixed and look at asymptotics when the number of pre-treatment periods goes to infinity, we treat the vector of unknown factor loads (μ_j) as fixed and the common factors (λ_t) as random variables. Alternatively, we can think that all results are conditional on $\{\mu_j\}_{j=0}^J$. In order to simplify the exposition of our main results, we consider the model without observed covariates Z_j. In Appendix Section A.5.2 we

[9] Bai (2009) and Moon and Weidner (2015) consider the asymptotic properties of estimators in linear factor models when the number of time periods and the number of cross-sectional units jointly go to infinity, without restricting to the particular case of estimation of treatment effects.

consider the model with covariates. The main goal of the SC method is to estimate the effect of the treatment for unit 0 for each post-treatment t, that is $\{\alpha_{01}, ..., \alpha_{0T_1}\}$.

Since the SC estimator is only well defined if it actually happened that one unit received treatment in a given period, all results of the paper are conditional on that. Let $D(j,t)$ be a dummy variable equal to 1 if unit j starts to be treated after period t, while all other units do not receive treatment. Without loss of generality, we consider a realization of the data in which unit 0 is treated and that treatment starts after $t = 0$, so $D(0,0) = 1$. Assumption 1 defines the sample a researcher observes in a SC application.

Assumption 1 (conditional sample) We observe a realization of $\{y_{0t}, ..., y_{Jt}\}_{t=-T_0+1}^{T_1}$ conditional on $D(0,0) = 1$.

We also impose that the treatment assignment is not informative about the first moment of the transitory shocks.

Assumption 2 (transitory shocks) $\mathbb{E}[\varepsilon_{jt}|D(0,0) = 1] = \mathbb{E}[\varepsilon_{jt}] = 0$ for all j and t.

Assumption 2 implies that transitory shocks are mean-independent from the treatment assignment. However, we still allow for the possibility that treatment assignment to unit 0 is correlated with the unobserved common factors. More specifically, we allow for $\mathbb{E}[\lambda_t|D(0,0) = 1] \neq \mathbb{E}[\lambda_t]$ for any t. While λ_t is a common shock, the fact that unit 0 is treated can still be informative about λ_t, because we are fixing (or conditioning on) μ_0. Suppose that the treatment is more likely to happen for unit j at time t if $\lambda_t\mu_j$ is high. In this case, the fact that unit 0 is treated after $t = 0$ is informative that $\lambda_t\mu_0$ should be high for $t \geq 0$ if λ_t is serially correlated. Since we are conditioning on μ_0, this in turn implies that the common factors that strongly affect unit 0 that we expect to be particularly high given that unit 0 is the treated one. As an illustration, consider a simple example in which there are two common factors $\lambda_t = [\lambda_t^1\ \lambda_t^2]$, with $\mu_j = (1,0)$ for $j = 0, ..., \frac{J}{2}$ and $\mu_j = (0,1)$ for $j = \frac{J}{2} + 1, ..., J$. Under these conditions, the fact that unit 0 is treated after $t = 0$ is informative about the common factor λ_t^1, because unit 0 is only affected by the first common factor. In this case, one should expect $\mathbb{E}[\lambda_t^1|D(0,0) = 1] > \mathbb{E}[\lambda_t^1]$ for $t > 0$. The assumptions we make are essentially the same as the ones considered by, for example, Gobillon and Magnac (2016) and Ben-Michael et al. (2018) (in their Section 4.1), where they assume unconfoundness conditional on the unobserved factor loadings. The difference is that we condition on μ_j, while they condition on λ_t. However, the essence of the assumptions in both cases are the same, in that we allow treatment assignment to be informative about the structure $\lambda_t\mu_j$, while the transitory shocks ϵ_{jt} are uncorrelated with treatment assignment.

Let $\boldsymbol{\mu} \equiv [\mu_1 \dots \mu_J]'$ be the $J \times F$ matrix that contains the information on the factor loadings of all control units, and $\mathbf{y}_t \equiv (y_{1t}, \dots, y_{Jt})$ and $\boldsymbol{\varepsilon}_t \equiv (\varepsilon_{1t}, \dots, \varepsilon_{Jt})$ be $J \times 1$ vectors with information on the control units' outcomes and transitory shocks at periods t. We define Φ as the set of weights such that a weighted average of the control units absorbs all time correlated shocks of unit 0, $\lambda_t\mu_0$. Following the original SC papers, we start restricting to convex combinations of the control units.

Therefore, $\Phi = \{\mathbf{w} \in \Delta^{J-1} \mid \mu_0 = \boldsymbol{\mu}'\mathbf{w}\}$, where $\Delta^{J-1} \equiv \{(w_1, ..., w_J) \in \mathbb{R}^J | w_j \geq 0 \text{ and } \sum_{j=1}^J w_j = 1\}$. Assuming $\Phi \neq \varnothing$, if we knew $\mathbf{w}^* \in \Phi$, then we could consider an infeasible SC estimator using these weights, $\hat{\alpha}_{0t}^* = y_{0t} - \mathbf{y}_t'\mathbf{w}^*$. For a given $t > 0$, we would have

$$\hat{\alpha}_{0t}^* = y_{0t} - \mathbf{y}_t'\mathbf{w}^* = \alpha_{0t} + \left(\varepsilon_{0t} - \boldsymbol{\varepsilon}_t'\mathbf{w}^*\right). \tag{2}$$

We consider the expected value of $\hat{\alpha}_{0t}^*$ conditional on $D(0,0) = 1$ (Assumption 1). Therefore, under Assumption 2, $\mathbb{E}[\hat{\alpha}_{0t}^* | D(0,0) = 1] = \alpha_{0t}$, which implies that this infeasible SC estimator is unbiased. Intuitively, the infeasible SC estimator constructs a SC unit for the counterfactual of y_{0t} that is affected in the same way as unit 0 by each of the common factors (that is, $\mu_0 = \boldsymbol{\mu}'\mathbf{w}$), but did not receive treatment. Therefore, the only difference between unit 0 and this SC unit, beyond the treatment effect, would be given by the transitory shocks, which are assumed not related to the treatment assignment (Assumption 2). This guarantees that a SC estimator, using these infeasible weights, provides an unbiased estimator. Since there might be multiple weights in Φ, we define the infeasible SC estimator from equation 2 considering $\mathbf{w}^* \in \Phi$ that minimizes $var(\hat{\alpha}_{0t}^*)$ for cases in which $\Phi \neq \varnothing$.

It is important to note that Abadie et al. (2010) do note make any assumption on $\Phi \neq \varnothing$. Instead, they consider that there is a set of weights $\widetilde{\mathbf{w}}^* \in \Delta^{J-1}$ that satisfies $y_{0t} = \mathbf{y}_t'\widetilde{\mathbf{w}}^*$ for all $t \in \mathcal{T}_0$. While subtle, this reflects a crucial difference between our setting and the setting considered in the original SC papers. Abadie et al. (2010) and Abadie et al. (2015) consider the properties of the SC estimator conditional on having a perfect pre-intervention fit. As stated by Abadie et al. (2015), they *"do not recommend using this method when the pretreatment fit is poor or the number of pretreatment periods is small"*. Abadie et al. (2010) provide conditions under which existence of $\widetilde{\mathbf{w}}^* \in \Delta^{J-1}$ such that $y_{0t} = \mathbf{y}_t'\widetilde{\mathbf{w}}^*$ for all $t \in \mathcal{T}_0$ (for large T_0) implies that $\mu_0 \approx \boldsymbol{\mu}'\widetilde{\mathbf{w}}^*$. In this case, the bias of the SC estimator would be bounded by a function that goes to zero when T_0 increases. We depart from the original SC setting in that we consider a setting with imperfect pre-treatment fit, meaning that we do not assume existence of $\widetilde{\mathbf{w}}^* \in \Delta^{J-1}$ such that $y_{0t} = \mathbf{y}_t'\widetilde{\mathbf{w}}^*$ for all $t \in \mathcal{T}_0$.[10] The motivation to analyze the SC method in our setting is that the SC method can still provide important improvement relative to alternative methods, even if the pre-intervention fit is imperfect.

In order to implement their method, Abadie et al. (2010) recommend a minimization problem using the pre-intervention data to estimate the SC weights. They define a set of K predictors where X_0 is a $(K \times 1)$ vector containing the predictors for the treated unit, and X_C is a $(K \times J)$ matrix of economic predictors for the control units.[11] The SC weights are estimated by minimizing $||X_0 - X_C\mathbf{w}||_V$ subject to $\mathbf{w} \in \Delta^{J-1}$, where V is a $(K \times K)$ positive semidefinite matrix. They

[10] Abadie et al. (2010) assume that such weights also provide perfect balance in terms of observed covariates. Botosaru and Ferman (2019) analyze the case in which the perfect balance on covariates assumption is dropped, but there is still perfect balance on pre-treatment outcomes. In Appendix A.5 we consider the case in which covariates are used in a setting with imperfect pre-treatment fit on both pre-treatment outcomes and covariates.

[11] Predictors can be, for example, linear combinations of the pre-intervention values of the outcome variable or other covariates not affected by the treatment.

discuss different possibilities for choosing the matrix V, including an iterative process where V is chosen such that the solution to the $||X_0 - X_C\mathbf{w}||_V$ optimization problem minimizes the pre-intervention prediction error. In other words, let $\mathbf{Y}_0$ be a $(T_0 \times 1)$ vector of pre-intervention outcomes for the treated unit, while $\mathbf{Y}_C$ be a $(T_0 \times J)$ matrix of pre-intervention outcomes for the control units. Then the SC weights would be chosen as $\widehat{\mathbf{w}}(V^*)$ such that V^* minimizes $||\mathbf{Y}_0 - \mathbf{Y}_C\widehat{\mathbf{w}}(V)||$.

We focus on the case where one includes all pre-intervention outcome values as predictors. In this case, the matrix V that minimizes the second step of the nested optimization problem would be the identity matrix (see Kaul et al. (2015) and Doudchenko and Imbens (2016)), so the optimization problem suggested by Abadie et al. (2010) to estimate the weights simplifies to

$$\widehat{\mathbf{w}} \quad = \quad \underset{\mathbf{w}\in\Delta^{J-1}}{\operatorname{argmin}} \frac{1}{T_0} \sum_{t\in\mathcal{T}_0} \left[y_{0t} - \mathbf{y}_t'\mathbf{w}\right]^2. \tag{3}$$

Ferman et al. (2017) provide conditions under which the SC estimator using all pre-treatment outcomes as predictors will be asymptotically equivalent, when $T_0 \to \infty$, to any alternative SC estimator such that the number of pre-treatment outcomes used as predictors goes to infinity with T_0, even for specifications that include other covariates. Therefore, our results are also valid for these SC specifications under these conditions. In Appendix A.5 we also consider SC estimators using (1) the average of the pre-intervention outcomes as predictor, and (2) other time-invariant covariates in addition to the average of the pre-intervention outcomes as predictors.

3 Model with "stationary" common factors

We start assuming that pre-treatment averages of the first and second moments of the common factors and the transitory shocks converge. Let $\epsilon_t = (\varepsilon_{0t}, ..., \varepsilon_{Jt})$.

Assumption 3 (convergence of pre-treatment averages) Conditional on $D(0,0) = 1$, $\frac{1}{T_0}\sum_{t\in\mathcal{T}_0}\lambda_t \overset{p}{\to} \omega_0$, $\frac{1}{T_0}\sum_{t\in\mathcal{T}_0}\epsilon_t \overset{p}{\to} 0$, $\frac{1}{T_0}\sum_{t\in\mathcal{T}_0}\lambda_t'\lambda_t \overset{p}{\to} \Omega_0$ positive semi-definite, $\frac{1}{T_0}\sum_{t\in\mathcal{T}_0}\epsilon_t\epsilon_t' \overset{p}{\to} \sigma_\varepsilon^2 I_{J+1}$, and $\frac{1}{T_0}\sum_{t\in\mathcal{T}_0}\epsilon_t\lambda_t \overset{p}{\to} 0$ when $T_0 \to \infty$.

Assumption 3 allows for serial correlation for both transitory shocks and common factors. We assume $\frac{1}{T_0}\sum_{t\in\mathcal{T}_0}\epsilon_t\epsilon_t' \overset{p}{\to} \sigma_\varepsilon^2 I_{J+1}$ in order to simplify the exposition of our results. However, this can be easily replaced by $\frac{1}{T_0}\sum_{t\in\mathcal{T}_0}\epsilon_t\epsilon_t' \overset{p}{\to} \Sigma$ for any positive definite $(J+1)\times(J+1)$ matrix Σ, so that transitory shocks are correlated across j. Assumption 3 would be satisfied if the processes ϵ_t and λ_t, conditional on $D(0,0) = 1$, are weakly stationary and second order ergodic in the pre-treatment period, and ϵ_t and λ_t are independent. However, such assumption would be too restrictive, and would not allow for important possibilities in the treatment selection process. Recall that Assumption 2 allows for $\mathbb{E}[\lambda_t|D(0,0) = 1] \neq \mathbb{E}[\lambda_t]$, even for $t \leq 0$, which will happen if treatment assignment to unit 1 is correlated with common factors before treatment starts. In this case, it would be too restrictive to impose the assumption that, conditional on $D(0,0) = 1$, λ_t is stationary, even if only for the pre-treatment periods.

We show first that, when the number of control units is fixed, $\widehat{\mathbf{w}}$ converges to

$$\bar{\mathbf{w}} = \underset{\mathbf{w}\in\Delta^{J-1}}{\text{argmin}} \left\{ \sigma_\varepsilon^2 \left(1+\mathbf{w}'\mathbf{w}\right) + \left(\mu_0 - \boldsymbol{\mu}'\mathbf{w}\right)' \Omega_0 \left(\mu_0 - \boldsymbol{\mu}'\mathbf{w}\right) \right\}, \tag{4}$$

where, in general, $\mu_0 \neq \boldsymbol{\mu}'\bar{\mathbf{w}}$.

Proposition 1 Under Assumptions 1, 2 and 3, $\widehat{\mathbf{w}} \overset{p}{\to} \bar{\mathbf{w}}$ when $T_0 \to \infty$, where $\mu_0 \neq \boldsymbol{\mu}'\bar{\mathbf{w}}$, unless $\sigma_\varepsilon^2 = 0$ or $\exists \mathbf{w} \in \Phi | \mathbf{w} \in \underset{\mathbf{w}\in\Delta^{J-1}}{\text{argmin}} \{\mathbf{w}'\mathbf{w}\}$

Proof. Details in Appendix A.1.1 ■

The intuition of Proposition 1 is that we can treat the SC weights as an M-estimator, so we have that $\widehat{\mathbf{w}}$ converges in probability to $\bar{\mathbf{w}}$, defined in (4). This objective function has two parts. The first one reflects that different choices of weights will generate different weighted averages of the idiosyncratic shocks ε_{jt}. In this simpler case, if we consider the specification that restricts weights to sum one, then this part would be minimized when we set all weights equal to $\frac{1}{J}$. The second part reflects the presence of common factors λ_t that would remain after we choose the weights to construct the SC unit. If $\Phi \neq \varnothing$, then we can set this part equal to zero by choosing $\mathbf{w}^* \in \Phi$. Now start from $\mathbf{w}^* \in \Phi$ and move in the direction of weights that minimize the first part of this expression. Since $\mathbf{w}^* \in \Phi$ minimizes the second part, there is only a second order loss in doing so. On the contrary, since we are moving in the direction of weights that minimize the first part, there is a first order gain in doing so. This will always be true, unless $\sigma_\varepsilon^2 = 0$ or $\exists \mathbf{w} \in \Phi$ such that $\mathbf{w} \in \text{argmin}_{\mathbf{w}\in\Delta^{J-1}} \{\mathbf{w}'\mathbf{w}\}$. Therefore, the SC weights will not generally converge to weights that reconstruct the factor loadings of the treated unit. If $\Phi = \varnothing$, then Proposition 1 trivially holds. Another intuition for this result is that the outcomes of the controls are proxy variables for the factor loadings, but they are measured with error. We present this interpretation in more detail in Appendix A.2.

For a given $t > 0$, the SC estimator is given by

$$\hat{\alpha}_{0t} = y_{0t} - \mathbf{y}_t'\widetilde{\mathbf{w}} \overset{p}{\to} \alpha_{0t} + \left(\varepsilon_{0t} - \boldsymbol{\varepsilon}_t'\bar{\mathbf{w}}\right) + \lambda_t \left(\mu_0 - \boldsymbol{\mu}'\bar{\mathbf{w}}\right) \text{ when } T_0 \to \infty. \tag{5}$$

Note that $\hat{\alpha}_{0t}$ converges in distribution to the parameter we want to estimate (α_{0t}) plus linear combinations of contemporaneous transitory shocks and common factors. Therefore, the SC estimator will be asymptotically unbiased if, conditional on $D(0,0) = 1$, the expected value of these linear combinations of transitory shocks and common factors are equal to zero.[12] More specifically, we need that $\mathbb{E}[\varepsilon_{0t} - \boldsymbol{\varepsilon}_t'\bar{\mathbf{w}} | D(0,0) = 1] = 0$ and $\mathbb{E}[\lambda_t(\mu_0 - \boldsymbol{\mu}'\bar{\mathbf{w}}) | D(0,0) = 1] = 0$. While the first equality is guaranteed by Assumption 2, the second one may not hold if treatment assignment is correlated with the unobserved heterogeneity.

[12] We consider the definition of asymptotic unbiasedness as the expected value of the asymptotic distribution of $\hat{\alpha}_{0t} - \alpha_{0t}$ equal to zero. An alternative definition is that $\mathbb{E}[\hat{\alpha}_{0t} - \alpha_{0t}] \to 0$. We show in Appendix A.4 that these two definitions are equivalent in this setting under standard assumptions.

