Revisiting the Synthetic Control Estimator

Ferman, Bruno and Pinto, Cristine

Sao Paulo School of Economics - FGV

23 November 2016

Revisiting the Synthetic Control Estimator*

Bruno Ferman[†] Cristine Pinto[‡]

Sao Paulo School of Economics - FGV

First Draft: June, 2016

This Draft: August, 2017

Please click here for the most recent version

Abstract

We analyze the conditions under which the Synthetic Control (SC) estimator is asymptotically unbiased when the number of pre-treatment periods goes to infinity. We show that the SC estimator is generally asymptotically biased if treatment assignment is correlated with time-varying unobserved confounders, and this may be true even if the pre-treatment fit is almost perfect. While we also show that the SC method can substantially improve relative to standard methods, our results suggest that researchers should be more careful in interpreting the identification assumptions required for this method.

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

*We would like to thank Alberto Abadie, Ivan Canay, Jin Hahn, Guido Imbens, Aureo de Paula, Tong Li, Ricardo Masini, Rodrigo Soares and seminar participants at the California Econometrics Conference 2016, Bristol Econometric Study Group 2016, IAAE conference 2017, EEA conference 2017, USP, George Washington University, University of Miami, PUC-Chile, IPEA and PUC-Rio for comments and suggestions. We also thank Deivis Angeli for excellent research assistance.

[†]bruno.ferman@fgv.br

[‡]cristine.pinto@fgv.br

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. Then they 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 of the method, Abadie et al. (2010) derive conditions under which the SC estimator would allow for confounding unobserved characteristics that vary with time, as long as we can fit a long set of pre-intervention periods.[1]

In this paper, we analyze the conditions under which the SC estimator is asymptotically unbiased when the number of pre-treatment periods (T_0) goes to infinity in a linear factor model setting. Differently from Abadie et al. (2010), we do not condition the analysis on a perfect pre-treatment match. In a model with "stationary" common factors, we show that the SC weights converge in probability to weights that do *not*, in general, reconstruct the factor loadings of the treated unit, even if such weights exist.[2,3] This happens because 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 will not generally be the solution to this problem, even if such weights exist. While, as argued in Doudchenko and Imbens (2016), in many SC applications the number of pre-treatment periods may not be large enough to justify large-T_0 asymptotics, our results should be interpreted as the SC weights not converging to weights that reconstruct the factor loadings of the treated unit *even when T_0 is large*. Based on our Monte Carlo (MC) simulation results, the SC weights should be even farther from weights that reconstruct the factor loadings of the treated unit when T_0 is small.

[1] Abadie et al. (2010) derive this result based on a linear factor model for the potential outcomes. However, they point out that the SC estimator can be useful in more general contexts.

[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 all pre-treatment periods as economic predictors. We also consider the case of the average of the pre-treatment periods and the average of the pre-treatment periods plus other covariates as predictors in Appendix A.4.

As a consequence, the SC estimator is, in general, biased if treatment assignment is correlated with the unobserved heterogeneity, even when the number of pre-treatment periods goes to infinity.[4] The intuition is the following: if the fact that unit 1 was treated after period T_0 is informative about the common factors, then we would need a SC unit that was 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 units. However, the fact that the SC weights do not converge, in general, to weights that reconstruct the factor loadings of the treated unit implies that 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.[5] These results do not rely on the fact the SC unit is constrained to convex combinations of control units, which implies that they also apply to the panel data approach suggested in Hsiao et al. (2012).

One important implication of the SC restriction to convex combinations of the control units is that the SC estimator 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) model. We therefore recommend a slight modification in the SC method where we demean the data using the pre-intervention period, and then construct the SC estimator using the demeaned data.[6] If selection into treatment is only correlated with time-invariant common factors, then this demeaned SC estimator is unbiased. Assuming further that time-varying common factors are stationary, we also guarantee that the asymptotic variance of this demeaned SC estimator is weakly lower than the asymptotic variance of the DID estimator. If selection into treatment is correlated with time-varying common factors, then both the demeaned SC and the DID estimators would be asymptotically biased. We show that the asymptotic bias of the demeaned SC estimator is lower than the bias of DID for a particular class of linear factor models.[7]

[4] We define the asymptotic bias as the difference between the expected value of the asymptotic distribution and the parameter of interest. We show in Appendix A.3 that, in the context of the SC estimator, the limit of the expected value converges to the expected value of the asymptotic distribution.

[5] Ando and Sävje (2013) point out that the SC estimator can be biased if the weights do not reconstruct the factor loadings of the treated unit. They argue that this may be the case if there is no set of weights that reconstructs the factor loadings of the treated unit with a weighted average of the factor loadings of the control units. 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.

[6] Demeaning the data before applying the SC estimator is equivalent to a generalization of the SC method suggested in Doudchenko and Imbens (2016) which includes an intercept parameter in the minimization problem to estimate the SC weights.