Since $\mu_0 \neq \boldsymbol{\mu}'\bar{\mathbf{w}}$, the SC estimator will only be asymptotically unbiased, in general, if we impose an additional assumption that $\mathbb{E}\left[\lambda_t^k | D(0,0) = 1\right] = 0$ for all common factors k such that $\mu_0^k \neq \sum_{j \neq 0} \bar{w}_j \mu_j^k$. In order to better understand the intuition behind this result, we consider a special case in which, conditional on $D(0,0) = 1$, λ_t is stationary for $t \leq 0$. In this case, we can assume, without loss of generality, that $\omega_0^1 = \mathbb{E}[\lambda_t^1] = 1$ and $\omega_0^k = \mathbb{E}[\lambda_t^k] = 0$ for $k > 0$. Therefore, the SC estimator will only be asymptotically unbiased if the weights turn out to recover unit 0 fixed effect (that is, $\mu_0^1 = \sum_{j \neq 0} \mu_j^1$) *and* treatment assignment is uncorrelated with time-varying unobserved common factors (that is, $\mathbb{E}[\lambda_t^k | D(0,0) = 1] = 0$ for $t > 0$) such that $\mu_0^k \neq \sum_{j \neq 0} \mu_j^k$ for $k > 1$. Importantly, once we relax the assumption of a perfect pre-treatment fit, this implies that the SC estimator may be asymptotically biased even in settings in which the DID estimator is unbiased, as the DID estimator takes into account unobserved characteristics that are fixed over time, while the SC estimator would not necessarily do so. We discuss this issue in more detail in Section 4. We also discuss in Section 4 the implications of this result for the asymptotic MSE of the SC estimator.

In the derivation of equation 5, we treat $\{\mu_j\}_{j=0}^J$ as fixed. An alternative way to think about this result is that we have the asymptotic distribution of $\hat{\alpha}_{0t}$ conditional on $\{\mu_j\}_{j=0}^J$, so we derive conditions in which $\hat{\alpha}_{0t}$ is asymptotically unbiased conditional on $\{\mu_j\}_{j=0}^J$. To check whether $\hat{\alpha}_{0t}$ is asymptotically unbiased unconditionally, we would have to integrate the conditional distribution of $\hat{\alpha}_{0t}$ over the distribution of $\{\mu_j\}_{j=0}^J$. Therefore, unless we are willing to impose restrictions on the distribution of $\{\mu_j\}_{j=0}^J$, we can only guarantee that $\hat{\alpha}_{0t}$ is asymptotically unbiased unconditionally if $\hat{\alpha}_{0t}$ is asymptotically unbiased conditional on every $\{\mu_j\}_{j=0}^J$. We show that this will generally not be the case if $\mathbb{E}[\lambda_t | D(0,0) = 1] \neq 0$.

While many SC applications does not have a large number of pre-treatment periods to justify large-T_0 asymptotics (see, for example, Doudchenko and Imbens (2016)), our results can also be interpreted as the SC weights not converging to weights that reconstruct the factor loadings of the treated unit when J is fixed *even when* T_0 *is large*. In Appendix A.2, we show that, with finite T_0, the SC weights will be biased estimators for $\mathbf{w}^*$. The intuition for this result is that the SC method uses the vector of control outcomes as a proxy for the vector common factors. That is, assuming $\Phi \neq \varnothing$, we can write the potential outcome of the treated unit as a linear combination of the control units using a set of weights $\mathbf{w}^* \in \Phi$. However, in this case the control outcomes will be, by construction, correlated with the error in this model. The intuition is that the transitory shocks would behave as a measurement error in these proxy variables, which leads to bias. In Section 7, we show that, in our MC simulations, the SC weights are, on average, even further from weights that reconstruct the factor loadings of the treated unit when T_0 is finite. In Section **??** we consider the case in which both J and T_0 diverge, which provides a better approximation to a setting in which J and T_0 are of similar magnitude, but both are large.

The discrepancy of our results with the results from Abadie et al. (2010) arises because we consider different frameworks. Abadie et al. (2010) consider the properties of the SC estimator conditional on having a perfect fit in the pre-treatment period in the data at hand. They do not consider the asymptotic properties of the SC estimator when T_0 goes to infinity. Instead, they provide conditions under which the bias of the SC estimator is bounded by a term that goes to zero when T_0 increases, *if there exist a set of weights that provide a perfect pre-treatment fit.* Our results are not as conflicting with the results from Abadie et al. (2010) as they may appear at first glance. In a model with "stationary" common factors, the probability that one would actually have a dataset at hand such that the SC weights provide a close-to-perfect pre-intervention fit with a moderate T_0 is close to zero, unless the variance of the transitory shocks is small. Therefore, our results agree with the theoretical results from Abadie et al. (2010) in that the aymptotic bias of the SC estimator should be small in situations where one would expect to have a close-to-perfect fit for a large T_0.

In Appendix A.5 we consider alternative specifications used in the SC method to estimate the weights. In particular, we consider the specification that uses the pre-treatment average of the outcome variable as predictor, and the specification that uses the pre-treatment average of the outcome variable and other time-invariant covariates as predictors. In both cases, we show that the objective function used to calculate the weights converge in probability to a function that can, in general, have multiple minima. If Φ is non-empty, then $\mathbf{w} \in \Phi$ will be one solution. However, there might be $\mathbf{w} \notin \Phi$ that also minimizes this function, so there is no guarantee that the SC weights in these specifications will converge in probability to weights in Φ.

4 Comparison to DID & the demeaned SC estimator

We show in Section 3 that the SC estimator can be asymptotically biased even in situations where the DID estimator is unbiased. In contrast to the SC estimator, the DID estimator for the treatment effect in a given post-intervention period $t > 0$, under Assumption 3, would be given by[13]

$$\begin{aligned} \hat{\alpha}_{0t}^{DID} &= y_{0t} - \frac{1}{J}\mathbf{y}_t'\mathbf{i} - \frac{1}{T_0}\sum_{\tau \in \mathcal{T}_0}\left[y_{0\tau} - \frac{1}{J}\mathbf{y}_\tau'\mathbf{i}\right] \\ &\overset{p}{\rightarrow} \alpha_{0t} + \left(\varepsilon_{0t} - \frac{1}{J}\boldsymbol{\varepsilon}_t'\mathbf{i}\right) + (\lambda_t - \omega_0)\left(\mu_0 - \frac{1}{J}\boldsymbol{\mu}'\mathbf{i}\right) \text{ when } T_0 \rightarrow \infty, \end{aligned} \quad (6)$$

where $\mathbf{i}$ is a $J \times 1$ vector of ones.

Therefore, the DID estimator will be asymptotically unbiased if $\mathbb{E}[\lambda_t | D(0,0) = 1] = \omega_0$ for the factors such that $\mu_0 \neq \frac{1}{J}\boldsymbol{\mu}'\mathbf{i}$, which means that the fact that unit 0 is treated after period $t = 0$ is not informative about the first moment of the common factors relative to their pre-treatment averages. Intuitively, the unit fixed effects control for any difference in unobserved variables that remain constant (in expectation) before and after the treatment. Moreover, the DID allows for arbitrary correlation between treatment assignment and δ_t (which is captured by the time effects).

[13] Since the goal in the SC literature is to estimate the effect of the treatment for unit 1 at a specific date t, this circumvents the problem of aggregating heterogeneous effects, as considered by Callaway and Sant'Anna (2018), Athey and Imbens (2018), and Goodman-Bacon (2018) in the DID setting.

However, the DID estimator will be biased if the fact that unit 0 is treated after period $t = 0$ is informative about variations in the common factors relative to their pre-treatment mean, and it turns out that the average of the factor loadings associated to such common factors are different from the factor loadings of the treated unit.

As an alternative to the standard SC estimator, we suggest a modification in which we calculate the pre-treatment average for all units and demean the data. This is equivalent to a generalization of the SC method suggested, in parallel to our paper, by Doudchenko and Imbens (2016), which includes an intercept parameter in the minimization problem to estimate the SC weights and construct the counterfactual. Here we formally consider the implications of this alternative on the bias and MSE of the SC estimator. Relaxing the non-intercept constraint was already a feature of Hsiao et al. (2012). The difference here is that we relax this constraint while maintaining the adding-up and non-negativity constraints, which allows us to rank the demeaned SC with the DID estimator.

The demeaned SC estimator is given by $\hat{\alpha}_{0t}^{SC'} = y_{0t} - \mathbf{y}_t'\widehat{\mathbf{w}}^{\text{SC}'} - (\bar{y}_0 - \bar{\mathbf{y}}'\widehat{\mathbf{w}}^{\text{SC}'})$, where $\bar{y}_0$ is the pre-treatment average of unit 0, and $\bar{\mathbf{y}}$ is an $J \times 1$ vector with the pre-treatment averages of the controls. The weights $\widehat{\mathbf{w}}^{\text{SC}'}$ are given by

$$\widehat{\mathbf{w}}^{\text{SC}'} = \underset{\mathbf{w} \in \Delta^{J-1}}{\text{argmin}} \frac{1}{T_0} \sum_{t \in \mathcal{T}_0} \left[y_{0t} - \mathbf{y}_t'\mathbf{w} - \left(\bar{y}_0 - \bar{\mathbf{y}}'\mathbf{w}\right) \right]^2. \tag{7}$$

Proposition 2 Under Assumptions 1, 2 and 3, $\widehat{\mathbf{w}}^{\text{SC}'} \overset{p}{\to} \bar{\mathbf{w}}^{\text{SC}'}$ when $T_0 \to \infty$, where $\mu_0 \neq \boldsymbol{\mu}'\bar{\mathbf{w}}^{\text{SC}'}$, unless $\sigma_\varepsilon^2 = 0$ or $\exists \mathbf{w} \in \Phi | \mathbf{w} \in \underset{\mathbf{w} \in \Delta^{J-1}}{\text{argmin}} \{\mathbf{w}'\mathbf{w}\}$. Moreover, for $t > 0$,

$$\hat{\alpha}_{0t}^{\text{SC}'} \overset{p}{\to} \alpha_{0t} + \left(\varepsilon_{0t} - \boldsymbol{\varepsilon}_t'\bar{\mathbf{w}}^{\text{SC}'}\right) + (\lambda_t - \omega_0)\left(\mu_0 - \boldsymbol{\mu}'\bar{\mathbf{w}}^{\text{SC}'}\right) \text{ when } T_0 \to \infty. \tag{8}$$

Proof.

See details in Appendix A.1.2 ■

Therefore, both the demeaned SC and the DID estimators are asymptotically unbiased when $\mathbf{E}[\lambda_t | D(0,0) = 1] = \omega_0$ for $t > 0$. Moreover, under this assumption, both estimators are unbiased even for finite T_0. With additional assumptions on $(\varepsilon_{0t}, ..., \varepsilon_{Jt}, \lambda_t')$ in the post-treatment periods, we can also assure that the demeaned SC estimator is asymptotically more efficient than DID.

Assumption 4 (Stability in the pre- and post-treatment periods) For $t > 0$, $\mathbb{E}[\lambda_t | D(0,0) = 1] = \omega_0$, $\mathbb{E}[\epsilon_t | D(0,0) = 1] = 0$, $\mathbb{E}[\lambda_t'\lambda_t | D(0,0) = 1] = \Omega_0$, and $\mathbb{E}[\epsilon_t\epsilon_t' | D(0,0) = 1] = \sigma_\varepsilon^2 I_{J+1}$, $cov(\epsilon_t, \lambda_t | D(0,0) = 1) = 0$.

Assumptions 3 and 4 imply that transitory shocks and common factors have the same first and second moments in the pre- and post-treatment periods. From Proposition 2, Assumption 4 implies that the demeaned SC estimator is asymptotically unbiased. We now show that this assumption also implies that the demeaned SC estimator has lower asymptotic MSE than both the DID estimator and the infeasible SC estimator.

Proposition 3 Under Assumptions 1, 2, 3, and 4, the demeaned SC estimator ($\hat{\alpha}_{0t}^{\text{SC}'}$) dominates both the DID estimator ($\hat{\alpha}_{0t}^{DID}$) and the infeasible SC estimator ($\hat{\alpha}_{0t}^{*}$) in terms of asymptotic MSE when $T_0 \to \infty$.

Proof.

See details in Appendix A.1.3 ■

The intuition of this result is that, under Assumption 4, the demeaned SC weights converge to weights that minimize a function $\Gamma(\mathbf{w})$ such that $\Gamma(\bar{\mathbf{w}}^{\text{SC}'}) = a.var(\hat{\alpha}_{0t}^{\text{SC}'}|D(0,0)=1)$, $\Gamma(\mathbf{w}^*) = a.var(\hat{\alpha}_{0t}^{*}|D(0,0)=1)$, and $\Gamma(\{\frac{1}{J},...,\frac{1}{J}\}) = a.var(\hat{\alpha}_{1t}^{\text{DID}}|D(0,0)=1)$. Therefore, it must be that the asymptotic variance of $\hat{\alpha}_{0t}^{\text{SC}'}$ is weakly lower than the variance of both $\hat{\alpha}_{0t}^{*}$ and $\hat{\alpha}_{1t}^{\text{DID}}$. Moreover, these three estimators are unbiased under these assumptions.

The demeaned SC estimator dominates the infeasible one, in terms of MSE, because the infeasible SC estimator focuses on eliminating the common factors, even if this means using a linear combination of the transitory shocks with higher variance. In contrast, the demeaned SC estimator provides a better balance in terms of the variance of the common factors and transitory shocks. This dominance of the demeaned SC estimator, however, relies crucially on the assumption that the first and second moments of the common factors and transitory shocks remain stable before and after the treatment. If we had that $\mathbb{E}[\lambda_t'\lambda_t|D(0,0)=1] \neq \Omega_0$ for $t>0$, then $\Gamma(\mathbf{w})$ would not provide the variance of the estimators with weights $\mathbf{w}$. Therefore, it would not be possible to guarantee that the demeaned SC estimator has lower variance, even if the three estimators are unbiased.

If we had that $\mathbb{E}[\lambda_t|D(0,0)=1] \neq \omega_0$ for $t>0$, then both the demeaned SC and the DID estimators would be asymptotically biased, while the infeasible SC estimator would remain unbiased. The asymptotic bias of $\hat{\alpha}_{0t}^{\text{SC}'}$ would be given by $(\mathbb{E}[\lambda_t|D(0,0)=1]-\omega_0)(\mu_0 - \boldsymbol{\mu}'\bar{\mathbf{w}}^{\text{SC}'})$. Therefore, provided $\mu_0 \neq \boldsymbol{\mu}'\bar{\mathbf{w}}^{\text{SC}'}$ (which, in general, will happen), the infeasible SC estimator will dominate the demeaned SC estimator in terms of asymptotic MSE if $(\mathbb{E}[\lambda_t|D(0,0)=1]-\omega_0)$ is large enough. In other words, once we relax Assumption 4, we cannot guarantee that the demeaned SC estimator provides a better prediction in terms of MSE relative to the infeasible one. Moreover, if the bias of the demeaned SC estimator is large enough, then the infeasible SC estimator will be better in terms of MSE relative to the demeaned SC estimator.

In general, it is not possible to rank the demeaned SC and the DID estimators in terms of bias and MSE if treatment assignment is correlated with time-varying common factors. We provide in Appendix A.3 a specific example in which the DID can have a smaller bias and MSE relative to the demeaned SC estimator. This might happen when selection into treatment depends on common factors with low variance, and it happens that a simple average of the controls provides a good match for the factor loadings associated with these common factors. For the particular class of linear factor models we present in Section 7, however, the asymptotic bias and the MSE of the demeaned SC estimator will always be lower relative to the DID estimator, provided that there is stability in the variance of common factors and transitory shocks before and after the treatment.

Importantly, it is not possible to, in general, compare the original and the demeaned SC estimator in terms of bias and variance. For example, the original SC estimator may lead to lower bias if we believe it is only possible to reproduce the trend of a series if we also reproduce its level. In this case, matching also on the levels would help provide a better approximation to the factor loadings of the treated unit associated with time-varying common trends. Moreover, being able to reproduce the trend and the level is a higher bar than reproducing the trend only. Therefore, it is not clear whether demeaning is the best option in all applications. Still, this demeaning process allows us to provide conditions under which the SC method dominates the DID estimator, which would not be the case if we consider the original SC estimator.

5 Other related estimators

We show in Appendix A.5.3 that our main result that the SC estimator will be asymptotically biased if there is selection on time-varying unobservables still apply if we also relax the non-negative and the adding-up constraints, which essentially leads to the panel data approach suggested by Hsiao et al. (2012), and further explored by Li and Bell (2017).[14] Our conditions for unbiasedness of the SC estimator also apply to the estimators proposed by Carvalho et al. (2018) and Carvalho et al. (2016) when J is fixed.

These papers rely on assumptions that essentially imply no selection on unobservables to derive consistency results, which reconciles our results with theirs. Hsiao et al. (2012) and Li and Bell (2017) implicitly rely on stability in the linear projection of the potential outcomes of the treated unit on the outcomes of the control units, before and after the intervention, to show that their proposed estimators are unbiasedness and consistency. See, for example, equation A.4 from Li and Bell (2017). The linear projection of y_{0t}^N on $\mathbf{y}_t$ for any given t is given by $\delta_1(t) + \mathbf{y}_t'\delta(t)$, where

$$\begin{cases} \delta(t) = \left[\boldsymbol{\mu} var(\lambda_t | D(0,0)=1)\boldsymbol{\mu}'\right]^{-1} \boldsymbol{\mu} var(\lambda_t | D(0,0)=1)\mu_0, \text{ and} \\ \delta_1(t) = \mathbb{E}[\lambda_t | D(0,0)=1](\mu_0 - \boldsymbol{\mu}'\delta(t)). \end{cases} \tag{9}$$

Therefore, in general, we will only have $(\delta_1(t), \delta(t))$ constant for all t if the distribution of λ_t conditional on $D(0,0) = 1$ is stable over time. However, the idea that treatment assignment is correlated with the factor model structure essentially means that the distribution of λ_t conditional on $D(0,0) = 1$ is different before and after the treatment assignment. Therefore, it would not be reasonable to assume that the parameters of the linear projection of y_{0t}^N on $\mathbf{y}_t$ are the same for $t \in \mathcal{T}_0$ and $t \in \mathcal{T}_1$ if we consider that treatment assignment is correlated with the factor model structure. Chernozhukov et al. (2018) assume that y_{0t}^N on $\mathbf{y}_t$ are covariance-stationary (see their Assumption 6), which implies that $(\delta_1(t), \delta(t))$ constant for all t. Therefore, they also implicitly imply that there is no selection on unobservables. Since they consider a setting with both large J and T, however, it is possible that their estimator is consistent when there is selection on unobservables

[14] In this case, since we do not constraint the weights to sum 1, we need to adjust Assumption 3 so that it also includes convergence of the pre-treatment averages of the first and second moments of δ_t.

under conditions similar to the ones considered by Ferman (2019).

Carvalho et al. (2018), Carvalho et al. (2016), and Masini and Medeiros (2016) assume that the outcome of the control units are independent from treatment assignment. If we consider the linear factor model structure from equation 1, then this essentially means that there is no selection on unobservables. Given Assumption 2, if treatment assignment is correlated with the potential outcomes of the treated unit, then it must be correlated with $\lambda_t \mu_0$. However, if this is the case, then treatment assignment must also be correlated with at least some control units given the common shocks λ_t, implying that their assumption that the outcome of the control units are independent from treatment assignment would be violated.

Overall, our results clarify what selection on unobservables means in this setting, and the conditions under which these estimators are asymptotically unbiased when J is fixed.

6 Model with "explosive" common factors

Many SC applications present time-series patterns that are not consistent with Assumption 3, including the applications considered by Abadie and Gardeazabal (2003), Abadie et al. (2010), and Abadie et al. (2015). This will be the case whenever we consider outcome variables that exhibit non-stationarities, such as GDP and average wages. We consider now the case in which the first and second moments of a subset of the common factors diverge. We modify the model to

$$\begin{cases} y_{jt}^N = \lambda_t \mu_j + \gamma_t \theta_j + \varepsilon_{jt} \\ y_{jt}^I = \alpha_{jt} + y_{jt}^N \end{cases} \tag{10}$$

where $\lambda_t = (\lambda_t^1, ..., \lambda_t^{F_0})$ is a $(1 \times F_0)$ vector of $I(0)$ common factors, and $\gamma_t = (\gamma_t^1, ..., \gamma_t^{F_1})$ is a $(1 \times F_1)$ vector of common factors that are $\mathcal{I}(1)$ and/or polynomial time trends t^f, while μ_j and θ_j are the vectors of factor loadings associated with these common factors. The time effect δ_t can be either included in vector λ_t or γ_t. Differently from the previous sections, in order to consider the possibility that treatment starts after a large number of periods in which some common factors may be $\mathcal{I}(1)$ and/or polynomial time trends, we label the periods as $t = 1, ..., T_0, T_0 + 1, ..., T$. We modify Assumption 3 to determine the behavior of the common factors and the transitory shocks in the pre-treatment periods.

Assumption 3′ (stochastic processes) Conditional on $D(0, T_0) = 1$, the process $z_t = (\varepsilon_{0t}, ..., \varepsilon_{Jt}, \lambda_t)$ is $I(0)$ and weakly stationary with finite fourth moments, while the components of γ_t are $I(1)$ and/or polynomial time trends t^f for $t = 1, ..., T_0$.

Assumption 3′ restricts the behavior of the common factors in the pre-treatment periods. However, this assumption allows for correlation between treatment assignment and common factors in the post-intervention periods. For example, if $\gamma_t^k = \gamma_{t-1}^k + \eta_t$, then Assumption 3′ implies that η_t has mean zero for all $t \leq T_0$. However, it may be that $\mathbb{E}[\eta_t | D(0, T_0)] \neq 0$ for $t > T_0$. This assumption could be easily relaxed to allow for $\mathbb{E}[\eta_t|D(0,T_0)] \neq 0$ for a fixed number of periods prior to the start of the treatment.

We also consider an additional assumption on the existence of weights that reconstruct the factor loadings of unit 1 associated with the non-stationary common trends.

Assumption 5 (existence of weights)

$$\exists \ \mathbf{w}^* \in W \mid \theta_1 = \sum_{j \neq 1} w_j^* \theta_j$$

where W is the set of possible weights given the constrains on the weights the researcher is willing to consider. For example, Abadie et al. (2010) suggest $W = \{\mathbf{w} \in \mathbb{R}^J \mid \sum_{j \neq 1} w_j^* = 1, \text{ and } w_j^* \geq 0\}$, while Hsiao et al. (2012) allows for $W = \mathbb{R}^J$. Let Φ_1 be the set of weights in W that reconstruct the factor loadings of unit 1 associated with the $I(1)$ common factors. Assumption 5 implies that $\Phi_1 \neq \varnothing$. In a setting in which γ_t is a vector of $I(1)$ common factors, Assumption 5 implies that the vector of outcomes $\mathbf{y}_t = (y_{0t}, ..., y_{J+1,t})'$ is co-integrated. However, we do *not* need to assume existence of weights in Φ_1 that also reconstruct the factor loadings of unit 1 associated with the $I(0)$ common factors, so it may be that $\Phi = \varnothing$, where Φ is the set of weights that reconstruct *all* factor loadings.

We consider an asymptotic exercise where $T_0 \to \infty$ with "explosive" common factors, so it is not possible to fix the label of the post-treatment periods, as we do in Sections 3 and 4. Instead, we consider the asymptotic distribution of the estimator for the treatment effect τ periods after the start of the treatment.

Proposition 4 Under Assumptions 1, 2, 3′, and 5, for $t = T_0 + \tau$, $\tau > 0$,

$$\hat{\alpha}_{0t}^{\mathrm{SC}'} \overset{d}{\to} \alpha_{0t} + \left(\varepsilon_{0t} - \sum_{j \neq 1} \bar{w}_j \varepsilon_{jt} \right) + (\lambda_t - \omega_0) \left(\mu_0 - \sum_{j \neq 1} \bar{w}_j \mu_j \right) \text{ when } T_0 \to \infty \tag{11}$$

where $\mu_0 \neq \sum_{j \neq 1} \bar{w}_j \mu_j$, unless $\sigma_\varepsilon^2 = 0$ or $\exists \mathbf{w} \in \Phi | \mathbf{w} \in \underset{\mathbf{w} \in W}{\text{argmin}} \left\{ \sum_{j \neq 1} (w_j)^2 \right\}$.

Proof.

Details in Appendix A.1.4. ■

Proposition 4 has two important implications. First, if Assumption 5 is valid, then the asymptotic distribution of the demeaned SC estimator does not depend on the non-stationary common trends. The intuition of this result is the following. The demeaned SC weights will converge to weights that reconstruct the factor loadings of the treated unit associated with the non-stationary common trends. Interestingly, while $\widehat{\mathbf{w}}$ will generally be only $\sqrt{T_0}$−consistent when Φ_1 is not a singleton, we show in Appendix A.1.4 that there are linear combinations of $\widehat{\mathbf{w}}$ that will converge at a faster rate, implying that $\gamma_t(\theta_1 - \sum_{j \neq 1} \hat{w}_j \theta_j) \overset{p}{\to} 0$, despite the fact that γ_t explodes when $T_0 \to \infty$. Therefore, such non-stationary common trends will not lead to asymptotic bias in the SC estimator. Second, the demeaned SC estimator will be biased if there is correlation between treatment assignment and the $I(0)$ common factors. The intuition is that the demeaned SC weights will converge in probability to weights in Φ_1 that minimize the variance of the $I(0)$ process $u_t = y_{0t} - \sum_{j \neq 1} w_j y_{jt} = \lambda_t(\mu_0 - \sum_{j \neq 1} w_j \mu_j) + (\varepsilon_{0t} - \sum_{j \neq 1} w_j \varepsilon_{jt})$. Following the same arguments as in Proposition 1, $\widehat{\mathbf{w}}$ will not eliminate the $I(0)$ common factors, unless we have that $\sigma_\varepsilon^2 = 0$ or it coincides that there is a $\mathbf{w} \in \Phi$ that also minimizes the linear combination of transitory shocks.

The result that the asymptotic distribution of the SC estimator does not depend on the non-stationary common trends depends crucially on Assumption 5. If there were no linear combination of the control units that reconstruct the factor loadings of the treated unit associated to the non-stationary common trends, then the asymptotic distribution of the SC estimator would trivially depend on these common trends, which might lead to bias in the SC estimator if treatment assignment is correlated with such non-stationary trends.

Proposition 4 remains valid when we relax the adding-up and/or the non-negativity constraints, with minor variations in the conditions for unbiasedness.[15] However, these results are not valid when we consider the no-intercept constraint, as the original SC estimator does. When the intercept is not included, it remains true that $\widehat{\mathbf{w}} \xrightarrow{p} \bar{\mathbf{w}} \in \Phi_1$. However, in this case, the weights will not converge fast enough to compensate the fact that γ_t explodes. We present an example in Appendix A.6.2.

The results from Proposition 4 suggest that correlation between treatment assignment and stationary common factors, beyond such non-stationary trends, may lead to bias in the SC estimator. Therefore, we recommend that researchers should *also* present the pre-treatment fit after eliminating non-stationary trends as an additional diagnosis test for the SC estimator, as this should be more indicative of potential bias from possible correlation between treatment assignment and stationary common factors. To illustrate this point, we consider the application presented by Abadie and Gardeazabal (2003).

We present in Figure 1.A the per capita GDP time series for the Basque Country and for other Spanish regions, while in Figure 1.B we replicate Figure 1 from Abadie and Gardeazabal (2003), which displays the per capita GDP of the Basque Country contrasted with the per capita GDP of a synthetic control unit constructed to provide a counterfactual for the Basque Country without terrorism. Figure 1.B displays a remarkably good pre-treatment fit. However, the per capita GDP series is clearly non-stationary, with all regions displaying similar trends before the intervention. Therefore, in light of Proposition 4, it may still be that correlation between treatment assignment and common factors beyond this non-stationary trend may lead to bias. In order to assess this possibility, we de-trend the data, so that we can have a better assessment on whether factor loadings associated with stationary common factors are also well matched. We subtract the outcome of the

[15] Relaxing the adding-up constraint makes the estimator biased if δ_t is correlated with treatment assignment and if it is $I(0)$. If δ_t is $I(1)$, then the weights will converge to sum one even when such restriction is not imposed, so this would not generate bias. Including or not the non-negative constraint does not alter the conditions for unbiasedness, although it may be that Assumption 5 is valid in a model without the non-negativity constraints, but not valid in a model with these constraints.

treated and control units by the average of the control units at time t ($a_t = \frac{1}{J}\sum_{j\neq 1} y_{jt}$).[16] If the non-stationarity comes from a common factor δ_t that affects every unit in the same way, then the series $\tilde{y}_{jt} = y_{jt} - \frac{1}{J}\sum_{j'\neq 1} y_{j't}$ would not display non-stationary trends. As shown in Figure 1.C, in this case, the treated and SC units do not display a non-stationary trend. The pre-treatment fit is still good for this de-trended series, but not as good as in the previous case, providing a better assessment of possible mismatches in factor loadings associated with stationary trends. In the presence of non-stationary common factors, a possible bias due to a correlation between treatment assignment and stationary common factors should become small relative to the scale of the outcome variable when $T_0 \rightarrow \infty$. However, this empirical illustration suggest that, for a finite T_0, a mismatch in factor loadings associated with stationary common factors might still be relevant, even when non-stationary common factors lead to graphs with seemingly perfect pre-treatment fit when we consider the variables in level.

Importantly, our results do not imply that one should not use the SC method when the data is non-stationary. On the contrary, we show that the SC method is very efficient in dealing with non-stationary trends. Indeed, the seemingly perfect pre-treatment fit when we consider the outcomes in level suggest that the method is being highly successful in taking into account non-stationary trends, which is an important advantage of the method relative to alternatives such as DID. Our only suggestion is to *also* present graphs with the de-trended series to have a better assessment of possible imbalances in the factor loadings associated with stationary common trends, beyond those non-stationary trends. Another possibility would be to apply the SC method on a transformation of the data that makes it stationary. In this case, however, the estimator would not be numerically the same as the estimator using the original data.

7 Particular Class of Linear Factor Models & Monte Carlo Simulations

We consider now in detail a particular class of linear factor models in which all units are divided into groups that follow different times trends. We present both theoretical and MC simulations for these models. In Section 7.1 we consider the case with stationary common factors, while in Section 7.2 we consider a case in which there are both $I(1)$ and $I(0)$ common factors.

7.1 Model with stationary common factors

We consider first a model in which the $J+1$ units are divided into K groups, where for each j we have that

$$y_{jt}(0) = \delta_t + \lambda_t^k + \varepsilon_{jt} \tag{12}$$

[16] Note that, under the adding-up constraint ($\sum_{j\neq 1} w_j = 1$), the SC weights with this de-trended data will be numerically the same as the original SC weights.

for some $k = 1, ..., K$. As in Section 3, let $t = -T_0+1, ..., 0, 1, ..., T_1$. We assume that $\frac{1}{T_0}\sum_{t=-T_0+1}^{0}\lambda_t^k \xrightarrow{p} 0$, $\frac{1}{T_0}\sum_{t=-T_0+1}^{0}(\lambda_t^k)^2 \xrightarrow{p} 1$, $\frac{1}{T_0}\sum_{t=-T_0+1}^{0}\varepsilon_{jt} \xrightarrow{p} 0$, $\frac{1}{T_0}\sum_{t=-T_0+1}^{0}\varepsilon_{jt}^2 \xrightarrow{p} \sigma_\varepsilon^2$ and $\frac{1}{T_0}\sum_{t=-T_0+1}^{0}\lambda_t^k\varepsilon_{jt} \xrightarrow{p} 0$
.

7.1.1 Asymptotic Results

Consider first an extreme case in which $K = 2$, so the first half of the $J + 1$ units follows the parallel trend given by λ_t^1, while the other half follows the parallel trend given by λ_t^2. In this case, an infeasible SC estimator would only assign positive weights to units in the first group.

We calculate, for this particular class of linear factor models, the asymptotic proportion of misallocated weights of the SC estimator using all pre-treatment lags as predictors. From the minimization problem 4, we have that, when $T_0 \to \infty$, the proportion of misallocated weights converges to

$$\gamma_2(\sigma_\varepsilon^2, J) = \sum_{j=\frac{J+1}{2}+1}^{J+1}\bar{w}_j = \frac{J+1}{J^2 + 2 \times J \times \sigma_\varepsilon^2 - 1} \times \sigma_\varepsilon^2 \tag{13}$$

where $\gamma_K(\sigma_\varepsilon^2, J)$ is the proportion of misallocated weights when the $J + 1$ groups are divided in K groups.

We present in Figure 2.A the relationship between asymptotic misallocation of weights, variance of the transitory shocks, and number of control units. For a fixed J, the proportion of misallocated weights converges to zero when $\sigma_\varepsilon^2 \to 0$, while this proportion converges to $\frac{J+1}{2J}$ (the proportion of misallocated weights of DID) when $\sigma_\varepsilon^2 \to \infty$. This is consistent with the results we have in Section 3. Moreover, for a given σ_ε^2, the proportion of misallocated weights converges to zero when the number of control units goes to infinity. This is consistent with Gobillon and Magnac (2016), who derive support conditions so that the assumptions from Abadie et al. (2010) for unbiasedness are satisfied when both T_0 and J go to infinity.

In this example, the SC estimator, for $t > 0$, converges to

$$\hat{\alpha}_{1t} \xrightarrow{d} \alpha_{1t} + \left(\varepsilon_{1t} - \sum_{j\neq 1}\bar{w}_j\varepsilon_{jt}\right) + \lambda_t^1 \times \gamma_2(\sigma_\varepsilon^2, J) - \lambda_t^2 \times \gamma_2(\sigma_\varepsilon^2, J), \tag{14}$$

so the potential bias due to correlation between treatment assignment and common factors (for example, $\mathbb{E}[\lambda_t^1|D(1,0) = 1] \neq 0$ for $t > 0$) will directly depend on the proportion of misallocated weights.

We consider now another extreme case in which the $J + 1$ units are divided into $K = \frac{J+1}{2}$ groups that follow the same parallel trend. In this case, each unit has a pair that follows its same parallel trend, while all other units follow different parallel trends. The proportion of misallocated weights converges to

$$\gamma_{\frac{J+1}{2}}(\sigma_\varepsilon^2, J) = \sum_{j=3}^{J+1} \bar{w}_j = \frac{J-1}{J(1+\sigma_\varepsilon^2)+1} \times \sigma_\varepsilon^2. \tag{15}$$

We present the relationship between misallocation of weights, variance of the transitory shocks, and number of control units in Figure 2.B. Again, the proportion of misallocated weights converges to zero when $\sigma_\varepsilon^2 \to 0$ and to the proportion of misallocated weights of DID when $\sigma_\varepsilon^2 \to \infty$ (in this case, $\frac{J-1}{J}$). Differently from the previous case, however, for a given σ_ε^2, the proportion of misallocated weights converges to $\frac{\sigma_\varepsilon^2}{1+\sigma_\varepsilon^2}$ when $J \to \infty$. Therefore, the SC estimator would remain asymptotically biased even when the number of control units is large. This happens because, in this model, the number of common factors increases with J, so the conditions derived by Gobillon and Magnac (2016) are not satisfied.

In both cases, the proportion of misallocated weights is always lower than the proportion of misallocated weights of DID. Therefore, in this particular class of linear factor models, the asymptotic bias of the SC estimator will always be lower than the asymptotic bias of DID. If we further assume that the variance of common factors and transitory shocks remain constant in the pre- and post-intervention periods, then we also have that the SC estimator will have lower variance and, therefore, lower MSE relative to the DID estimator. However, this is not a general result, as we show in Appendix A.3.