[7] This result is only valid for a particular set of linear factor models. Outside this set of linear factor models, we provide a very specific example in which the asymptotic bias of the SC can be larger. This might happen when selection into treatment depends on common factors with low variance and a simple average of the control units provides a good approximation for the

Overall, while we argue that the SC method is, in general, asymptotically biased if treatment assignment is correlated with time-varying confounders, it can still provide important improvement over DID, even if a close-to-perfect pre-treatment match is not achieved. Our results from Monte Carlo (MC) simulations suggest that such improvement can be attained even if T_0 is small.[8]

Note that our results for models with "stationary" common factors are not as conflicting with the results in Abadie et al. (2010) as it might appear at first glance. The asymptotic bias of the SC estimator, in this case, goes to zero when the variance of the transitory shocks is small, in which case one should expect to have a good pre-treatment match and, therefore, Abadie et al. (2010) would recommend using the SC method. When a subset of the common factors is non-stationary, however, we show that the asymptotic bias may not go to zero even in situations where one would expect a close-to-perfect pre-treatment fit. In a model with both $I(1)$ and $I(0)$ common factors, we show that the SC weights will converge to weights that reconstruct the factor loadings associated to the $I(1)$ common factors of the treated unit, but that will generally fail to reconstruct the factor loadings associated with the $I(0)$ common factors.[9] Therefore, in this setting, asymptotic unbiasedness requires that treatment assignment is uncorrelated with the $I(0)$ common factors. The same is true when we consider a model with unit specific polynomial time trends. Importantly, these cases show that, when a subset of the common factors is non-stationary, a close-to-perfect pre-treatment match for a long set of pre-intervention periods does not guarantee that the asymptotic bias of the SC estimator is close to zero. Therefore, we recommend that researchers applying the SC method should also assess the pre-treatment fit of the SC estimator after de-trending the data.[10] We show that prominent SC applications that display a seemingly perfect pre-treatment fit in the original data does not provide such a perfect pre-treatment fit once the data is de-trended.

Our paper is related to a recent literature that analyzes the asymptotic properties of the SC estimator and of generalizations of the method. Gobillon and Magnac (2016) derive conditions under which the assumption

factor loadings associated with these common factors.

[8]We also provide in Appendix A.4.4 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.

[9]We assume that the vector of outcomes is cointegrated. In the SC setting, this assumption is equivalent to the existence of weights that reconstruct the factor loadings of unit 1 associated with the $I(1)$ common factors. See Carvalho et al. (2016) for a discussion on the construction of counter-factual units with $I(1)$ data with no cointegration.

[10]Note that 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. The only caveat is that measures of pre-intervention fit could be misleading as diagnostic tests, as they may hide important discrepancies in the factor loadings associated to stationary common factors beyond these non-stationary trends. Given that, we recommend alternative diagnostic tests.

of perfect match in Abadie et al. (2010) can be satisfied when both the number of pre-treatment periods *and* the number of control units go to infinity.[11] Xu (2017) proposes an alternative to the SC method in which in a first step he estimates the factor loadings, and then in a second step he constructs the SC unit to match the estimated factor loadings of the treated unit. This method also requires a large number of both control units and pre-treatment units, so that the factor loadings are consistently estimated. Differently from Gobillon and Magnac (2016) and Xu (2017), we consider the case with a finite number of control units and let the number of pre-intervention periods go to infinity. We show that, in this case, the SC estimator can be asymptotically biased when $T_0 \to \infty$ even when the pre-treatment fit is almost perfect.[12] Carvalho et al. (2015) and Carvalho et al. (2016) also propose an alternative method that is related to the SC estimator, and derive conditions under which their estimator yields a consistent estimator. However, in a linear factor model as the one we consider, their assumptions would essentially exclude the possibility that treatment assignment is correlated with the unobserved heterogeneity.[13] Finally, building on the results from our paper, 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. Note, however, that we also show that the standard 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 the strategy proposed in Powell (2017).[14]

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 and consider different asymptotics

[11]They require that the matching variables (factor loadings and exogenous covariates) of the treated units belong to the support of the matching variables of control units. In this case, the SC estimator would be equivalent to the interactive effect methods they recommend.

[12]Wong (2015) also considers the asymptotic properties of the SC estimator when T_0 goes to infinity while holding the number of control units constant. He argues that the SC estimator would be asymptotically unbiased. However, we show in Appendix A.5 that the conditions we find such that the SC estimator is asymptotically biased also lead to an asymptotically biased estimator in his settings.

[13]Their main assumption is that the outcomes of the control units are independent of treatment assignment. However, in our setting, if we assume that transitory shocks are uncorrelated with the treatment assignment, then the potential outcomes of the treated unit being correlated with treatment assignment implies that treatment assignment is correlated with the common factors. If this is the case, then it cannot be that the outcomes of the control units are independent of the treatment assignment. In an extension, Carvalho et al. (2015) consider the case in which the intervention also affects the control units. They model that as a structural change in the common factors after the treatment, in which case they find that their estimator would be biased. Note, however, that they do not treat such change in the common factors as selection on unobservables. Instead, they consider this as a case in which the intervention *affects* all units.

[14]In fact, the 2-step procedure proposed in Powell (2017) can exacerbate the bias of the SC estimator if there is a correlation between treatment assignment and stationary common factors. For example, consider the case in which the variance of the transitory shocks is close to zero, so that the bias of the standard SC estimator is also close to zero. Since the procedure proposed in Powell (2017) essentially discards all variation beyond the time trends, it will generally fail to provide weights that match the factor loadings of these stationary common factors, leading to a biased estimator.

than Abadie et al. (2010). In Section 3, we show that, in a model such that the first and second moments of the common factors converge, the SC estimator is, in general, asymptotically biased.[15] In Section 4, we contrast the SC estimator with the DID estimator, and propose the demeaned SC estimator. In Section 5, we consider a setting in which pre-treatment averages of the common factor diverge, and we show that, in this case, the SC estimator can be asymptotically biased even if we have a close-to-perfect pre-treatment match. We revisit the applications in Abadie and Gardeazabal (2003), Abadie et al. (2010), and Abadie et al. (2015) in light of these results. In Section 6, we present a particular class of linear factor models in which we consider the asymptotic properties of the SC estimator and Monte Carlo simulations with finite T_0. We conclude in Section 7.

2 Base Model

Suppose we have a balanced panel of $J+1$ units indexed by i observed on $t = 1, ..., T$ periods. We want to estimate the treatment effect of a policy change that affected only unit $j = 1$ from period $T_0 + 1 \leq T$ to T. The potential outcomes are given by:

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

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_i is a $(F \times 1)$ vector of unknown factor loadings, and the error terms ϵ_{it} are unobserved transitory shocks. We only observe $y_{it} = d_{it} y_{it}(1) + (1 - d_{it}) y_{it}(0)$, where $d_{it} = 1$ if unit i is treated at time t. Since 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_i) as fixed and the common factors (λ_t) as random variables. In order to simplify the exposition of our main results, we consider the model without observed covariates Z_i. In Appendix Section A.4.2 we consider the model with covariates.

An important feature of our setting is that the SC estimator is only well defined if it actually happened that one unit received treatment in a given period. We define $D(j, T_0)$ as a dummy variable equal to 1 if

[15] We focus on the SC specification that uses all pre-treatment outcome lags as economic predictors. Asymptotic properties of alternative specifications of the SC estimator are considered in Section A.4.

unit j is treat after T_0 while all other units do not receive treatment.[16] Assumption 1 makes it clear that the sample a researcher observes when considering the SC estimator is always conditional on the fact that one unit was treated in a given period. Without loss of generalization, we consider that unit 1 is treated.

Assumption 1 (conditional sample) We observe a realization of $\{y_{1t}, ..., y_{J+1,t}\}$ for $t = 1, ..., T$ conditional on $D(1, T_0) = 1$.

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

Assumption 2 (transitory shocks) $E[\epsilon_{jt}|D(1, T_0)] = E[\epsilon_{jt}] = 0$

Assumption 2 implies that transitory shocks are mean-independent from the treatment assignment. However, we still allow for the possibility that the treatment assignment to unit 1 is correlated with the unobserved common factors. More specifically, we allow for $E[\lambda_t|D(1, T_0)] \neq E[\lambda_t]$. To better understand the implications of this possibility, suppose that the treatment is more likely to happen in unit j at time t if $\lambda_t \mu_j$ is high, and let λ_t^1 be a common factor that strongly affects unit 1.[17] Under these conditions, the fact that unit 1 is treated after T_0 is informative about the common factor λ_t^1, because one should expect $E[\lambda_t^1|D(1, T_0) = 1] > E[\lambda_t^1]$. Note that we allow for dependence between treatment assignment and common factors both before and after the start of the treatment. So we can consider, for example, a case in which treatment is triggered in unit 1 by a sequence of positive shocks on $\lambda_t \mu_1$ even before T_0.

In order to present the main intuition of the SC estimator, we assume that there exists a stable linear combination of the control units that absorbs all time correlated shocks of unit 1, $\lambda_t \mu_1$. Note, however, that this assumption is not necessary for any of our main results. Following the original SC papers, we restrict to convex combinations of the control units. We relax these constraints in Section 4.

Assumption 3 (existence of weights)

$$\exists \ \mathbf{w}^* \in \mathbb{R}^J \mid \mu_1 = \sum_{j \neq 1} w_j^* \mu_j, \ \sum_{j \neq 1} w_j^* = 1, \text{ and } w_j^* \geq 0$$

[16] That is, one can think of $D(j, T_0)$ as a product between two indicator variables, one for the event that the treated unit is unit j, and the other one that the treatment starts after T_0.

[17] That is, the factor loading of unit 1 associated with this common factor, μ_1^1, is large.

There is no guarantee that there is only one set of weights that satisfies Assumption 3, so we define $\Phi = \{\mathbf{w} \in \mathbb{R}^J \mid \mu_1 = \sum_{j \neq 1} w_j \mu_j, \ \sum_{j \neq 1} w_j = 1, \text{ and } w_j \geq 0\}$ as the set of weights that satisfy this condition.

If we knew $\mathbf{w}^* \in \Phi$, then we could consider an infeasible SC estimator using these weights, $\hat{\alpha}^*_{1t} = y_{1t} - \sum_{j \neq 1} w^*_j y_{it}$. For a given $t > T_0$, we would have that:

$$\hat{\alpha}^*_{1t} = y_{1t} - \sum_{j \neq 1} w^*_j y_{it} = \alpha_{1t} + \left(\epsilon_{1t} - \sum_{j \neq 1} w^*_j \epsilon_{jt} \right) \tag{2}$$

Therefore, under Assumption 2, we have that $E[\hat{\alpha}^*_{1t} | D(1, T_0) = 1] = \alpha_{1t}$, which implies that this infeasible SC estimator is unbiased. Note that we have to consider the expected value of $\hat{\alpha}^*_{1t}$ conditional on $D(1, T_0) = 1$, since we only observe a conditional sample (Assumption 1). Intuitively, the infeasible SC estimator constructs a SC unit for the counterfactual of y_{1t} that is affected in the same way as unit 1 by each of the common factors (that is, $\mu_1 = \sum_{j \neq 1} w^*_j \mu_j$), but did not receive treatment. Therefore, the only difference between unit 1 and this SC unit, beyond the treatment effect, would be given by the transitory shocks, which we assumed are not related to the treatment assignment. This guarantees that a SC estimator, using these infeasible weights, provides an unbiased estimator.

It is important to note that Abadie et al. (2010) do note make any assumption on the existence of weights that reconstruct the factor loadings of the treated unit. Instead, they consider that there is a set of weights that satisfies $y_{1t} = \sum_{j \neq 1} w^*_j y_{jt}$ for all $t \leq T_0$. 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 good pre-intervention fit. As stated in 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 $y_{1t} = \sum_{j \neq 1} w^*_j y_{jt}$ for all $t \leq T_0$ (for large T_0) implies that Assumption 3 holds approximately. In this case, the bias of the SC estimator would be bounded by a term that goes to zero when T_0 increases. We depart from the original SC setting in that we do not condition on having a perfect pre-intervention fit. The motivation to analyze the SC method in our setting is that, even if Assumption 3 is valid, in a model with only "stationary" factors the probability that we find a perfect pre-intervention fit in the data converges to zero when $T_0 \to \infty$, unless the variance of the transitory shocks is equal to zero. Still, we show that the SC method can provide important improvement over alternative methods even if the pre-intervention fit is imperfect. Moreover, we also show in Section 5 that, if a subset of the common factors is non-stationary, then the SC estimator may be asymptotically biased even if the pre-treatment fit is almost perfect.