Finally, we compare the asymptotic MSE between the feasible and the infeasible SC estimator in this particular class of linear factor models. As outlined in Section 4, assuming that common factors and transitory shocks are stable before and after the intervention, the feasible SC estimator has a lower asymptotic MSE relative to the infeasible one. However, if the feasible SC estimator is asymptotically biased, and the bias is large enough, then it will have a higher asymptotic MSE relative to the infeasible SC estimator. We illustrate these features in Table 1. Considering 20 units divided in 10 groups of 2 (columns 1 to 3), the feasible SC estimator has a lower asymptotic MSE for the estimator of α_{1t}, for $t > 0$, when $\mathbb{E}[\lambda_t|D(1,0)=1]=1$. However, when the correlation between treatment assignment and common factors is larger, then the feasible SC estimator has a higher asymptotic MSE relative to the infeasible one. When the number of post-treatment periods is greater than one (that is, $T_1 > 1$), if we consider estimators for the average treatment effect across all post-treatment periods, then the ratio of asymptotic MSE for the feasible and infeasible SC estimators would be substantially higher. In this case, the infeasible SC estimator dominates the feasible one in terms of asymptotic MSE even when $\mathbb{E}[\lambda_t|D(1,0)=1]=1$ (panel ii of Table 1). This happens because averaging across post-treatment periods does not affect the asymptotic bias, while it reduces the variance of both estimators. In columns 4 to 6, we present the case in which 20 units are divided in 2 groups of 10. In this case, the difference between the two estimators is much smaller, although it also shows that the feasible SC estimator has a higher asymptotic MSE when its bias is large enough. While, of course, the infeasible SC estimator would not be available in real applications, these results highlight that researchers applying the SC estimator should be aware that it may have a non-trivial asymptotic MSE if there is correlation between treatment assignment and unobserved common factors.

7.1.2 Monte Carlo Simulations

The results presented in Section 7.1.1 are based on large-T_0 asymptotics. We now consider, in MC simulations, the finite T_0 properties of the SC estimator. We present MC simulations using a data generating process (DGP) based on equation 12, with $K = 10$ (that is, 10 groups of 2). We consider in our MC simulations $J + 1 = 20$, λ_t^k normally distributed following an AR(1) process with 0.5 serial correlation parameter, $\varepsilon_{jt} \sim N(0, \sigma_\varepsilon^2)$, and $T - T_0 = 10$. We also impose that there is no treatment effect, i.e., $y_{jt} = y_{jt}(0) = y_{jt}(1)$ for each time period $t \in \{-T_0 + 1, ..., 0, 1, ..., T_1\}$. We consider variations in DGP in the following dimensions:

- The number of pre-intervention periods: $T_0 \in \{5, 20, 50, 100\}$.
- The variance of the transitory shocks: $\sigma_\varepsilon^2 \in \{0.1, 0.5, 1\}$.

For each simulation, we calculate the SC estimator that uses all pre-treatment outcome lags as predictors, and calculate the proportion of misallocated weights. For each scenario, we generate 20,000 simulations.

In columns 1 to 3 of Table 2, we present the proportion of misallocated weights when $K = 10$ for different values of T_0 and σ_ε^2. Consistent with our analytical results from Section 7.1.1, the misallocation of weights is increasing with the variance of the transitory shocks. The misallocation of weights goes to zero when $\sigma_\varepsilon^2 \to 0$, which is the case in which we should expect to find applications with a good pre-treatment fit. With $T_0 = 100$, the proportion of misallocated weights is close to the asymptotic values, while the proportion of misallocated weights is substantially higher when T_0 is small. From equation 14, there is a direct link between misallocation of weights and the bias of the SC estimator (for a given $\mathbb{E}[\lambda_t | D(1,0) = 1]$). Therefore, if there is correlation between treatment assignment and common factors, then the bias of the SC estimator should be expected to be larger than its asymptotic values when T_0 is small.

In this particular class of linear factor models, the proportion of misallocated weights is always lower than the proportion of misallocated weights of the DID estimator, which implies in a lower bias if treatment assignment is correlated with common factors. This is true even when the pre-treatment match is not perfect and when the number of pre-treatment periods is very small. From Section 4, we also know that, if common factors are stationary for both pre- and post-treatment periods, then a demeaned SC estimator is unbiased and has a lower asymptotic variance than DID. Since this DGP has no time-invariant factor, this is true for the standard SC estimator as well. Columns 4 to 6 of Table 2 present the DID/SC ratio of standard errors. With $T_0 = 100$, the DID standard error is 2.4 times higher than the SC standard errors when $\sigma_\varepsilon^2 = 0.1$. When σ_ε^2 is higher, the advantage of the SC estimator is reduced, although the DID standard error is still 1.3 (1.1) times higher when σ_ε^2 is equal to 0.5 (1). This is expected given that, in this model, the

SC estimator converges to the DID estimator when $\sigma_\varepsilon^2 \to \infty$. More strikingly, the variance of the SC estimator is lower than the variance of DID even when the number of pre-treatment periods is small. This suggests that the SC estimator can still improve relative to DID even when the number of pre-treatment periods is not large and when the pre-treatment fit is not perfect, situations in which Abadie et al. (2015) suggest the method should not be used. However, a very important qualification is that, in these cases, the SC estimator requires stronger identification assumptions than stated in the original SC papers. More specifically, it is generally asymptotically biased if treatment assignment is correlated with time-varying confounders.

7.2 Model with "explosive" common factors

We consider now a model in which a subset of the common factors is $I(1)$. We consider the following DGP:

$$y_{jt}(0) = \delta_t + \lambda_t^k + \gamma_t^r + \varepsilon_{jt} \tag{16}$$

for some $k = 1, ..., K$ and $r = 1, ..., R$. We maintain that λ_t^k is stationary, while γ_t^r follows a random walk.

7.2.1 Asymptotic results

Based on our results from Section 6, the SC weights will converge to weights in Φ_1 that minimize the second moment of the $I(0)$ process that remains after we eliminate the $I(1)$ common factor. Consider the case $K = 10$ and $R = 2$. Therefore, units $j = 2, ..., 10$ follow the same non-stationary path γ_t^1 as the treated unit, although only unit $j = 2$ also follows the same stationary path λ_t^1 as the treated unit. In this case, asymptotically, all weights would be allocated among units 2 to 10, eliminating the relevance of the $I(1)$ common factor. However, the allocation of weights within these units will not assign all weights to unit 2, so the $I(0)$ common factor will remain relevant.

7.2.2 Monte Carlo simulations

In our MC simulations, we maintain that λ_t^k is normally distributed following an AR(1) process with 0.5 serial correlation parameter, while γ_t^r follows a random walk. We consider the case $K = 10$ and $R = 2$.

The proportion of misallocated weights (in this case, weights not allocated to unit 2) is very similar to the proportion of misallocated weights in the stationary case (columns 1 to 3 of Table 3). If we consider the misallocation of weights only for the $I(1)$ factors, then the misallocation of weights is remarkably low with moderate T_0, even when the variance of the transitory shocks is high (columns 4 to 6 of Table 3). The reason is that, with a moderate T_0, the $I(1)$ common factors dominate the transitory shocks, so the SC method is extremely efficient selecting control units that follow the same non-stationary trend as the treated unit.

This suggests that the SC method works remarkably well to control for $I(1)$ common factors. However, we might still have misallocation of weights for the $I(0)$ common factors. Taken together, these results suggest that the SC method provides substantial improvement relative to DID in this scenario, as the SC estimator is extremely efficient in capturing the $I(1)$ factors. Also, if the DID and SC estimators are unbiased, then the variance of the DID relative to the variance of the SC estimator would be substantially higher, as presented in columns 7 to 9 Table 3. However, one should be aware that, in this case, the identification assumption only allows for correlation of treatment assignment with the $I(1)$ factors. Still, this potential bias of the SC estimator due to a correlation between treatment assignment and the $I(0)$ common shocks, in this particular class of linear factor models, would be lower than the bias of DID.

8 Conclusion

We consider the properties of the SC and related estimators, in a linear factor model setting, when the pre-treatment fit is imperfect. We show that, in this framework, the SC estimator is generally biased if treatment assignment is correlated with the unobserved heterogeneity, and that such bias does not converge to zero even when the number of pre-treatment periods is large. Still, we also show that a modified version of the SC method can substantially improve relative to currently available methods, even if the pre-treatment fit is not close to perfect and if T_0 is not large. Moreover, we suggest that, in addition to the standard graph comparing treated and SC units, researchers should also present a graph comparing the treated and SC units after de-trending the data, so that it is possible to better assess whether there might be relevant possibilities for bias arising due to a correlation between treatment assignment and common factors beyond non-stationary trends. Overall, we show that the SC method can provide substantial improvement relative to alternative methods, even in settings where the method was not originally designed to work. However, researchers should be more careful in the evaluation of the identification assumptions in those cases.

References

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller, "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program," *Journal of the American Statiscal Association*, 2010, *105* (490), 493–505.

_ , _ , and _ , "Comparative Politics and the Synthetic Control Method," *American Journal of Political Science*, 2015, *59* (2), 495–510.

_ and Javier Gardeazabal, "The Economic Costs of Conflict: A Case Study of the Basque Country," *American Economic Review*, 2003, *93* (1), 113–132.

Amjad, Muhammad Jehangir, Devavrat Shah, and Dennis Shen, "Robust Synthetic Control," 2017.

Ando, Michihito and Fredrik Sävje, "Hypothesis Testing with the Synthetic Control Method," 2013. Working Paper.

Arkhangelsky, Dmitry, Susan Athey, David A. Hirshberg, Guido W. Imbens, and Stefan Wager, "Synthetic Difference in Differences," *arXiv e-prints*, Dec 2018, p. arXiv:1812.09970.

Athey, Susan and Guido Imbens, "Design-based Analysis in Difference-In-Differences Settings with Staggered Adoption," Working Paper, arXiv:1808.05293 2018.

_ and Guido W. Imbens, "The State of Applied Econometrics: Causality and Policy Evaluation," *Journal of Economic Perspectives*, May 2017, *31* (2), 3–32.

_ , Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi, "Matrix Completion Methods for Causal Panel Data Models," 2017.

Bai, Jushan, "Panel Data Models With Interactive Fixed Effects," *Econometrica*, 2009, *77* (4), 1229–1279.

Ben-Michael, Eli, Avi Feller, and Jesse Rothstein, "The Augmented Synthetic Control Method," Working Paper, arXiv:1811.04170 2018.

Botosaru, Irene and Bruno Ferman, "On the Role of Covariates in the Synthetic Control Method," *Econometrics Journal*, 1 2019.

Callaway, Brantly and Pedro H. C. Sant'Anna, "Difference-in-Differences with Multiple Time Periods and an Application on the Minimum Wage and Employment," Working Paper, arXiv:1803.09015 2018.

Carvalho, Carlos V., Ricardo Masini, and Marcelo C. Medeiros, "The Perils of Counterfactual Analysis with Integrated Processes," December 2016. Working Paper.

_ , _ , and _ , "ArCo: An Artificial Counterfactual Approach for Aggregate Data," *Journal of Econometrics*, 2018. Forthcoming.

Chernozhukov, Victor, Han Hong, and Elie Tamer, "Estimation and Confidence Regions for Parameter Sets in Econometric Models," *Econometrica*, 2007, *75* (5), 1243–1284.

Chernozhukov, Victor, Kaspar Wuthrich, and Yinchu Zhu, "Practical and robust t-test based inference for synthetic control and related methods," *arXiv e-prints*, Dec 2018, p. arXiv:1812.10820.

Doudchenko, Nikolay and Guido Imbens, "Balancing, regression, difference-in-differences and synthetic control methods: A synthesis," 2016.

Ferman, Bruno, "On the Properties of the Synthetic Control Estimator with Many Periods and Many Controls," *arXiv e-prints*, Jun 2019, p. arXiv:1906.06665.

Ferman, Bruno, Cristine Pinto, and Vitor Possebom, "Cherry Picking with Synthetic Controls," MPRA Paper 78213, University Library of Munich, Germany August 2017.

Gobillon, Laurent and Thierry Magnac, "Regional Policy Evaluation: Interactive Fixed Effects and Synthetic Controls," *The Review of Economics and Statistics*, 2016, *98* (3), 535–551.

Goodman-Bacon, Andrew, "Difference-in-Differences with Variation in Treatment Timing," Working Paper 25018, National Bureau of Economic Research September 2018.

Hamilton, James Douglas, *Time series analysis*, Princeton, NJ: Princeton Univ. Press, 1994.

Heckman, James and Jose Scheinkman, "The Importance of Bundling in a Gorman-Lancaster Model of Earnings," *The Review of Economic Studies*, 1987, *54* (2), 243–255.

Hsiao, Cheng, H. Steve Ching, and Shui Ki Wan, "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*, 2012, *27* (5), 705–740.

Kaul, Ashok, Stefan Klöbner, Gregor Pfeifer, and Manuel Schieler, "Synthetic Control Methods: Never Use All Pre-Intervention Outcomes as Economic Predictors," May 2015. Working Paper.

Li, Kathleen T. and David R. Bell, "Estimation of average treatment effects with panel data: Asymptotic theory and implementation," *Journal of Econometrics*, 2017, *197* (1), 65 – 75.

Masini, Ricardo and Marcelo C. Medeiros, "Counterfactual Analysis with Artificial Controls: Inference, High Dimensions and Nonstationarity," 2016. Working Paper.

Moon, Hyungsik Roger and Martin Weidner, "Linear Regression for Panel With Unknown Number of Factors as Interactive Fixed Effects," *Econometrica*, 2015, *83* (4), 1543–1579.

Newey, Whitney K., "Uniform Convergence in Probability and Stochastic Equicontinuity," *Econometrica*, 1991, *59* (4), 1161–1167.

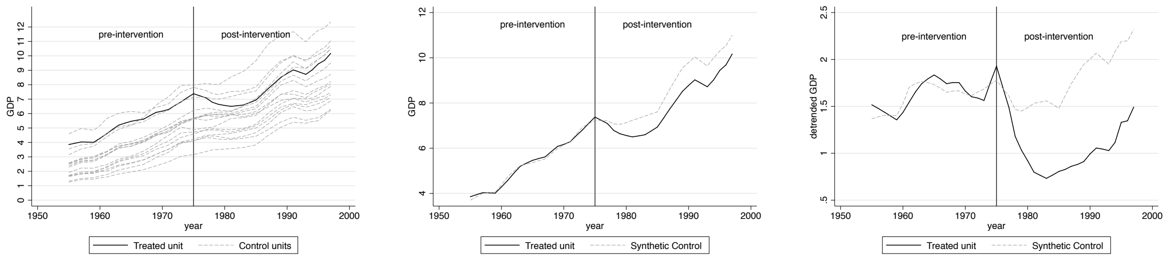
_ **and Daniel McFadden**, "Chapter 36 Large sample estimation and hypothesis testing," in "in," Vol. 4 of *Handbook of Econometrics*, Elsevier, 1994, pp. 2111 – 2245.

Powell, David, "Imperfect Synthetic Controls: Did the Massachusetts Health Care Reform Save Lives?," 2017.

Xu, Yiqing, "Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models," *Political Analysis*, 2017, *25* (1), 57?76.

Figure 1: **Abadie and Gardeazabal (2003) application**

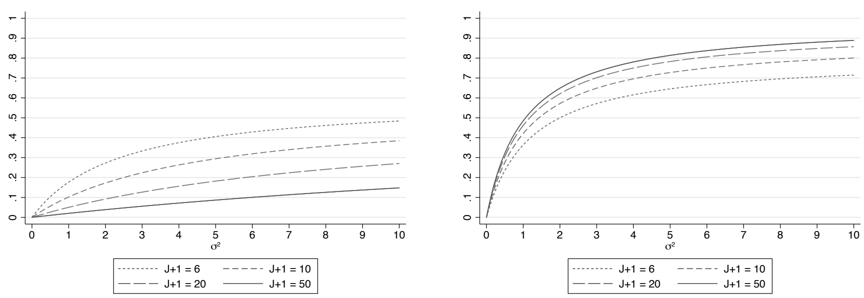
1.A: Raw data | 1.B: GDP in level | 1.C: GDP de-trended



Notes: Figure A presents time series for the treated and for the control units used in the empirical application from Abadie and Gardeazabal (2003). In Figure B we present the time series for the treated and for the SC units. In Figure C we present the same information as in Figure B after subtracting the control groups' averages for each time period.

Figure 2: **Asymptotic Misallocation of Weights**

2.A: $K = 2$ | 2.B: $K = \frac{J+1}{2}$



Notes: these figures present the asymptotic misallocation of weights of the SC estimator as a function of the variance of the transitory shocks for different numbers of control units. Figures 2.A presents results when there are 2 groups of $\frac{J+1}{2}$ units each, while Figure 2.B presents results when there are $\frac{J+1}{2}$ groups of 2 units each. The misallocation of weights is defined as the proportion of weight allocated to units that do not belong to the group of treated unit.

Table 1: **Asymptotic MSE (Feasible SC estimator / Infeasible SC estimator)**

	$K = \frac{J+1}{2} = 10$			$K = 2$		
	$\sigma^2_\varepsilon = 0.1$	$\sigma^2_\varepsilon = 0.5$	$\sigma^2_\varepsilon = 1$	$\sigma^2_\varepsilon = 0.1$	$\sigma^2_\varepsilon = 0.5$	$\sigma^2_\varepsilon = 1$
$\mathbb{E}[\lambda_t \vert D(0,0) = 1]$	(1)	(2)	(3)	(4)	(5)	(6)
			Panel i: $T_1 = 1$			
1	0.99	0.94	0.88	1.00	1.00	1.00
2	1.09	1.22	1.20	1.00	1.00	1.00
4	1.50	2.34	2.47	1.00	1.02	1.03
			Panel ii: $T_1 = 10$			
1	1.07	1.14	1.11	1.00	1.00	1.00
2	1.39	2.02	2.12	1.00	1.01	1.02
4	2.67	5.56	6.16	1.01	1.06	1.11

Notes: this table presents the ration of the asymptotic MSE of the feasible and infeasible SC estimator for the model presented in Section 7.1. We set $J+1 = 20$. Columns 1 to 3 present the case in which these 20 units are divided in 10 groups of 2 units each, while columns 4 to 6 present the case in which units are divided in 2 groups of 10. Different columns present different values of σ^2_ϵ, while $\sigma^2_\lambda = 1$. Different rows present different values of $\mathbb{E}[\lambda_t|D(0,0)]$ for $t > 0$ (that is, in the post-treatment periods. Panel i displays the results when there is only one post-treatment periods. Panel ii assumes 10 post-treatment periods, considering an estimator for the average treatment effect across all post-treatment periods.