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_1 is a $(K \times 1)$ vector containing the predictors for the treated unit and X_0 is a $(K \times J)$ matrix of economic predictors for the control units.[18] The SC weights are estimated by minimizing $||X_1 - X_0\mathbf{w}||_V$ subject to $\sum_{i=2}^{J+1} w_j = 1$ and $w_j \geq 0$, where V is a $(K \times K)$ positive semidefinite matrix. They discuss different possibilities for choosing the matrix V, including an iterative process where V is chosen such that the solution to the $||X_1 - X_0\mathbf{w}||_V$ optimization problem minimizes the pre-intervention prediction error. In other words, let $\mathbf{Y}_1^P$ be a $(T_0 \times 1)$ vector of pre-intervention outcomes for the treated unit, while $\mathbf{Y}_0^P$ 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 $||\mathbf{Y}_1^P - \mathbf{Y}_0^P \widehat{\mathbf{w}}(V)||$.

We focus on the case where one includes all pre-intervention outcome values as economic 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)), so the optimization problem suggested by Abadie et al. (2010) to estimate the weights simplifies to:

$$\widehat{\mathbf{w}} = \underset{\mathbf{w} \in W}{\operatorname{argmin}} \frac{1}{T_0} \sum_{t=1}^{T_0} \left[y_{1t} - \sum_{j \neq 1} w_j y_{jt} \right]^2 \tag{3}$$

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

In Appendix A.4 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 Asymptotic Bias 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 $z_t = (\epsilon_{1t}, ..., \epsilon_{J+1,t}, \lambda_t')$.

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

Assumption 4 (convergence of pre-treatment averages) $\frac{1}{T_0}\sum_{t=1}^{T_0} z_t \xrightarrow{p} a$ with $||a|| < \infty$, and $\frac{1}{T_0}\sum_{t=1}^{T_0} z_t z_t' \xrightarrow{p} A$, where A is a positive-definite matrix.

In order to simplify the exposition of our results, we consider an alternative set of assumptions that is stronger than necessary for our main results.

Assumption 4′ (convergence of pre-treatment averages) $\frac{1}{T_0}\sum_{t=1}^{T_0} \lambda_t \xrightarrow{p} \omega_0$, $\frac{1}{T_0}\sum_{t=1}^{T_0} \lambda_t'\lambda_t \xrightarrow{p} \Omega_0$, $\frac{1}{T_0}\sum_{t=1}^{T_0} \epsilon_{jt} \xrightarrow{p} 0$, $\frac{1}{T_0}\sum_{t=1}^{T_0} \epsilon_{jt}^2 \xrightarrow{p} \sigma_\epsilon^2$, and $\epsilon_{jt} \perp \lambda_s$ for all s, t and for all j.

Note that assumption 4 would be satisfied if the process z_t is weakly stationary and second order ergodic in the pre-treatment period conditional on $D(1, T_0) = 1$. 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 $E[\lambda_t|D(1, T_0)] \neq E[\lambda_t]$, even for $t < T_0$, which will happen if treatment assignment to unit 1 is correlated with common factors before T_0. In this case, it would be too restrictive to impose the assumption that, conditional on $D(1, T_0) = 1$, λ_t is stationary, even if only the pre-treatment periods.

We show first the convergence of $\hat{\mathbf{w}}$.

Proposition 1 Under assumptions 1, 2 and 4′, we have that $\hat{\mathbf{w}} \xrightarrow{p} \bar{\mathbf{w}}$ where $\mu_1 \neq \sum_{j\neq 1} \bar{w}_j \mu_j$, unless $\sigma_\epsilon^2 = 0$ or $\exists \mathbf{w} \in \Phi | \mathbf{w} \in \underset{\mathbf{w}\in W}{\operatorname{argmin}} \left\{\sum_{j\neq 1}(w_j)^2\right\}$

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 $\hat{\mathbf{w}}$ will converge in probability to $\bar{\mathbf{w}}$ such that:

$$\bar{\mathbf{w}} = \underset{\mathbf{w}\in W}{\operatorname{argmin}} \left\{ \sigma_\epsilon^2 \left(1 + \sum_{j\neq 1}(w_j)^2\right) + \left(\mu_1 - \sum_{j\neq 1} w_j\mu_j\right)' \Omega_0 \left(\mu_1 - \sum_{j\neq 1} w_j\mu_j\right) \right\} \tag{4}$$

which is the probability limit of the M-estimator objective function (equation 3).

Note that 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_{it}. 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 $\frac{1}{J}$.[19] The second part reflects the presence of common factors λ_t that would remain after we choose

[19] If we do not impose this restriction, then this part would be minimized setting all weights equal to zero, and our main argument would remain valid.

the weights to construct the SC unit. If assumption 3 is satisfied, 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 $\sigma_\epsilon^2 = 0$ or $\exists \mathbf{w} \in \Phi$ such that $\mathbf{w} \in \underset{\mathbf{w} \in W}{\text{argmin}} \left\{ \sum_{j \neq 1} (w_j)^2 \right\}$. Therefore, the SC weights will not generally converge to weights that reconstruct the factor loadings of the treated unit. Note that it may be that $\Phi = \varnothing$, in which case Proposition 1 trivially holds.

For a given $t > T_0$, the SC estimator will be given by:

$$\hat{\alpha}_{1t} = y_{1t} - \sum_{j \neq 1} \hat{w}_j y_{it} \xrightarrow{d} \alpha_{1t} + \left(\epsilon_{1t} - \sum_{j \neq 1} \bar{w}_j \epsilon_{jt} \right) + \lambda_t \left(\mu_1 - \sum_{j \neq 1} \bar{w}_j \mu_j \right) \tag{5}$$

Note that $\hat{\alpha}_{1t}$ converges in distribution to the parameter we want to estimate (α_{1t}) plus linear combinations of contemporaneous transitory shocks and common factors. Therefore, the SC estimator will be asymptotically unbiased if, conditional on the fact that unit 1 was treated in period t, the expected values of these linear combinations of transitory shocks and common factors are equal to zero.[20] More specifically, we need that $E\left[\epsilon_{1t} - \sum_{j \neq 1} \bar{w}_j \epsilon_{jt} | D(1, T_0) = 1 \right] = 0$ and $E\left[\lambda_t \left(\mu_1 - \sum_{j \neq 1} \bar{w}_j \mu_j \right) | D(1, T_0) = 1 \right] = 0$. While the first equality is guaranteed by Assumption 2, the second one will generally not hold if treatment assignment is correlated with the unobserved heterogeneity.

Since $\mu_1 \neq \sum_{j \neq 1} \bar{w}_j \mu_j$, the SC estimator will only be asymptotically unbiased, in general, if we impose an additional assumption that $E\left[\lambda_t^k | D(1, T_0) = 1 \right] = 0$ for all common factors k such that $\mu_1^k \neq \sum_{j \neq 1} \bar{w}_j \mu_j^k$. In order to better understand the intuition behind this result, we consider a special case in which, unconditionally, λ_t is stationary and the pre-treatment averages of the conditional process converge in probability to the unconditional expectations.[21] In this case, we can assume, without loss of generality, that $E[\lambda_t^1] = 1$ and $E[\lambda_t^k] = 0$ for $k > 0$. Therefore, the SC estimator will only be asymptotically unbiased if the weights turn out to recover unit 1 fixed effect (that is, $\mu_1^1 = \sum_{j \neq 1} \mu_j^1$) *and* treatment assignment is uncorrelated with time-varying unobserved common factors.[22]

[20] We consider the definition of asymptotic unbiasedness as the expected value of the asymptotic distribution of $\hat{\alpha}_{1t} - \alpha_{1t}$ equal to zero. An alternative definition is that $E[\hat{\alpha}_{1t} - \alpha_{1t}] \to 0$. We show in Appendix A.3 that these two definitions are equivalent in our setting under standard assumptions.

[21] This allows for correlation between common factors and treatment assignment prior to T_0, but limits this dependence in the sense that this dependence becomes irrelevant for the pre-treatment average once we consider a long history before treatment.

[22] While, as argued in Doudchenko and Imbens (2016), many SC applications does not have a large number of pre-treatment

Abadie et al. (2010) argue that, in contrast to the usual DID model, the SC model would allow the effects of confounding unobserved characteristics to vary with time. It is important to note that the discrepancy of our results arises because we rely on different assumptions. Abadie et al. (2010) consider the properties of the SC estimator conditional on having a good 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 SC estimator is bounded by a term that goes to zero when T_0 increases, *if the pre-treatment fit is perfect*. Note that our results are not as conflicting with the results in 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 in 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 good fit for a large T_0. In Section 5, however, we show that this result may not hold if we have non-stationary common factors.

In Appendix A.4 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 economic predictor, and the specification that uses the pre-treatment average of the outcome variable and other time-invariant covariates as economic 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 & alternative SC estimators

Our results from Session 3 show 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

periods to justify large-T_0 asymptotics, our results should be interpreted as the SC weights not converging to weights that reconstruct the factor loadings of the treated unit *even when T_0 is large*. In Section 6 we consider in MC simulations the behavior of the SC estimator when T_0 is finite.

in a given post-intervention period $t > T_0$, under Assumption 4′, would be given by:

$$\begin{aligned} \hat{\alpha}_{1t}^{DID} &= y_{1t} - \frac{1}{J}\sum_{j\neq 1} y_{jt} - \frac{1}{T_0}\sum_{\tau=1}^{T_0}\left[y_{1\tau} - \frac{1}{J}\sum_{j\neq 1} y_{j\tau}\right] \\ &\overset{d}{\rightarrow} \epsilon_{1t} - \frac{1}{J}\sum_{j\neq 1}\epsilon_{jt} + (\lambda_t - \omega_0)\left(\mu_1 - \frac{1}{J}\sum_{j\neq 1}\mu_j\right) \end{aligned} \tag{6}$$

Therefore, the DID estimator will be asymptotically unbiased if $E[\lambda_t | D(1,T_0) = 1] = \omega_0$, which means that the fact that unit 1 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). However, the DID estimator will be asymptotically biased if the fact that unit 1 is treated after period T_0 is informative about variations in the common factors relative to their pre-treatment mean.