Table 2: **Monte Carlo Simulations - Stationary Model**

	Misallocation of weights			DID/SC ratio of standard errors		
	$\sigma^2_\varepsilon = 0.1$	$\sigma^2_\varepsilon = 0.5$	$\sigma^2_\varepsilon = 1$	$\sigma^2_\varepsilon = 0.1$	$\sigma^2_\varepsilon = 0.5$	$\sigma^2_\varepsilon = 1$
	(1)	(2)	(3)	(4)	(5)	(6)
$T_0 = 5$	0.418	0.714	0.807	1.585	1.082	1.005
$T_0 = 20$	0.197	0.495	0.653	2.232	1.231	1.074
$T_0 = 50$	0.150	0.415	0.573	2.327	1.294	1.101
$T_0 = 100$	0.130	0.384	0.539	2.389	1.314	1.123

Notes: this table presents MC simulations from a stationary model. We consider the SC estimator that uses all pre-treatment outcome lags as economic predictors for a given $(T_0, \sigma^2_\varepsilon)$. In all simulations, we set $J+1 = 20$ and $K = 10$, which means that the 20 units are divided into 10 groups of 2 units that follow the same common factor λ^k_t. Columns 1 to 3 present the proportion of misallocated weights, which is given by the sum of weights allocated to units 3 to 20. Columns 4 to 6 present the ratio of standard errors of the DID estimator vs. the SC estimator.

Table 3: **Monte Carlo Simulations - Non-Stationary Model**

	Misallocation of weights			Misallocation of weights (non-stationary factors)			DID/SC ratio of standard errors		
	$\sigma_\varepsilon^2 = 0.1$	$\sigma_\varepsilon^2 = 0.5$	$\sigma_\varepsilon^2 = 1$	$\sigma_\varepsilon^2 = 0.1$	$\sigma_\varepsilon^2 = 0.5$	$\sigma_\varepsilon^2 = 1$	$\sigma_\varepsilon^2 = 0.1$	$\sigma_\varepsilon^2 = 0.5$	$\sigma_\varepsilon^2 = 1$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$T_0 = 5$	0.372	0.661	0.762	0.107	0.192	0.232	2.072	1.263	1.115
$T_0 = 20$	0.176	0.441	0.589	0.029	0.069	0.095	4.374	2.155	1.680
$T_0 = 50$	0.136	0.373	0.518	0.015	0.036	0.050	6.649	3.190	2.420
$T_0 = 100$	0.120	0.346	0.489	0.009	0.022	0.030	9.462	4.494	3.369

Notes: this table presents MC simulations results from a model with non-stationary and stationary common factors. We consider the SC estimator that uses all pre-treatment outcome lags as economic predictors for a given $(T_0, \sigma_\varepsilon^2, K)$. In all simulations, we set $J+1 = 20$, $K = 10$ (which means that the 20 units are divided into 10 groups of 2 units each that follow the same stationary common factor λ_t^k) and $R = 2$ (which means that the 20 units are divided into 2 groups of 10 units each that follow the same non-stationary common factor γ_t^r). Columns 1 to 3 present the proportion of misallocated weights, which is given by the sum of weights allocated to units 3 to 20. Columns 4 to 6 present the proportion of misallocated weights considering only the non-stationary common factor, which is given by the sum of weights allocated to units 11 to 20. Columns 7 to 9 present the ratio of standard errors of the DID estimator vs. the SC estimator.

A Supplemental Appendix: Revisiting the Synthetic Control Estimator (For Online Publication)

A.1 Proof of the Main Results

A.1.1 Proposition 1

Proof.

Let $\mathbf{y}_{0t} = (y_{2t}, ..., y_{J+1,t})'$, $\varepsilon_{0t} = (\varepsilon_{2t}, ..., \varepsilon_{J+1,t})'$, and $\mu_0 = (\mu_2, ..., \mu_{J+1})$. The SC weights $\widehat{\mathbf{w}} \in \mathbb{R}^J$ are given by

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}\in W} \frac{1}{T_0} \sum_{t\in\mathcal{T}_0} \left(y_{0t} - \mathbf{y}_{0t}'\mathbf{w}\right)^2 \tag{17}$$

where $W = \{\mathbf{w} \in \mathbb{R}^J | w_j \geq 0 \text{ and } \sum_{j\neq 1} w_j = 1\}$.[17]

Under Assumptions 1 and 3, the objective function $\widehat{Q}_{T_0}(\mathbf{w}) \equiv \frac{1}{T_0}\sum_{t\in\mathcal{T}_0}\left(y_{0t} - \mathbf{y}_{0t}'\mathbf{w}\right)^2$ converges pointwise in probability to

$$Q_0(\mathbf{w}) \equiv \sigma_\varepsilon^2(1 + \mathbf{w}'\mathbf{w}) + (\mu_1 - \mu_0\mathbf{w})'\,\Omega_0\,(\mu_1 - \mu_0\mathbf{w}) \tag{18}$$

which is a continuous and strictly convex function. Therefore, $Q_0(\mathbf{w})$ is uniquely minimized over W, and we define its minimum as $\bar{\mathbf{w}} \in W$.

We show that this convergence in probability is uniform over $\mathbf{w} \in W$. Define $\tilde{y}_{0t} = y_{0t} - \delta_t$ and $\tilde{\mathbf{y}}_{0t} = \mathbf{y}_{0t} - \delta_t\mathbf{i}$, where $\mathbf{i}$ is a $J \times 1$ vector of ones. For any $\mathbf{w}', \mathbf{w} \in W$, using the mean value theorem, we can find a $\widetilde{\mathbf{w}} \in W$ such that

$$\begin{aligned}\left|\widehat{Q}_{T_0}(\mathbf{w}') - \widehat{Q}_{T_0}(\mathbf{w})\right| &= \left|2\left(\frac{1}{T_0}\sum_{t\in\mathcal{T}_0}\tilde{\mathbf{y}}_{0t}\tilde{y}_{0t} - \frac{1}{T_0}\sum_{t\in\mathcal{T}_0}\tilde{\mathbf{y}}_{0t}\tilde{\mathbf{y}}_{0t}'\widetilde{\mathbf{w}}\right)\cdot(\mathbf{w}' - \mathbf{w})\right| \\ &\leq \left[\left(2\left|\left|\frac{1}{T_0}\sum_{t\in\mathcal{T}_0}\tilde{\mathbf{y}}_{0t}\tilde{y}_{0t}\right|\right| + \left|\left|\frac{1}{T_0}\sum_{t\in\mathcal{T}_0}\tilde{\mathbf{y}}_{0t}\tilde{\mathbf{y}}_{0t}'\right|\right| \times ||\widetilde{\mathbf{w}}||\right)\left|\left|\mathbf{w}' - \mathbf{w}\right|\right|\right]^{\frac{1}{2}}.\end{aligned} \tag{19}$$

Define $B_{T_0} = 2\left|\left|\frac{1}{T_0}\sum_{t\in\mathcal{T}_0}\tilde{\mathbf{y}}_{0t}\tilde{y}_{0t}\right|\right| + \left|\left|\frac{1}{T_0}\sum_{t\in\mathcal{T}_0}\tilde{\mathbf{y}}_{0t}\tilde{\mathbf{y}}_{0t}'\right|\right| \times C$. Since W is compact, $||\widetilde{\mathbf{w}}||$ is bounded, so we can find a constant C such that $\left|\widehat{Q}_{T_0}(\mathbf{w}') - \widehat{Q}_{T_0}(\mathbf{w})\right| \leq B_{T_0}\left(||\mathbf{w}' - \mathbf{w}||\right)^{\frac{1}{2}}$. Since $\tilde{y}_{0t}\tilde{\mathbf{y}}_{0t}$ and $\tilde{\mathbf{y}}_{0t}\tilde{\mathbf{y}}_{0t}'$ are linear combinations of cross products of λ_t and ε_{it}, from Assumptions 1 and 3 we have that B_{T_0} converges in probability to a positive constant, so $B_{T_0} = O_p(1)$. Note also that $Q_0(\mathbf{w})$ is uniformly continuous on W. Therefore, from Corollary 2.2 of Newey (1991), we have that $\widehat{Q}_{T_0}$ converges uniformly in probability to Q_0. Since Q_0 is uniquely minimized at $\bar{\mathbf{w}}$, W is a

[17] If the number of control units is greater than the number of pre-treatment periods, then the solution to this minimization problem might not be unique. However, since we consider the asymptotics with $T_0 \to \infty$, then we guarantee that, for large enough T_0, the solution will be unique.

compact space, Q_0 is continuous and $\widehat{Q}_{T_0}$ converges uniformly to Q_0, from Theorem 2.1 of Newey and McFadden (1994), $\widehat{\mathbf{w}}$ exists with probability approaching one, and $\widehat{\mathbf{w}} \xrightarrow{p} \bar{\mathbf{w}}$.

Now we show that $\bar{\mathbf{w}}$ does not generally reconstruct the factor loadings. Note that Q_0 has two parts. The first one reflects that different choices of weights will generate different weighted averages of the idiosyncratic shocks ε_{it}. In this simpler case, this part would be minimized when we set all weights equal to $\frac{1}{J}$. Let the $J \times 1$ vector $\mathbf{j_J} = \left(\frac{1}{J}, ..., \frac{1}{J}\right)' \in W$. The second part reflects the presence of common factors λ_t that would remain after we choose the weights to construct the SC unit. This part is minimized if we choose a $\mathbf{w}^* \in \Phi = \{\mathbf{w} \in W \mid \mu_0 = \mu_0 \mathbf{w}\}$. Suppose that we start at $\mathbf{w}^* \in \Phi$ and move in the direction of $\mathbf{j_J}$, with $\mathbf{w}(\Delta) = \mathbf{w}^* + \Delta(\mathbf{j_J} - \mathbf{w}^*)$. Note that, for all $\Delta \in [0, 1]$, these weights will continue to satisfy the constraints of the minimization problem. If we consider the derivative of function 18 with respect to Δ at $\Delta = 0$, we have that:

$$\Gamma'(\mathbf{w}^*) = 2\sigma_\varepsilon^2 \left(\frac{1}{J} - \mathbf{w}^{*\prime}\mathbf{w}^*\right) < 0 \text{ unless } \mathbf{w}^* = \mathbf{j_J} \text{ or } \sigma_\varepsilon^2 = 0$$

Therefore, $\mathbf{w}^*$ will not, in general, minimize Q_0. This implies that, when $T_0 \to \infty$, the SC weights will converge in probability to weights $\bar{\mathbf{w}}$ that does not reconstruct the factor loadings of the treated unit, unless it turns out that $\mathbf{w}^*$ also minimizes the variance of this linear combination of the idiosyncratic errors or if $\sigma_\varepsilon^2 = 0$. ■

A.1.2 Proposition 2

Proof.

The demeaned SC estimator is given by $\widehat{\mathbf{w}}^{\text{SC}'} = \underset{\mathbf{w} \in W}{\text{argmin}} \, \widehat{Q}'_{T_0}(\mathbf{w})$, where

$$\begin{aligned}
\widehat{Q}'_{T_0}(\mathbf{w}) &= \frac{1}{T_0} \sum_{t \in \mathcal{T}_0} \left(y_{0t} - \mathbf{y}'_{0t}\mathbf{w} - \left(\frac{1}{T_0} \sum_{t \in \mathcal{T}_0} y_{0t} - \frac{1}{T_0} \sum_{t \in \mathcal{T}_0} \mathbf{y}'_{0t}\mathbf{w} \right) \right)^2 \\
&= \widehat{Q}_{T_0}(\mathbf{w}) - \left(\frac{1}{T_0} \sum_{t \in \mathcal{T}_0} y_{0t} - \frac{1}{T_0} \sum_{t \in \mathcal{T}_0} \mathbf{y}'_{0t}\mathbf{w} \right)^2 . \qquad (20)
\end{aligned}$$

$\widehat{Q}'_{T_0}(\mathbf{w})$ converges pointwise in probability to

$$Q'_0(\mathbf{w}) \equiv \sigma_\varepsilon^2(1 + \mathbf{w}'\mathbf{w}) + (\mu_1 - \mu_0\mathbf{w})' \left(\Omega_0 - \omega'_0\omega_0\right) (\mu_1 - \mu_0\mathbf{w}) \qquad (21)$$

where $\Omega_0 - \omega'_0\omega_0$ is positive semi-definite, so $Q'_0(\mathbf{w})$ is a continuous and convex function.

The proof that $\widehat{\mathbf{w}}^{\text{SC}'} \xrightarrow{p} \bar{\mathbf{w}}^{\text{SC}'}$ where $\bar{\mathbf{w}}^{\text{SC}'}$ will generally not reconstruct the factor loadings of the treated unit follows exactly the same steps as the proof of Proposition 1. Therefore

$$\hat{\alpha}_{0t}^{\text{SC}'} = y_{0t} - \mathbf{y}_{0t}\widehat{\mathbf{w}}^{\text{SC}'} - \left[\frac{1}{T_0}\sum_{t'=-T_0+1}^{0} y_{0t} - \frac{1}{T_0}\sum_{t'=-T_0+1}^{0} \mathbf{y}_{0t}'\widehat{\mathbf{w}}^{\text{SC}'}\right] \tag{22}$$

$$\xrightarrow{d} \alpha_{0t} + \left(\varepsilon_{0t} - \varepsilon_{0t}'\bar{\mathbf{w}}^{\text{SC}'}\right) + (\lambda_t - \omega_0)\left(\mu_0 - \mu_0\bar{\mathbf{w}}^{\text{SC}'}\right). \tag{23}$$

■

A.1.3 Proposition 3

Proof.

For any estimator $\hat{\alpha}_{0t}(\widetilde{\mathbf{w}}) = y_{0t} - \mathbf{y}_{0t}\widetilde{\mathbf{w}} - \left[\frac{1}{T_0}\sum_{t'=-T_0+1}^{0} y_{0t} - \frac{1}{T_0}\sum_{t'=-T_0+1}^{0}\mathbf{y}_{0t}'\widetilde{\mathbf{w}}\right]$ such that $\widetilde{\mathbf{w}} \xrightarrow{p} \mathbf{w}$, we have that, under Assumptions 1, 2, 3 and 4,

$$a.var(\hat{\alpha}_{0t}(\widetilde{\mathbf{w}})|D(0,0)=1) = \sigma_\varepsilon^2(1+\mathbf{w}'\mathbf{w}) + (\mu_1 - \mu_0\mathbf{w})'\left(\Omega_0 - \omega_0'\omega_0\right)(\mu_1 - \mu_0\mathbf{w}) = Q_0'(\mathbf{w}), \tag{24}$$

which implies that $a.var(\hat{\alpha}_{0t}^{\text{SC}'}|D(0,0)=1) = Q_0'(\hat{\alpha}_{0t}^{\text{SC}'})$, $a.var(\hat{\alpha}_{0t}^{DID}|D(0,0)=1) = Q_0'(\hat{\alpha}_{0t}^{DID})$, and $a.var(\hat{\alpha}_{0t}^{*}|D(0,0)=1) = Q_0'(\hat{\alpha}_{0t}^{*})$. By definition of $\hat{\alpha}_{0t}^{\text{SC}'}$, it must be that $Q_0'(\hat{\alpha}_{0t}^{\text{SC}'}) \leq Q_0'(\hat{\alpha}_{0t}^{DID})$ and $Q_0'(\hat{\alpha}_{0t}^{\text{SC}'}) \leq Q_0'(\hat{\alpha}_{0t}^{*})$. ■

A.1.4 Proposition 4

Proof.

We show this result for the case without the adding-up, non-negativity, and no intercept constraints. In Appendix A.6.1 we extend these results for the cases with the adding-up and/or non-negativity constraints. In Appendix A.6.2 we show that this result is not valid when we use the no intercept constraint.

Note first that we can re-write model 10 as

$$\mathbf{Y}_t = \begin{bmatrix} \theta_1' \\ \vdots \\ \theta_{J+1}' \end{bmatrix} \gamma_t' + \tilde{\epsilon}_t = \Theta\gamma_t' + \tilde{\epsilon}_t, \tag{25}$$

where $\gamma_t = (\gamma_t^1, ..., \gamma_t^{F_1})$, and Θ is a $J+1 \times F$ matrix with the factor loadings associated with γ_t for all units and $\tilde{\epsilon}_t$ is an $\mathcal{I}(0)$ vector that includes the stationary common factors and the transitory shocks. Without loss of generality, we assume that the elements of γ_t are ordered so that its first element of γ_t is the deterministic polynomial trend with highest power, and the last elements are the $\mathcal{I}(1)$ common factors.

Suppose there are h linearly independent vectors $\mathbf{b} \in \mathbb{R}^{J+1}$ such that $\mathbf{b}'\Theta = 0$. In this case, we can consider the triangular representation

$$\mathbf{y}_{1t} = \Gamma' \mathbf{y}_{2t} + \mu_0^* + \mathbf{z}_t^*, \tag{26}$$

where $\mathbf{y}_{1t}$ is $h \times 1$, $\mathbf{y}_{2t}$ is $g \times 1$, and Γ' is $h \times g$; $\mathbf{z}_t^*$ is a $h \times 1$ $\mathcal{I}(0)$ series with mean zero and μ_0^* is an $h \times 1$ vector of constants. Given Assumption 5, we can write this representation with unit 1 in the vector $\mathbf{y}_{1t}$. Without loss of generality, we consider the case where $\mathbf{y}_{1t} = (y_{0t}, ..., y_{ht})'$ and $\mathbf{y}_{2t} = (y_{h+1,t}, ..., y_{J+1,t})'$. We define the matrix Θ_i^j as a submatrix with the lines i to j of matrix Θ. Importantly, note that equation 26 implies that $\Theta_1^h = \Gamma' \Theta_{h+1}^{J+1}$.

From the definition of $\mathbf{y}_{2t}$, we have that $rank(\Theta_{h+1}^{J+1}) = g$. Otherwise, it would be possible to find another linearly independent vector $v \in \mathbb{R}^{J+1}$ such that $v'\mathbf{y}_t$ is stationary, which contradicts the fact that the dimension of such space is h. We consider a linear transformation $\tilde{\mathbf{y}}_{2t} \equiv A\mathbf{y}_{2t}$ for some invertible $g \times g$ matrix A such that the matrix $\tilde{\Theta}_{h+1}^{J+1} \equiv A\Theta_{h+1}^{J+1}$ with elements $\tilde{\theta}_{j,f}$ has the following property: there exist integers $1 = f_1 < ... < f_g \leq F_1$ such that $\tilde{\theta}_{j,f_j} \neq 0$ and $\tilde{\theta}_{j,f} = 0$ if $f > f_j$. In words, this transformed vector $\tilde{\mathbf{y}}_{2t}$ is such that its n^{th} element does not contain a common factor of higher order than the highest order common factors for any element $j < n$ of $\tilde{\mathbf{y}}_{2t}$.

We show that it is possible to construct such matrix given the definition of $\mathbf{y}_{2t}$. We start setting $\tilde{y}_{1,t} = y_{j,t}$ for some $j \in \{h+1, .., J+1\}$ such that $\theta_{j,1} \neq 0$. For the second row, consider linear combinations $b'\mathbf{y}_{2t}$ for some $b \in \mathbb{R}^g$ and let $\tilde{\theta}_f(b)$ be the f-component of the $(1 \times F_1)$ row vector $b'\Theta_{h+1}^{J+1}$. Consider now the set of all linear combinations $b'\mathbf{y}_{2t}$ such that $\tilde{\theta}_1(b) = 0$, and let f_2 be the largest $f \in \{1, ..., F_1\}$ such that $\tilde{\theta}_{f_2}(b) \neq 0$ for some b in this set. We pick one b such that $\tilde{\theta}_1(b) = 0$ and $\tilde{\theta}_{f_2}(b) \neq 0$ and set $\tilde{y}_{2,t} = b'\mathbf{y}_{2t}$. For the third row, we consider linear combinations of $\mathbf{y}_{2t}$ such that $\tilde{\theta}_f(b) = 0$ for all $f \leq f_2$, and choose $\tilde{y}_{3,t}$ as a linear combination $b'\mathbf{y}_{2t}$ such that $\tilde{\theta}_{f_3}(b) \neq 0$. Since, $rank(\Theta_{h+1}^{J+1}) = g$, we can continue this construction until we get $\tilde{y}_{g,t} = b'\mathbf{y}_{2t}$ for a linear combination b such that $\tilde{\theta}_f(b) = 0$ for all $f \leq f_{g-1}$ with $\tilde{\theta}_f(b) \neq 0$ for at least one $f > f_{g-1}$.

Therefore, we have that

$$\mathbf{y}_{1t} = \Gamma' A^{-1} \tilde{\mathbf{y}}_{2t} + \mu_0^* + \mathbf{z}_t^*. \tag{27}$$

Now closely following the proof of proposition 19.3 in Hamilton (1994), we consider the OLS regression

$$z_{1t}^* = \alpha + \boldsymbol{\beta}' \mathbf{z}_{2t}^* + \phi' \tilde{\mathbf{y}}_{2t} + u_t \tag{28}$$

where z_{1t}^* is the first element of $\mathbf{z}_t^*$, and $\mathbf{z}_{2t}^* = (z_{2t}^*, ..., z_{ht}^*)'$.

Now let $\tilde{f}_k$ be equal to the order of the polynomial common factor $\gamma_t^{f_k}$ or equal to $\frac{1}{2}$ is $\gamma_t^{f_k}$ is an $\mathcal{I}(1)$ common factor. Then OLS estimator for this model is

$$\begin{bmatrix} \widehat{\boldsymbol{\beta}}-\boldsymbol{\beta} \\ \widehat{\alpha} \\ T_0^{\tilde{f}_1}\widehat{\phi}_1 \\ \vdots \\ T_0^{\tilde{f}_g}\widehat{\phi}_g \end{bmatrix} = \begin{bmatrix} \frac{\sum \mathbf{z}_{2t}^*\mathbf{z}_{2t}^{*\prime}}{T_0} & \frac{\sum \mathbf{z}_{2t}^*}{T_0} & \frac{\sum \mathbf{z}_{2t}^*\tilde{y}_{1,t}}{T_0^{\tilde{f}_1+1}} & \cdots & \frac{\sum \mathbf{z}_{2t}^*\tilde{y}_{g,t}}{T_0^{\tilde{f}_g+1}} \\ \frac{\sum \mathbf{z}_{2t}^{*\prime}}{T_0} & 1 & \frac{\sum \tilde{y}_{1,t}}{T_0^{\tilde{f}_1+1}} & \cdots & \frac{\sum \tilde{y}_{g,t}}{T_0^{\tilde{f}_g+1}} \\ \frac{\sum \tilde{y}_{1,t}\mathbf{z}_{2t}^{*\prime}}{T_0^{\tilde{f}_1+1}} & \frac{\sum \tilde{y}_{1,t}}{T_0^{\tilde{f}_1+1}} & \frac{\sum \tilde{y}_{1,t}^2}{T_0^{2\tilde{f}_1+1}} & \cdots & \frac{\sum \tilde{y}_{1,t}\tilde{y}_{g,t}}{T_0^{\tilde{f}_1+\tilde{f}_g+1}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\sum \tilde{y}_{g,t}\mathbf{z}_{2t}^{*\prime}}{T_0^{\tilde{f}_g+1}} & \frac{\sum \tilde{y}_{g,t}}{T_0^{\tilde{f}_g+1}} & \frac{\sum \tilde{y}_{g,t}\tilde{y}_{1,t}}{T_0^{\tilde{f}_1+\tilde{f}_g+1}} & \cdots & \frac{\sum \tilde{y}_{g,t}^2}{T_0^{2\tilde{f}_g+1}} \end{bmatrix}^{-1} \times \begin{bmatrix} T_0^{-1}\sum \mathbf{z}_{2t}^* u_t \\ T_0^{-1}\sum u_t \\ T_0^{-(1+\tilde{f}_1)}\sum \tilde{y}_{1,t}u_t \\ \vdots \\ T_0^{-(1+\tilde{f}_g)}\sum \tilde{y}_{g,t}u_t \end{bmatrix}. \quad (29)$$

Suppose that $\tilde{y}_{jt}$ has non-negative coefficients for at least one polynomial common factor for $j = 1, ..., g'$, while $\tilde{y}_{jt}$ has non-negative coefficients only for $\mathcal{I}(1)$ common factors for $j = g' + 1, ..., g$. We start showing that the first matrix in the right hand side of equation 29 converges to a matrix that is almost surely non-singular. Note that the terms $T_0^{-1}\sum \mathbf{z}_{2t}^*$ and $T_0^{-(\tilde{f}_j+1)}\sum \mathbf{z}_{2t}^*\tilde{y}_{j,t}$ converge in probability to zero, while $T_0^{-1}\sum \mathbf{z}_{2t}^*\mathbf{z}_{2t}^{*\prime} \xrightarrow{p} \mathbb{E}[\mathbf{z}_{2t}^*\mathbf{z}_{2t}^{*\prime}]$. Also, for $j \in \{1, ..., g'\}$, $\sum \tilde{y}_{j,t}$ is dominated by $\sum \tilde{\theta}_{j,f_j} t^{\tilde{f}_j}$, which implies that $T_0^{-(\tilde{f}_j+1)}\sum \tilde{y}_{j,t} \xrightarrow{p} \tilde{\theta}_{j,f_j}/(\tilde{f}_j+1)$. Similarly, for $(i, j) \in \{1, ..., g'\}$, $\sum \tilde{y}_{j,t}\tilde{y}_{i,t}$ is dominated by $\sum \tilde{\theta}_{j,f_j}\tilde{\theta}_{i,f_i} t^{\tilde{f}_i+\tilde{f}_j}$, which implies that $T_0^{-(\tilde{f}_j+\tilde{f}_i+1)}\sum \tilde{y}_{j,t}\tilde{y}_{i,t} \xrightarrow{p} \tilde{\theta}_{j,f_j}\tilde{\theta}_{i,f_i}/(\tilde{f}_i+\tilde{f}_j+1)$. Finally, the terms that include interactions with $\tilde{y}_{j,t}$ for $j \in \{g'+1, ..., g\}$ will converge in law to functions of an $(g-g')$-dimensional Brownian motion (with exception of those interacted with $\mathbf{z}_{2t}^*$, which, in this case, converge in probability to zero).[18] Putting these results together, we have that

$$\begin{bmatrix} \frac{\sum \mathbf{z}_{2t}^*\mathbf{z}_{2t}^{*\prime}}{T_0} & \frac{\sum \mathbf{z}_{2t}^*}{T_0} & \frac{\sum \mathbf{z}_{2t}^*\tilde{y}_{1,t}}{T_0^{\tilde{f}_1+1}} & \cdots & \frac{\sum \mathbf{z}_{2t}^*\tilde{y}_{g,t}}{T_0^{\tilde{f}_g+1}} \\ \frac{\sum \mathbf{z}_{2t}^{*\prime}}{T_0} & 1 & \frac{\sum \tilde{y}_{1,t}}{T_0^{\tilde{f}_1+1}} & \cdots & \frac{\sum \tilde{y}_{g,t}}{T_0^{\tilde{f}_g+1}} \\ \frac{\sum \tilde{y}_{1,t}\mathbf{z}_{2t}^{*\prime}}{T_0^{\tilde{f}_1+1}} & \frac{\sum \tilde{y}_{1,t}}{T_0^{\tilde{f}_1+1}} & \frac{\sum \tilde{y}_{1,t}^2}{T_0^{2\tilde{f}_1+1}} & \cdots & \frac{\sum \tilde{y}_{1,t}\tilde{y}_{g,t}}{T_0^{\tilde{f}_1+\tilde{f}_g+1}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\sum \tilde{y}_{g,t}\mathbf{z}_{2t}^{*\prime}}{T_0^{\tilde{f}_g+1}} & \frac{\sum \tilde{y}_{g,t}}{T_0^{\tilde{f}_g+1}} & \frac{\sum \tilde{y}_{g,t}\tilde{y}_{1,t}}{T_0^{\tilde{f}_1+\tilde{f}_g+1}} & \cdots & \frac{\sum \tilde{y}_{g,t}^2}{T_0^{2\tilde{f}_g+1}} \end{bmatrix} \xrightarrow{L} \begin{bmatrix} \mathbb{E}[\mathbf{z}_{2t}^*\mathbf{z}_{2t}^{*\prime}]_{h\times h} & \mathbf{0}_{h\times(g'+1)} & \mathbf{0}_{h\times(g-g')} \\ \mathbf{0}_{(g'+1)\times h} & \mathbf{C}_{(g'+1)\times(g'+1)} & \mathbf{D}'_{(g'+1)\times(g-g')} \\ \mathbf{0}_{(g-g')\times h} & \mathbf{D}_{(g-g')\times(g'+1)} & \mathbf{E}_{(g-g')\times(g-g')} \end{bmatrix} \equiv \mathbf{V} (30)$$

where $\mathbf{C}$ is a non-random matrix with the limits of the terms $T_0^{-(\tilde{f}_j+\tilde{f}_i+1)}\sum \tilde{y}_{j,t}\tilde{y}_{i,t}$ and $T_0^{-(\tilde{f}_i+1)}\sum \tilde{y}_{i,t}$ for $(i, j) \in \{1, ..., g'\}$, $\mathbf{E}$ is a random matrix for where the terms $T_0^{-(\tilde{f}_j+\tilde{f}_i+1)}\sum \tilde{y}_{j,t}\tilde{y}_{i,t}$ for $(i, j) \in \{g'+1, ..., g\}$ converge in law, and $\mathbf{D}$ is a random matrix for where the terms $T_0^{-(\tilde{f}_j+\tilde{f}_i+1)}\sum \tilde{y}_{j,t}\tilde{y}_{i,t}$ and $T_0^{-(\tilde{f}_j+1)}\sum \tilde{y}_{j,t}$ for $i \in \{1, ..., g'+1\}$ and $j \in \{g'+1, ..., g\}$ converge in law. Note that $\mathbb{E}[\mathbf{z}_{2t}^*\mathbf{z}_{2t}^{*\prime}]$ is non-singular by definition of $\mathbf{z}_{2t}^*$. It is also easy to show that $\mathbf{C}$ is non-singular.[19] Following the

[18] See the proof of proposition 19.3 in Hamilton (1994) for details.

[19] When $\tilde{\theta}_{j,f_j} \neq 0$ and $0 < f_1 < ... < f_{g'}$, which will be the case by construction, it is possible to diagonalize this matrix. For each row $j = 2, ..., g'+1$, we can subtract it by row 1 multiplied by $\frac{\theta_j}{1+f_j}$, and then divide that by $\frac{-f_j}{1+f_j}$. This will result in a matrix with the same entries as the original one, except that rows 2 to $g'+1$ in the first column

proof of Proposition 19.3 in Hamilton (1994), we also have that $\mathbf{E}$ is nonsingular with probability one. Therefore, we have that $\mathbf{V}$ is non-singular with probability one.[20]

Now we show that the second matrix in the right hand side of equation 29 converges in probability to zero. In this case, note that $\sum \tilde{y}_{j,t} u_t$ for $j = g'+1, ..., g$ is dominated by terms $\sum \xi_t u_t$ where ξ_t is $\mathcal{I}(1)$, which implies that $T_0^{-\frac{3}{2}} \sum \tilde{y}_{j,t} u_t \xrightarrow{p} 0$. For $j \in \{1, ..., g'\}$, note that $\sum \tilde{y}_{j,t} u_t$ is dominated by a term $\sum t^{\tilde{f}_j} u_t$. Therefore, $T_0^{-(1+\tilde{f}_j)} \sum \tilde{y}_{j,t} u_t$ converges in probability to zero. Finally, we also have that $T^{-1} \sum u_t$ and $T^{-1} \sum \mathbf{z}^*_{2t} u_t$ converge in probability to zero. Therefore, $\hat{\alpha} \xrightarrow{p} 0$, $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$, and $T^{\tilde{f}_i} \widehat{\phi}'_i \xrightarrow{p} 0$. From equations 27 and 28, we have that OLS estimator of y_{0t} on a constant and $y_{2t}, ..., y_{ht}, \tilde{y}_{h+1,t}, ..., \tilde{y}_{J+1,t}$ is given by $(\hat{\beta}' \quad \widehat{\phi}' + [1 \quad \hat{\beta}']\Gamma' A^{-1})$.[21] This implies that the OLS estimator of y_{0t} on a constant and $y_{2t}, ..., y_{J+1,t}$ is given by $\widehat{\mathbf{w}}' = (\hat{\beta}' \quad \widehat{\phi}' A + [1 \quad \hat{\beta}']\Gamma')$.

We are interested in the limiting distribution of $\hat{\alpha}_{0t}$, which is the effect of the treatment $\tau = t - T_0$ periods after the treatment started $(t > T_0)$. Note that

$$\begin{aligned}
\hat{\alpha}^{SC'}_{0t} &= \alpha_{0t} + \lambda_t \left(\mu_0 - \sum_{j \neq 1} \hat{w}_j \mu_j \right) + \gamma_t \left(\theta_1 - \sum_{j \neq 1} \hat{w}_j \theta_j \right) + \left(\varepsilon_{0t} - \sum_{j \neq 1} \hat{w}_j \varepsilon_{jt} \right) \\
& - \frac{1}{T_0} \sum_{t'=1}^{T_0} \left[\lambda'_t \left(\mu_0 - \sum_{j \neq 1} \hat{w}_j \mu_j \right) + \gamma'_t \left(\theta_1 - \sum_{j \neq 1} \hat{w}_j \theta_j \right) + \left(\varepsilon_{1t'} - \sum_{j \neq 1} \hat{w}_j \varepsilon_{jt'} \right) \right]. \quad (31)
\end{aligned}$$

For the term $\gamma_t \left(\theta_1 - \sum_{j \neq 1} \hat{w}_j \theta_j \right)$, note that

$$\begin{aligned}
\sum_{j \neq 1} \hat{w}_j \theta_j &= \begin{bmatrix} \Theta_2^{h'} & \Theta_{h+1}^{J+1'} \end{bmatrix} \widehat{\mathbf{w}} = \begin{bmatrix} \Theta_2^{h'} & \Theta_{h+1}^{J+1'} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ A' \hat{\phi} + \Gamma \begin{bmatrix} 1 \\ -\hat{\beta} \end{bmatrix} \end{bmatrix} \\
&= \Theta_2^{h'} \hat{\beta} + \Theta_{h+1}^{J+1'} A' \hat{\phi} + \Theta_{h+1}^{J+1'} \Gamma \begin{bmatrix} 1 \\ -\hat{\beta} \end{bmatrix} \quad (32) \\
&= \Theta_2^{h'} \hat{\beta} + \Theta_{h+1}^{J+1'} A' \hat{\phi} + \Theta_1^{h'} \begin{bmatrix} 1 \\ -\hat{\beta} \end{bmatrix} = \theta_1 + \Theta_{h+1}^{J+1'} A' \hat{\phi}.
\end{aligned}$$

Let $\Lambda = diag(T_0^{a_1}, ..., T_0^{a_F})$, where a_k is defined such that $\gamma^k_{T_0} T_0^{-a_k}$ converge either to a constant (when γ^k_t is a deterministic time trend) or to a distribution (when γ^k_t is an $\mathcal{I}(1)$ common factor).

will be equal to zero. Then for each row $j = 3, ..., g'+1$ we can subtract it by row 2 multiplied by $\frac{\theta_j}{\theta_1} \frac{1+2f_1}{1+f_1+f_j}$, and then divide it by $-\frac{f_j - f_1}{1+f_1+f_j}$. This will transform rows 3 to $g'+1$ in column 2 to zero. Continuing this procedure, we have an upper triangular matrix with diagonal elements different from zero.