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 Doudchenko and Imbens (2016) which includes an intercept parameter in the minimization problem to estimate the SC weights. The demeaned SC estimator is given by $\hat{\alpha}_{1t}^{SC'} = y_{1t} - \sum_{j\neq 1}\hat{w}_j^{\text{SC}'} y_{jt} - (\bar{y}_1 - \sum_{j\neq 1}\hat{w}_j^{\text{SC}'}\bar{y}_j)$, where $\bar{y}_j$ is the pre-treatment average of unit j, and the weights $\widehat{\mathbf{w}}^{\text{SC}'} = \{\hat{w}_j^{\text{SC}'}\}_{j=2}^{J+1}$ are given by:

$$\widehat{\mathbf{w}}^{\text{SC}'} = \underset{\mathbf{w}\in W}{\text{argmin}}\frac{1}{T_0}\sum_{t=1}^{T_0}\left[y_{1t} - \sum_{j\neq 1} w_j y_{jt} - \left(\bar{y}_1 - \sum_{j\neq 1} w_j \bar{y}_j\right)\right]^2 \tag{7}$$

Proposition 2 Under assumptions 1, 2 and 4′, we have that $\widehat{\mathbf{w}}^{\text{SC}'} \overset{p}{\rightarrow} \bar{\mathbf{w}}^{\text{SC}'}$ where $\mu_1 \neq \sum_{j\neq 1}\bar{w}_j^{\text{SC}'}\mu_j$, unless $\sigma_\epsilon^2 = 0$ or $\exists \mathbf{w} \in \Phi | \mathbf{w} \in \underset{\mathbf{w}\in W}{\text{argmin}}\left\{\sum_{j\neq 1}(w_j)^2\right\}$. Moreover:

$$\hat{\alpha}_{1t}^{\text{SC}'} \overset{d}{\rightarrow} \alpha_{1t} + \left(\epsilon_{1t} - \sum_{j\neq 1}\bar{w}_j^{\text{SC}'}\epsilon_{jt}\right) + (\lambda_t - \omega_0)\left(\mu_1 - \sum_{j\neq 1}\bar{w}_j^{\text{SC}'}\mu_j\right) \tag{8}$$

Proof.

See details in Appendix A.1.2 ∎

Therefore, the demeaned SC estimator is asymptotically unbiased under the same conditions as the DID estimator. Under the stronger assumption that the conditional process $z_t = (\epsilon_{1t}, ..., \epsilon_{J+1,t}, \lambda_t')$ is stationary, we can also assure that the demeaned SC estimator is asymptotically more efficient than DID. Note that stationarity of the conditional process of λ_t implies that both the demeaned SC and the DID estimators are asymptotically unbiased.

Assumption 4″ (stationarity) The process $z_t = (\epsilon_{1t}, ..., \epsilon_{J+1,t}, \lambda_t')$, conditional on $D(1, T_0) = 1$, is weakly stationary stationary and second-order ergodic for $t = 1, ..., T$.

Proposition 3 Under assumptions 1, 2 and 4″, the demeaned SC estimator ($\hat{\alpha}_{1t}^{\text{SC}'}$) is more efficient than the DID estimator ($\hat{\alpha}_{1t}^{DID}$).

Proof.

See details in Appendix A.1.3 ∎

The intuition of this result is the following. For any $t > T_0$, we have that:

$$a.var(\hat{\alpha}_{1t}^{\text{SC}'} - \alpha_{1t}) = E\left[\left(\epsilon_{1t} - \sum_{j\neq 1} \bar{w}_j^{\text{SC}'} \epsilon_{jt}\right) + \tilde{\lambda}_t \left(\tilde{\mu}_1 - \sum_{j\neq 1} \bar{w}_j^{\text{SC}'} \tilde{\mu}_j\right) |D(1, T_0) = 1\right]^2 \tag{9}$$

while:

$$a.var(\hat{\alpha}_{1t}^{\text{DID}} - \alpha_{1t}) = E\left[\left(\epsilon_{1t} - \sum_{j\neq 1} \frac{1}{J} \epsilon_{jt}\right) + \tilde{\lambda}_t \left(\tilde{\mu}_1 - \sum_{j\neq 1} \frac{1}{J} \tilde{\mu}_j\right) |D(1, T_0) = 1\right]^2 \tag{10}$$

where $\tilde{\lambda}_t$ and $\tilde{\mu}_j$ exclude the time-invariant common factor if there is one. We show in Appendix A.1.3 that the demeaned SC weights converge to weights that minimize a function $\Gamma(w)$ such that $\Gamma(w_j^{\text{SC}'}) = a.var(\hat{\alpha}_{1t}^{\text{SC}'} - \alpha_{1t})$ and $\Gamma(\{\frac{1}{J}, ..., \frac{1}{J}\}) = a.var(\hat{\alpha}_{1t}^{\text{DID}} - \alpha_{1t})$. Therefore, it must be that the variance of the demeaned SC estimator is weakly lower than the variance of the DID estimator. Notice that this result relies on stationarity of the common factors for both pre- and post-intervention periods. Under assumption 4′, if we have that $var(\lambda_t) \neq \Omega_0$ for $t > T_0$, then it would not be possible to guarantee that the demeaned SC estimator is more efficient than DID, even if both estimators are asymptotically unbiased.

If treatment assignment is correlated with time-varying common factors, then both the demeaned SC and the DID estimators will be asymptotically biased. In general, it is not possible to rank these two estimators in terms of their bias. We provide in Appendix A.2 an example in which the DID bias can be smaller than the bias of the SC. This might happen when selection into treatment depends on common factors with low variance. We show in Section 6 a particular class of linear factor models in which the asymptotic bias of the demeaned SC estimator will always be lower.

In addition to including an intercept, Doudchenko and Imbens (2016) also consider the possibility of relaxing the non-negative and the adding-up constraints in the SC model. We show in Appendix A.4.3 that our main result that the SC estimator will be asymptotically biased if there is selection on time-varying unobservables still apply if we relax these conditions.[23] Notice that the panel data approach suggested in Hsiao et al. (2012) is essentially the same as the SC estimator using all outcome lags as economic predictor and relaxing the no-intercept, adding-up, and non-negativity constraints. Therefore, our result on asymptotic bias is also valid for the Hsiao et al. (2012) estimator. Note also that relaxing the adding-up constraint implies that the SC estimator may be biased if the time effect δ_t is correlated with the treatment assignment.