[20]Note that $det(\mathbf{V}) = det(\mathbb{E}[\mathbf{z}^*_{2t} \mathbf{z}^{*'}_{2t}]) det(\mathbf{C} - \mathbf{D}'\mathbf{E}^{-1}\mathbf{D}) det(\mathbf{E})$. We have that $det(\mathbb{E}[\mathbf{z}^*_{2t} \mathbf{z}^{*'}_{2t}]) \neq 0$ and that $det(\mathbf{E}) \neq 0$ with probability one (which also implies that $\mathbf{E}^{-1}$ exists with probability one). Therefore, we only need that $det(\mathbf{C} - \mathbf{D}'\mathbf{E}^{-1}\mathbf{D}) \neq 0$ to guarantee that $\mathbf{V}$ is non-singular. Since $\mathbf{C}$ is non-singular, the realizations of $\mathbf{D}'\mathbf{E}^{-1}\mathbf{D}$ such that $\mathbf{C} - \mathbf{D}'\mathbf{E}^{-1}\mathbf{D}$ is singular will have measure zero, which implies that $\mathbf{V}$ is non-singular with probability one.

[21]Those are the estimators associated with $\mathbf{z}^*_{2t}$ and $\tilde{\mathbf{y}}_{2t}$. The estimator for the constant is given by $\hat{\alpha} + [1 \quad -\hat{\beta}']\mu^*_0$.

Then

$$\gamma_t \left(\theta_1 - \sum_{j \neq 1} \hat{w}_j \theta_j \right) = -\gamma_t \Theta_{h+1}^{J+1\prime} A' \hat{\phi} = -\gamma_t \Lambda^{-1} \Lambda \Theta_{h+1}^{J+1\prime} A' \hat{\phi}. \tag{33}$$

If $\gamma_t = t^k$, then $\gamma_t = (T_0 + (t - T_0))^k$, which implies that $T_0^{-k} \gamma_t = (1 + \frac{(t-T_0)}{T_0})^k \to 1$ when $T_0 \to \infty$. If γ_t is $\mathcal{I}(1)$, then $\gamma_t = \gamma_{T_0} + \sum_{t'=T_0+1}^{t} \eta_t$, which implies that $T_0^{-\frac{1}{2}} \gamma_t$ converges in distribution to a normal variable when $T_0 \to \infty$. Using the properties of $A\Theta_{h+1}^{J+1}$, we also have that the n^{th} row of $\Lambda \Theta_{h+1}^{J+1\prime} A' \hat{\phi}$ will be given by $T_0^{a_n}$ multiplied by a linear combination of elements $\hat{\phi}_j$ such that $f_j \geq a_n$. Therefore, the random variables $\hat{\phi}_j$ that are present in row n converge to zero at a faster rate than $T_0^{a_n}$, so $\Lambda \Theta_{h+1}^{J+1\prime} A' \hat{\phi} \xrightarrow{p} 0$. That is, we show that the SC weights will converge to weights that reconstruct the factor loadings of the treated unit associated with the non-stationary common factors, and the convergence in this case will be fast enough to compensate the fact that the non-stationary factors explode. Similarly, we have that $\frac{1}{T_0} \sum_{t'=1}^{T_0} \gamma_t' (\theta_1 - \sum_{j \neq 1} \hat{w}_j \theta_j) \xrightarrow{p} 0$.

Finally, by definition of u_t in equation 28, the OLS estimator converges to weights that minimize $var[u_t^2]$ subject to $\mathbf{w} \in \Phi_1$, where $u_t = \lambda_t (\mu_0 - \sum_{j \neq 1} w_j \mu_j) + (\varepsilon_{0t} - \sum_{j \neq 1} w_j \varepsilon_{jt})$. Therefore, the proof that $\widehat{\mathbf{w}} \xrightarrow{p} \bar{\mathbf{w}} \notin \Phi$ is essentially the same as the proof of Proposition 1.

Combining these results, we have that:

$$\hat{\alpha}_{0t} \xrightarrow{d} \alpha_{0t} + \left(\varepsilon_{0t} - \sum_{j \neq 1} \bar{w}_j \varepsilon_{jt} \right) + (\lambda_t - \omega_0) \left(\mu_0 - \sum_{j \neq 1} \bar{w}_j \mu_j \right) \tag{34}$$

where $\omega_0 = plim_{T_0 \to \infty} \frac{1}{T_0} \sum_{t'=1}^{T_0} \lambda_t$. ■

A.2 Case with finite T_0

We consider here the case with T_0 fixed. For weights $\{w_j^*\}_{j \neq 1} \in \Phi$, note that:

$$y_{0t} = \sum_{j=1}^{J+1} w_j^* y_{jt} + \eta_t, \text{ for } t \leq 0, \text{ where } \eta_t = \varepsilon_{0t} - \sum_{j=1}^{J+1} w_j^* \varepsilon_{jt} \tag{35}$$

Since $\sum_{j=2}^{J+1} w_j^* = 1$, we can write:

$$\tilde{y}_{0t} = \sum_{j=1}^{J} w_j^* \tilde{y}_{jt} + \eta_t \tag{36}$$

where $\tilde{y}_{jt} = y_{jt} - y_{J+1,t}$. The SC weights will be given by the OLS regression in 36 with the non-negativity constraints. We ignore for now the non-negativity constraints. If we let $\tilde{y}_{0t} =$

$(\tilde{y}_{2t}, ..., \tilde{y}_{Jt})'$, $\mathbf{w}_0^* = (w_2^*, ..., w_J^*)'$ and $\widehat{\mathbf{w}}_0 = (\widehat{w}_2, ..., \widehat{w}_J)'$, then we have that

$$\widehat{\mathbf{w}}_0 = \left(\sum_{t=-T_0+1}^{0} \tilde{y}_{0t} \tilde{y}_{0t}' \right)^{-1} \sum_{t=-T_0+1}^{0} \tilde{y}_{0t} \tilde{y}_{0t}.$$

We assume that T_0 is large enough so that $\sum_{t=-T_0+1}^{0} \tilde{y}_{0t} \tilde{y}_{0t}'$ has full rank. Therefore:

$$\mathbb{E}[\widehat{\mathbf{w}}_0 | \tilde{y}_{0,-T_0+1}, ..., \tilde{y}_{0,0}] = \mathbf{w}_0^* + \left(\sum_{t=-T_0+1}^{0} \tilde{y}_{0t} \tilde{y}_{0t}' \right)^{-1} \sum_{t=-T_0+1}^{0} \tilde{y}_{0t} \mathbb{E}[\eta_t | \tilde{y}_{0,-T_0+1}, ..., \tilde{y}_{0,0}] \quad (37)$$

By definition of η_t, we have that $\mathbb{E}[\eta_t | \tilde{y}_{0,-T_0+1}, ..., \tilde{y}_{0,0}] \neq 0$ for $t \leq 0$, which implies that $\widehat{\mathbf{w}}_0$ is a biased estimator of $\mathbf{w}_0^*$. Intuitively, the transitory shocks behave as a measurement error when we use the control outcomes as a proxy for the common factors. Considering the non-negativity constraints would affect the distribution of $\widehat{\mathbf{w}}_0$ because, with finite T_0, there will be a positive probability that the solution to the unrestricted OLS problem will not satisfy the non-negativity constraints. However, this would not change the conclusion that $\widehat{\mathbf{w}}_0$ is a biased estimator of $\mathbf{w}_0^*$.

A.3 Example: SC Estimator vs DID Estimator

We provide an example in which the asymptotic bias of the SC estimator can higher than the asymptotic bias of the DID estimator. Assume we have 1 treated and 4 control units in a model with 2 common factors. For simplicity, assume that there is no additive fixed effects and that $\mathbb{E}[\lambda_t] = 0$. We have that the factor loadings are given by:

$$\mu_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \mu_2 = \begin{pmatrix} 0.5 \\ 1 \end{pmatrix}, \mu_3 = \begin{pmatrix} 1.5 \\ 1 \end{pmatrix}, \mu_4 = \begin{pmatrix} 0.5 \\ 0 \end{pmatrix}, \mu_5 = \begin{pmatrix} 1.5 \\ 1 \end{pmatrix} \quad (38)$$

Note that the linear combination $0.5\mu_2 + w_1^3\mu_3 + w_1^5\mu_5 = \mu_0$ with $w_1^3 + w_1^5 = 0.5$ satisfy Assumption **??**. Note also that DID equal weights would set the first factor loading to 1, which is equal to μ_0^1, but the second factor loading would be equal to $0.75 \neq \mu_0^2$. We want to show that the SC weights would improve the construction of the second factor loading but it will distort the combination for the first factor loading. If we set $\sigma_\varepsilon^2 = \mathbb{E}[(\lambda_t^1)^2] = \mathbb{E}[(\lambda_t^2)^2] = 1$, then the factor loadings of the SC unit would be given by $(1.038, 0.8458)$. Therefore, there is small loss in the construction of the first factor loading and a gain in the construction of the second factor loading. Therefore, if selection into treatment is correlated with the common shock λ_t^1, then the SC estimator would be more asymptotically biased than the DID estimator.

A.4 Definition: Asymptotically Unbiased

We now show that the expected value of the asymptotic distribution will be the same as the limit of the expected value of the SC estimator in the setting described in Section 3. Let γ be the expected value of the asymptotic distribution of $\hat{\alpha}_{0t} - \alpha_{0t}$. Therefore, we have that:

$$\begin{aligned}\mathbb{E}[\hat{\alpha}_{0t} - \alpha_{0t}] &= \gamma + E\left[\sum_{j\neq 1}(\bar{w}_j - \hat{w}_j)\varepsilon_{jt}\right] + E\left[\lambda_t \sum_{j\neq 1}(\bar{w}_j - \hat{w}_j)\mu_j\right] \\ &= \gamma + \sum_{j\neq 1} E\left[(\bar{w}_j - \hat{w}_j)\varepsilon_{jt}\right] + \sum_{j\neq 1} E\left[\lambda_t(\bar{w}_j - \hat{w}_j)\right]\mu_j\end{aligned}$$

Therefore:

$$|E\left[(\bar{w}_j - \hat{w}_j)\varepsilon_{jt}\right]| \leq E\left[|(\bar{w}_j - \hat{w}_j)\varepsilon_{jt}|\right] \leq \sqrt{E\left[(\bar{w}_j - \hat{w}_j)^2\right] E\left[(\varepsilon_{jt})^2\right]}$$

Now note that $\hat{w}_j$ is a consistent estimator for $\bar{w}_j$ and the random variable $(\bar{w}_j - \hat{w}_j)^2$ is bounded, because W is compact. Therefore, the sequence $(\bar{w}_j - \hat{w}_j)^2$ is asymptotically uniformly integrable, which implies that $E\left[(\bar{w}_j - \hat{w}_j)^2\right] \to 0$. If we also assume that ε_{it} and λ_t^f for all $f = 1, ..., F$ have finite variance, then $\mathbb{E}[\hat{\alpha}_{0t} - \alpha_{0t}] \to \gamma$ when $T_0 \to \infty$.

A.5 Alternatives specifications and alternative estimators

A.5.1 Average of pre-intervention outcome as economic predictor

We consider now another very common specification in SC applications, which is to use the average pre-treatment outcome as the economic predictor. Note that if one uses only the average pre-treatment outcome as the economic predictor then the choice of matrix V would be irrelevant. In this case, the minimization problem would be given by:

$$\begin{aligned}\{\hat{w}_j\}_{j\neq 1} &= \text{argmin}_{w\in W}\left[\frac{1}{T_0}\sum_{t\in\mathcal{T}_0}\left(y_{0t} - \sum_{j\neq 1} w_j y_{jt}\right)\right]^2 \\ &= \text{argmin}_{w\in W}\left[\frac{1}{T_0}\sum_{t\in\mathcal{T}_0}\left(\varepsilon_{0t} - \sum_{j\neq 1} w_j\varepsilon_{jt} + \lambda_t\left(\mu_0 - \sum_{j\neq 1} w_j\mu_j\right)\right)\right]^2 \qquad (39)\end{aligned}$$

where $W = \{\{w_j\}_{j\neq 1} \in \mathbb{R}^J | w_j \geq 0 \text{ and } \sum_{j\neq 1} w_j = 1\}$.

Therefore, under Assumptions 1, 2 and 3, the objective function converges in probability to:

$$\Gamma(\mathbf{w}) = \left[E\left[\lambda_t | D(0,0) = 1\right]\left(\mu_0 - \sum_{j\neq 1} w_j\mu_j\right)\right]^2 \qquad (40)$$

Assuming that there is a time-invariant common factor (that is, $\lambda_t^1 = 1$ for all t) and that the pre-treatment average of the conditional process λ_t converges to $\mathbb{E}[\lambda_t^k] = 0$ for $k > 1$, the objective function collapses to:

$$\Gamma(\mathbf{w}) = \left[\left(\mu_1^1 - \sum_{j\neq 1} w_j \mu_j^1\right)\right]^2 \tag{41}$$

Therefore, even if we assume that there exists at least one set of weights that reproduces all factor loadings ($\Phi \neq \varnothing$), the objective function will only look for weights that approximate the first factor loading. This is problematic because there might be $\{\tilde{w}_j\}_{j\neq 1} \notin \Phi$ that satisfy $\mu_1^1 = \sum_{j\neq 1} \tilde{w}_j \mu_j^1$. In this case, there is no guarantee that the SC control method will choose weights that are close to the correct ones. This result is consistent with the MC simulations in Ferman et al. (2017), who show that this specification performs particularly bad in allocating the weights correctly.

A.5.2 Adding other covariates as predictors

Most SC applications that use the average pre-intervention outcome value as economic predictor also consider other time invariant covariates as economic predictors. Let Z_i be a $(R \times 1)$ vector of observed covariates (not affected by the intervention). Model 1 changes to:

$$\begin{cases} y_{it}(0) = \delta_t + \theta_t Z_i + \lambda_t \mu_i + \varepsilon_{it} \\ y_{it}(1) = \alpha_{it} + y_{it}(0) \end{cases} \tag{42}$$

We redefine the set $\Phi = \{\mathbf{w} \in W \mid \mu_0 = \sum_{j\neq 1} w_j^* \mu_j,\ Z_1 = \sum_{j\neq 1} w_j^* Z_j\}$. Let X_1 be an $(R+1\times 1)$ vector that contains the average pre-intervention outcome and all covariates for unit 1, while X_0 is a $(R+1 \times J)$ matrix that contains the same information for the control units. For a given V, the first step of the nested optimization problem suggested in Abadie et al. (2010) would be given by:

$$\widehat{\mathbf{w}}(V) \in \text{argmin}_{\mathbf{w}\in W} ||X_1 - X_0\mathbf{w}||_V \tag{43}$$

where $W = \{\{w_j\}_{j\neq 1} \in \mathbb{R}^J | w_j \geq 0 \text{ and } \sum_{j\neq 1} w_j = 1\}$. Assuming again that there is a time-invariant common factor (that is, $\lambda_t^1 = 1$ for all t) and that the pre-treatment average of the unconditional process λ_t converges to $\mathbb{E}[\lambda_t^k] = 0$ for $k > 1$, objective function of this minimization problem converges to $||\bar{X}_1 - \bar{X}_0\mathbf{w}||_V$, where:

$$\bar{X}_1 - \bar{X}_0\mathbf{w} = \begin{bmatrix} \mathbb{E}[\theta_t | D(0, T_0) = 1] \left(Z_1 - \sum_{j\neq 1} w_j Z_j\right) + \left(\mu_1^1 - \sum_{j\neq 1} w_j \mu_j^1\right) \\ \left(Z_1^1 - \sum_{j\neq 1} w_j Z_j^1\right) \\ \vdots \\ \left(Z_1^R - \sum_{j\neq 1} w_j Z_j^R\right) \end{bmatrix} \tag{44}$$

Similarly to the case with only the average pre-intervention outcome value as economic predictor, it might be that $\Phi \neq \varnothing$, but there are weights $\{\tilde{w}_j\}_{j\neq 1}$ that satisfy $\mu_1^1 = \sum_{j\neq 1} \tilde{w}_j \mu_j^1$ and $Z_1 = \sum_{j\neq 1} \tilde{w}_j Z_j$, although $\mu_1^k \neq \sum_{j\neq 1} \tilde{w}_j \mu_j^k$ for some $k > 1$. Therefore, there is no guarantee that an estimator based on this minimization problem would converge to weights in Φ for any given matrix V, even if $\Phi \neq \varnothing$.

The second step in the nested optimization problem is to choose V such that $\widehat{\mathbf{w}}(V)$ minimizes the pre-intervention prediction error. Note that this problem is essentially given by:

$$\widehat{\mathbf{w}} \quad = \quad \text{argmin}_{w \in \widetilde{W}} \left[\frac{1}{T_0} \sum_{t \in \mathcal{T}_0} \left(y_{0t} - \sum_{j \neq 1} w_j y_{jt} \right) \right]^2 \tag{45}$$

where $\widetilde{W} \subseteq W$ is the set of $\mathbf{w}$ such that $\mathbf{w}$ is the solution to problem 43 for some positive semidefinite matrix V. Similarly to the SC estimator that includes all pre-treatment outcomes, there is no guarantee that this minimization problem will choose weights in Φ, even when $T_0 \rightarrow \infty$. Therefore, it is not possible to guarantee that this SC estimator would be asymptotically unbiased. MC simulation presented by Ferman et al. (2017) confirm that this SC specification systematically misallocates more weight than alternatives that use a large number of pre-treatment outcome lags as predictors.

A.5.3 Relaxing constraints on the weights

If we assume that $W = \mathbb{R}^J$ instead of the compact set $\{\widehat{\mathbf{w}} \in \mathbb{R}^J | w_j \geq 0 \text{ and } \sum_{j \neq 1} w_j = 1\}$, then we can still guarantee consistency of the SC weights. The only difference is that we also need to assume convergence of the pre-treatment averages of δ_t. In Proposition 1 this was not necessary because the adding-up restriction implies that δ_t was always eliminated. Consider the model

$$y_{it}(0) = \dot{\lambda}_t \dot{\mu}_i + \varepsilon_{it} \tag{46}$$

where $\dot{\lambda}_t = (\delta_t, \lambda_t)$ and $\dot{\mu}_i = (1, \mu_i)'$. We modify Assumption 3 to include assumptions on the convergence of δ_t.