We also present in Appendix A.4.4 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 main idea is that, under these assumptions, one could use the lag outcome of the control units as instrumental variables to estimate parameters that reconstruct the factor loadings of the treated unit.

5 Model with "explosive" common factors

We consider now the case in which the first and second moments of a subset of the common factors diverge. Consider first a model with $I(1)$ and $I(0)$ factors:

$$\begin{cases} y_{it}(0) = \lambda_t \mu_i + \gamma_t \theta_i + \epsilon_{it} \\ y_{it}(1) = \alpha_{it} + y_{it}(0) \end{cases} \tag{11}$$

[23] In this case, since we do not constraint the weights to sum 1, we need to adjust assumption 4′ so that it also includes convergence of the pre-treatment averages of the first and second moments of δ_t.

where λ_t is a $(1 \times F_0)$ vector of $I(0)$ common factors, and γ_t is a $(1 \times F_1)$ vector of $I(1)$ common factors, while μ_i and θ_i are the vectors of factor loadings associated with these common factors. Note that the time effect δ_t can be either included in vector λ_t or γ_t.

We modify assumption 4$'$ to state that the pre-treatment processes λ_t and γ_t remain, respectively, $I(0)$ and $I(1)$ even conditional on $D(1, T_0) = 1$. We also assume that ϵ_{jt} is $I(0)$, which will allow for the possibility of cointegration.

Assumption 4$'''$ (stochastic processes) Conditional on $D(1, T_0) = 1$, the processes λ_t and ϵ_{jt} are $I(0)$ while the processes γ_t is $I(1)$ in the pre-treatment periods.

We also modify assumption 3 to state that there are weights that reconstruct the factor loadings of unit 1 associated with the $I(1)$ common factors.

Assumption 3$'$ (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 3$'$ implies that $\Phi_1 \neq \varnothing$.

Note that, in this setting with $I(1)$ common factors, assumption 3$'$ implies that the vector of outcomes $\mathbf{y}_t = (y_{1t}, ..., y_{J+1,t})'$ is co-integrated. Importantly, differently from our results in Session 3, assumption 3$'$ is key for our results.[24] Note, however, that 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. .

Proposition 4 Under assumptions 1, 2, 3$'$, and 4$'''$, we have that:

- In a model with no-intercept: $\hat{\alpha}_{1t} \xrightarrow{d} \alpha_{1t} + \left(\epsilon_{1t} - \sum_{j\neq 1} \bar{w}_j \epsilon_{jt}\right) + \lambda_t \left(\mu_1 - \sum_{j\neq 1} \bar{w}_j \mu_j\right)$

[24]See Carvalho et al. (2016) for the case of construction of artificial counterfactuals when data is $I(1)$ and there is no cointegration relation.

- In a model with intercept: $\hat{\alpha}_{1t} \xrightarrow{d} \alpha_{1t} + \left(\epsilon_{1t} - \sum_{j\neq 1} \bar{w}_j \epsilon_{jt}\right) + (\lambda_t - \omega_0) \left(\mu_1 - \sum_{j\neq 1} \bar{w}_j \mu_j\right)$

where $\mu_1 \neq \sum_{j\neq 1} \bar{w}_j \mu_j$, unless $\sigma^2_\epsilon = 0$ 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. ■

The intuition of this result is that the weights will converge in probability to $\bar{\mathbf{w}} \in \Phi_1$ that minimizes the second moment of the $I(0)$ process $u_t = y_{1t} - \sum_{j\neq 1} w_j y_{jt} = \lambda_t(\mu_1 - \sum_{j\neq 1} w_j \mu_j) + (\epsilon_{1t} - \sum_{j\neq 1} w_j \epsilon_{jt})$.[25] Following the same arguments as in Proposition 1, $\bar{\mathbf{w}}$ will not eliminate the $I(0)$ common factors, unless we have that $\sigma^2_\epsilon = 0$ or it coincides that there is a $\mathbf{w} \in \Phi$ that also minimizes the linear combination of transitory shocks.

Proposition 4 has two important implications. First, if outcomes are indeed cointegrated (that is, assumption $3'$ is valid), then correlation between treatment assignment and $I(1)$ common factors will not generate bias in the SC control and related estimators. However, these estimators may be biased if there is correlation between treatment assignment and the $I(0)$ common factors. The SC estimator (which includes the no-intercept, adding-up, and non-negative constraints) will be asymptotically biased if $\mu^1_1 \neq \sum_{j\neq 1} \bar{w}_j \mu^1_j$ (that is, the weighted average of the control units does not reconstruct the time invariant unobserved variables) and/or if treatment assignment is correlated with time-varying $I(0)$ common factors.[26]

We also consider the case in which γ_t is a deterministic polynomial trend of order F_1 instead of being $I(1)$ processes. We modify assumption $4'''$.

Assumption $4''''$ (stochastic processes) Conditional on $D(1, T_0) = 1$, the processes λ_t and ϵ_{jt} are $I(0)$ in the pre-treatment periods while the $\gamma_t = (t, t^2, ..., t^{F_1})$.

Proposition 5 Under assumptions 1, 2, $3'$, and $4''''$, we have that:

- In a model with no-intercept: $\hat{\alpha}_{1t} \xrightarrow{d} \alpha_{1t} + \left(\epsilon_{1t} - \sum_{j\neq 1} \bar{w}_j \epsilon_{jt}\right) + \lambda_t \left(\mu_1 - \sum_{j\neq 1} \bar{w}_j \mu_j\right)$

[25] This is the case for the model with no intercept. For the model with intercept, weights will converge to β and $\mathbf{w} \in \Phi_1$ that minimize the variance of $u_t = y_{1t} - \beta - \sum_{j\neq 1} w_j y_{jt}$. See Proposition 19.3 in Hamilton (1994) for the case without constraints. In Appendix A.1.4 we show that this result is also valid for any combination of the constraints considered in the SC method.