Assumption 3$''$ (convergence of pre-treatment averages) Conditional on $D(0,0) = 1$, $\frac{1}{T_0} \sum_{t \in \mathcal{T}_0} \dot{\lambda}_t \xrightarrow{p} \dot{\omega}_0$, $\frac{1}{T_0} \sum_{t \in \mathcal{T}_0} \dot{\lambda}_t' \dot{\lambda}_t \xrightarrow{p} \dot{\Omega}_0$, $\frac{1}{T_0} \sum_{t \in \mathcal{T}_0} \epsilon_t \xrightarrow{p} 0$, $\frac{1}{T_0} \sum_{t \in \mathcal{T}_0} \epsilon_t \epsilon_t' \xrightarrow{p} \sigma_\varepsilon^2 I_{J+1}$, $\varepsilon_{jt} \perp \dot{\lambda}_s$, and $\frac{1}{T_0} \sum_{t \in \mathcal{T}_0} \epsilon_t \dot{\lambda}_t \xrightarrow{p} 0$ when $T_0 \rightarrow \infty$.

Note first that, under assumptions 1 and 3$''$, the objective function converges in probability to

$$\widehat{Q}_{T_0}(\mathbf{w}) \xrightarrow{p} \dot{Q}_0(\mathbf{w}) \equiv \sigma_\varepsilon^2 (1 + \mathbf{w}'\mathbf{w}) + (\dot{\mu}_1 - \dot{\mu}_0 \mathbf{w})' \, \dot{\Omega}_0 \, (\dot{\mu}_1 - \dot{\mu}_0 \mathbf{w}) \,, \tag{47}$$

where $\dot{Q}_0(\mathbf{w})$ is continuous and strictly convex. Since W is a convex space, $\dot{Q}_0(\mathbf{w})$ has a unique minimum that is in the interior of W. Therefore, by Theorem 2.7 of Newey and McFadden (1994), $\widehat{\mathbf{w}}$ exists with probability approaching one and $\widehat{\mathbf{w}} \xrightarrow{p} \mathbf{w}_0$.

For the case $W = \{\mathbf{w} \in \mathbb{R}^J \mid \sum_{j=2}^{J+1} w_j = 1\}$, note that the transformed model with $y_{0t} - y_{2t}$ as the outcome of the treated unit and $y_{3t} - y_{2t}, ..., y_{J+1,t} - y_{2t}$ as the outcomes of the control units is equivalent to the original model. Then we can use the same arguments on this modified model.

Consistency when we impose only the non-negativity constraint follows from the same arguments as in Appendix A.6.1.

Given that we assure convergence of $\widehat{\mathbf{w}}$, the fact that $\widehat{\mathbf{w}}$ does not converge to weights that reconstruct the factor loadings of the treated unit follows from the same arguments as the proof of Proposition 1. Note that, without the adding-up constraint, it might be that the asymptotic distribution of the SC estimator depends on δ_t.

A.5.4 IV-Like SC Estimator

Consider again equation 35. The key problem is that η_t is correlated with y_{jt}, which implies that the restricted OLS estimators are biased and inconsistent. Imposing strong assumptions on the structure of the idiosyncratic error and the common factors, we show that it is possible to consider moment equations that will be equal to zero if, and only if, $\{w_j\}_{j\neq 1} \in \Phi$.

Let $\mathbf{y}_{0t} = (y_{2,t}, ..., y_{J+1,t})'$, $\mu_{\mathbf{0}}$ be a $(F \times J)$ matrix with columns μ_j, $\epsilon_{0t} = (\varepsilon_{2,t}, ..., \varepsilon_{J+1,t})$, and $\mathbf{w} = (w_2, ..., w_{J+1})'$. In this case, we can look at

$$\begin{aligned} \mathbf{y}_{t-1}(y_{0t} - \mathbf{y}_{0t}'\mathbf{w}) &= (\mu_0'\lambda_{t-1}' + \epsilon_{0,t-1})\lambda_t\,(\mu_1 - \mu_{\mathbf{0}}\mathbf{w}) + (\mu_0'\lambda_{t-1}' + \epsilon_{0,t-1})(\varepsilon_{0t} - \epsilon_{0t}'\mathbf{w}) \\ &= \mu_0'\lambda_{t-1}'\lambda_t\,(\mu_1 - \mu_{\mathbf{0}}\mathbf{w}) + \epsilon_{0,t-1}\lambda_t\,(\mu_1 - \mu_{\mathbf{0}}\mathbf{w}) + \mu_0'\lambda_{t-1}'(\varepsilon_{0t} - \epsilon_{0t}'\mathbf{w}) + \epsilon_{0,t-1}(\varepsilon_{0t} - \epsilon_{0t}'\mathbf{w}). \end{aligned} \tag{48}$$

Under Assumptions 1 and 3, and assuming further that ε_{it} is independent across t, then the objective function given by $\frac{1}{T_0}\sum_{t=-T_0+1}^{0} \mathbf{y}_{t-1}(y_{0t} - \mathbf{y}_{0t}'\mathbf{w})$ converges uniformly to $\mathbb{E}[\mathbf{y}_{0,t-1}(y_{0t} - \mathbf{y}_{0t}'\mathbf{w})] = \mu_0'\mathbb{E}[\lambda_{t-1}'\lambda_t](\mu_1 - \mu_{\mathbf{0}}\mathbf{w})$

Therefore, if the $(J \times F)$ matrix $\mu_0'\mathbb{E}[\lambda_{t-1}'\lambda_t]$ has full rank, then the moment conditions equal to zero if, and only if, $\mathbf{w} \in \Phi$. One particular case in which this assumption is valid is if λ_t^f and $\lambda_t^{f'}$ are uncorrelated and λ_t^f is serially correlated for all $f = 1, ..., F$. Intuitively, under these assumptions, we can use the lagged outcome values of the control units as instrumental variables for the control units' outcomes.[22] One challenge to analyze this method is that there might be multiple solutions to the moment condition. Based on the results by Chernozhukov et al. (2007), it is possible to consistently estimate this set. Therefore, it is possible to generate an IV-like SC estimator that is, under additional assumptions, asymptotically unbiased.

A.6 Extensions on Proposition 4

A.6.1 Relaxing the adding-up and non-negativity constraints

To show that this result is also valid for the case with adding-up constraint we just have to consider the OLS regression of $y_{0t} - y_{2t}$ on a constant and $y_{3t} - y_{2t}, ..., y_{J+1,t} - y_{2t}$. Under Assumption 5,

[22] The idea of SC-IV is very similar to the IV estimator used in dynamic panel data. In the dynamic panel models, lags of the outcome are used to deal with the endogeneity that comes from the fact the idiosyncratic errors are correlated with the lagged depend variable included in the model as covariates. The number of lags that can be used as instruments depends on the serial correlation of the error terms.

this transformed model is also cointegrated, so we can apply our previous result.

We now consider the case with the non-negative constraints. We prove the case $W = \{\mathbf{w} \in \mathbb{R}^J \mid w_j \geq 0\}$. Including an adding-up constraint then follows directly from a change in variables as we did for the case without non-negative constraints.

We first show that $\widehat{\mathbf{w}} \xrightarrow{p} \bar{\mathbf{w}}$ where $\bar{\mathbf{w}}$ minimizes $\mathbb{E}[u_t^2]$ subject to $\mathbf{w} \in \Phi_1 \cap W$. Suppose that $\bar{\mathbf{w}} \in int(W)$. This implies that $\bar{\mathbf{w}} \in int(\Phi_1 \cap W)$ relative to Φ_1. By convexity of $E[u_t^2]$, $\bar{\mathbf{w}}$ also minimizes $E[u_t^2]$ subject to Φ_1. We know that OLS without the non-negativity constraints converges in probability to $\bar{\mathbf{w}}$. Let $\widehat{\mathbf{w}}_u$ be the OLS estimator without the non-negativity constraints and $\widehat{\mathbf{w}}_r$ be the OLS estimator with the non-negativity constraint. Since $\bar{\mathbf{w}} \in int(W)$, then it must be that, for all $\varepsilon > 0$, $Pr(|\widehat{\mathbf{w}}_u - \bar{\mathbf{w}}| > \varepsilon) = 0$ with probability approaching to 1 (w.p.a.1). Since $\widehat{\mathbf{w}}_u = \widehat{\mathbf{w}}_r$ when $\widehat{\mathbf{w}}_u \in int(W)$ (due to convexity of the OLS objective function), these two estimators are asymptotically equivalent.

Consider now the case in which $\bar{\mathbf{w}}$ is on the boundary of W. This means that $\bar{w}_j = 0$ for at least one j. Let $A = \{j | w_j^* = 0\}$. Note first that $\bar{\mathbf{w}}$ also minimizes $E[u_t^2]$ subject to $\mathbf{w} \in \Phi \cap \{\mathbf{w} | w_j = 0 \ \forall j \in A\}$. That is, if we impose the restriction $w_j = 0$ for all j such that $\bar{w}_j = 0$, then we would have the same minimizer, even if we ignore the other non-negative constraints. Suppose there is an $\tilde{\mathbf{w}} \neq \bar{\mathbf{w}}$ that minimizes $E[u_t^2]$ subject to $\mathbf{w} \in \Phi \cap \{\mathbf{w} | w_j = 0 \ \forall j \in A\}$. By convexity of the objective function and the fact that $\bar{\mathbf{w}}$ is in the interior of $\Phi \cap W \cap \{\mathbf{w} | w_j = 0 \ \forall j \in A\}$ relative to $\Phi \cap \{\mathbf{w} | w_j = 0 \ \forall j \in A\}$, there must be $\mathbf{w}' \in \Phi \cap W \cap \{\mathbf{w} | w_j = 0 \ \forall j \in A\} \subset \Phi \cap W$ that attains a lower value in the objective function than $\bar{\mathbf{w}}$. However, this contradicts the fact that $\bar{\mathbf{w}} \in \Phi \cap W$ is the minimum.

Now let $\widehat{\mathbf{w}}'$ be the OLS estimator subject to $\{\mathbf{w} | w_j = 0 \ \forall j \in A\}$. We have that $\widehat{\mathbf{w}}'$ is consistent for $\bar{\mathbf{w}}$ (Lemma **??**). Now we show that $\widehat{\mathbf{w}}'$ is asymptotically equivalent to $\widehat{\mathbf{w}}''$, the OLS estimator subject to $\{\mathbf{w} | w_j \geq 0 \ \forall j \in A\}$. We prove the case in which $A = \{j\}$ (there is only one restriction that binds). The general case follows by induction. Suppose these two estimators are not asymptotically equivalent. Then there is $\varepsilon > 0$ such that $LimPr(|\widehat{\mathbf{w}}' - \widehat{\mathbf{w}}''| > \varepsilon) \neq 0$. There are two possible cases.

First, suppose that $\text{Lim}Pr\left(|\widehat{w}_j''| > \varepsilon'\right) = 0$ for all $\varepsilon' > 0$ (that is, the OLS subject to $\{\mathbf{w} | w_j \geq 0 \ \forall j \in A\}$ converges in probability to $\bar{\mathbf{w}}$ such that $\bar{w}_j = 0$). However, since the two estimators are not asymptotically equivalent, for all T_0', we can always find a $T_0 > T_0'$ such that, with positive probability, $|\widehat{\mathbf{w}}' - \widehat{\mathbf{w}}''| > \varepsilon$. Since $\{\mathbf{w} | w_j = 0 \ \forall j \in A\} \subset \{\mathbf{w} | w_j \geq 0 \ \forall j \in A\}$ and $\widehat{\mathbf{w}}' \neq \widehat{\mathbf{w}}''$, then $Q_{T_0}(\widehat{\mathbf{w}}'') < Q_{T_0}(\widehat{\mathbf{w}}')$, where $Q_{T_0}()$ is the OLS objective function. Now using the continuity of the OLS objective function and the fact that $\widehat{w}_j''$ converges in probability to zero, we can always find T_0' such that there will be a positive probability that $Q_{T_0}(\widehat{\mathbf{w}}'' - e_j \hat{w}_j'') < Q_{T_0}(\widehat{\mathbf{w}}')$. Since $\widehat{\mathbf{w}}'' - e_j \hat{w}_j'' \in \{\mathbf{w} | w_j = 0 \ \forall j \in A\}$, this contradicts $\widehat{\mathbf{w}}'$ being OLS subject to $\{\mathbf{w} | w_j = 0 \ \forall j \in A\}$.

Alternatively, suppose that there exists $\varepsilon' > 0$ such that $\text{Lim}Pr\left(|\widehat{w}_j''| > \varepsilon'\right) \neq 0$. This means that, for all T_0', we can find $T_0 > T_0'$ such that there is a positive probability that the solution to OLS on $\{\mathbf{w} | w_j \geq 0 \ \forall j \in A\}$ is in an interior point $\widehat{\mathbf{w}}''$ with $\hat{w}_j'' > \varepsilon' > 0$. By convexity of $Q_{T_0}()$, this would imply that $\widehat{\mathbf{w}}''$ is also the solution to the OLS without any restriction. However, this contradicts the fact that OLS without non-negativity restriction is consistent (see proof of

Proposition 4).

Finally, we show that $\widehat{\mathbf{w}}''$ and $\widehat{\mathbf{w}}_r$ are asymptotically equivalent. Note that $\bar{\mathbf{w}}$ is in the interior of W relative to $\{\mathbf{w}|w_j \geq 0 \; \forall j \in A\}$. Therefore, w.p.a.1, $\widehat{\mathbf{w}}'' \in W$, which implies that $\widehat{\mathbf{w}}'' = \widehat{\mathbf{w}}_r$.

We still need to show that linear combinations of $\widehat{\mathbf{w}}^r$ converge fast enough to reconstruct the factor loadings of the treated unit associated with the non-stationary common factors, so that $\gamma_t(\theta_1 - \sum_{j\neq 1} \hat{w}_j^r \theta_j) \xrightarrow{p} 0$. Let $Q_{T_0}()$ be the OLS objective function, and let $\widetilde{\mathcal{W}} = \{\widetilde{\mathbf{w}}_1, ..., \widetilde{\mathbf{w}}_{2^J}\}$ be the set of all possible OLS estimators when we consider some of the non-negative constraints as equality and ignore the other ones. Let $\widetilde{\mathcal{W}}' \subset \widetilde{\mathcal{W}}$ be the set of estimators in $\widetilde{\mathcal{W}}$ such that all non-negative constraints are satisfied. Then we know that $\widehat{\mathbf{w}}^r = argmin_{\mathbf{w}\in\widetilde{\mathcal{W}}'} Q_{T_0}(\mathbf{w})$.

Suppose first that, for any of the 2^J combinations of restrictions, there is at least one $\mathbf{w} \in \Phi_1$ that satisfy these restrictions. In this case, we know from the first part of the proof that $\gamma_t \left(\theta_1 - \sum_{j\neq 1} \widetilde{w}_j^h \theta_j\right) \xrightarrow{p} 0$ for all $h = 1, ..., 2^J$, where $\widetilde{\mathbf{w}}_h = (\widetilde{w}_2^h, ..., \widetilde{w}_{J+1}^h)'$. Moreover, since $\widetilde{\mathcal{W}}$ is finite, then this convergence is uniform in $\widetilde{\mathcal{W}}$. Therefore, it must be that $\gamma_t(\theta_1 - \sum_{j\neq 1} \hat{w}_j^r \theta_j) \xrightarrow{p} 0$. Suppose now that for the combination of restrictions considered for $\widetilde{\mathbf{w}}_h$, with $h \in \{1, ..., 2^J\}$, there is no $\mathbf{w} \in \Phi_1$ that satisfies these restrictions. Since the parameter space with this combination of restrictions is closed, then $\exists \eta > 0$ such that $||\theta_1 - \sum_{j\neq 1} w_j \theta_0|| > \eta$ for all $\mathbf{w}$ that satisfy this combinations of restrictions.[23] Therefore, $Q_{T_0}(\widetilde{\mathbf{w}}_h)$ diverge when $T_0 \to \infty$, implying that, w.p.a.1, $\widehat{\mathbf{w}}^r \neq \widetilde{\mathbf{w}}_h$.

A.6.2 Example with no intercept

We consider now a very simple example to show that it is not possible to guarantee that $\gamma_t \left(\theta_1 - \sum_{j\neq 1} \hat{w}_j \theta_j\right) \xrightarrow{p} 0$ if we do not include the intercept. Consider the case in which there are only one treated and one control unit, and $y_{0t} = \mu_0 + t + u_{1t}$ while $y_{2t} = \mu_2 + t + u_{2t}$. We consider a regression of y_{0t} on y_{2t} without the intercept. Note that $y_{0t} = (\mu_0 - \mu_2) + y_{2t} + u_{1t} - u_{2t} = \mu + y_{2t} + u_t$. Then we have that:

$$\hat{\beta} = \frac{\sum_{t=1}^{T_0} y_{2t} y_{0t}}{\sum_{t=1}^{T_0} y_{2t}^2} = 1 + \frac{\sum_{t=1}^{T_0} (\mu\mu_2 + \mu t + \mu u_{2t} + \mu_2 u_t + t u_t + u_t u_{2t})}{\sum_{t=1}^{T_0} (t^2 + \mu_2^2 + u_{2t}^2 + \text{“cross terms”})} \tag{49}$$

which implies that:

$$T(\hat{\beta} - 1) = \frac{\frac{1}{T^2} \sum_{t=1}^{T_0} (\mu\mu_2 + \mu t + \mu u_{2t} + \mu_2 u_t + t u_t + u_t u_{2t})}{\frac{1}{T^3} \sum_{t=1}^{T_0} (t^2 + \mu_2^2 + u_{2t}^2 + \text{“cross terms”})} \xrightarrow{p} \frac{\frac{1}{2}\mu}{\frac{1}{3}} \tag{50}$$

Therefore, while $\hat{\beta} \xrightarrow{p} 1$, it does not converge fast enough so that $T(\hat{\beta} - 1) \xrightarrow{p} 0$, except when $\mu_0 = \mu_2$.

[23] Otherwise, there would be $\mathbf{w} \in \Phi_1$ that satisfies this combination of restrictions.