[26] Relaxing the no-intercept constraint implies in an estimator that is asymptotically unbiased provided that treatment assignment is uncorrelated with time-varying $I(0)$ common factors, although treatment assignment may still be correlated with δ_t (whether it is an $I(0)$ or $I(1)$ common factor). Relaxing the adding-up constraint makes the estimator biased if δ_t is correlated with treatment assignment and 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 these conclusions.

- In a model with intercept: $\hat{\alpha}_{1t} \overset{d}{\to} \alpha_{1t} + \left(\epsilon_{1t} - \sum_{j\neq 1} \bar{w}_j \epsilon_{jt}\right) + (\lambda_t - \omega_0)\left(\mu_1 - \sum_{j\neq 1} \bar{w}_j \mu_j\right)$

where $\mu_1 \neq \sum_{j\neq 1} \bar{w}_j \mu_j$, unless $\sigma^2_\epsilon = 0$ 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.1.5. ■

Therefore, as in the case with $I(1)$ common factors, when $T_0 \to \infty$ the SC method will construct a SC unit that follows exactly the same polynomial trend as the treated unit. However, the SC estimator will also be, in general, asymptotically biased if treatment assignment is correlated with the stationary common factors, λ_t.

An important feature of these settings is that, as $T_0 \to \infty$, the pre-treatment fit will become close to perfect, which is the case in which Abadie et al. (2010) recommend that the SC method should be used. As a measure of goodness of pre-treatment fit, we consider a pre-treatment normalized mean squared error index, as suggested in Ferman et al. (2016):

$$\widetilde{R}^2 = 1 - \frac{\frac{1}{T_0}\sum_{t=1}^{T_0}\left(y_{1t} - \widehat{y}_{1t}\right)^2}{\frac{1}{T_0}\sum_{t=1}^{T_0}\left(y_{1t} - \overline{y}_1\right)^2} \tag{12}$$

where $\widehat{y}_{1t}$ is the outcome of the SC unit and $\overline{y}_1 = \frac{\sum_{t=1}^{T_0} y_{1t}}{T_0}$. This measure is always lower than one, and it is close to one when the pre-treatment fit is good. If this measure is equal to one, then we have a perfect fit.[27] Note that, in both cases analyzed in this session, the numerator will converge to the variance of an $I(0)$ process, while the denominator will diverge as $T_0 \to \infty$. Therefore, in these cases, we show that the SC estimator can be asymptotically biased *even conditional on a close-to-perfect pre-treatment fit.*[28]

Our results show that, in a setting with non-stationary trends, a seemingly perfect pre-treatment fit might hide important possibilities for asymptotic bias in the SC method. While this perfect pre-treatment fit would be indicative that the SC estimator was able to eliminate potential bias coming from correlations between treatment assignment and non-stationary common factors, this would not guarantee unbiasedness if

[27] Differently from the R^2 measure, this measure can be negative, which would suggest a poor pre-treatment fit.

[28] Note that, in their proof, Abadie et al. (2010) assume that there exists a constant $\bar{\lambda}$ such that $|\lambda_t^f| \leq \bar{\lambda}$ for all $t = 1, ..., T$ and $f = 1, ..., F$, where $\lambda_t = (\lambda_t^1, ..., \lambda_t^F)$ is the vector of common factors. Under this and other additional assumptions, they show that the bias of the SC estimator can be bounded by a function that depends on $\bar{\lambda}$ and T_0 if we have a perfect match in the pre-treatment outcoms. In order to guarantee that this function goes to zero when T_0 increases, however, we need to assume that the condition on $\bar{\lambda}$ remains valid when T_0 increases. This will not be the case if some components of λ_t increase without bound when T_0 increases. Therefore, our result does not contradicts the result from Abadie et al. (2010) on the bias of the SC estimator.

there is a correlation between treatment assignment and common factors beyond such non-stationary trends. Therefore, we recommend that researchers should also present the pre-treatment match after eliminating non-stationary trends as a diagnosis test for the SC estimator. To illustrate this point, we consider the three influential applications presented in Abadie and Gardeazabal (2003), Abadie et al. (2010) and Abadie et al. (2015). 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 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. The pre-treatment fit in this case is seemingly perfect, with an $\tilde{R}^2 = 0.99$. However, the per capita GDP series is clearly non-stationary, with all regions displaying similar trends before the intervention. Therefore, based on our results presented in Proposition 4, despite the seemingly perfect pre-treatment fit, it may still be that the SC estimator is biased if there is a correlation between treatment assignment and common factors beyond this non-stationary trend.

In order to assess this possibility, we consider two different ways to de-trend the data, so we can have a better assessment on whether factor loadings associated with stationary common factors are well matched. In both cases, we subtract the outcome of the treated and control units by constant terms $\{a_t\}_{t=1}^{T}$. Note that, under the adding-up constraint ($\sum_{j\neq 1} w_j = 1$), the SC weights and the SC estimator will be numerically the same whether we estimate with the original data or with $\tilde{y}_{jt} = y_{jt} - a_t$. We first subtract the average of the control units at time t ($a_t = \frac{1}{J}\sum_{j\neq 1} y_{jt}$) for both treated and control units. Therefore, 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, and the pre-treatment fit for this de-trended series would not be as good as in the previous case, with an $\tilde{R}^2 = 0.65$. We get similar results if we de-trend by fitting a polynomial $a(t)$ to the synthetic control series, with an $\tilde{R}^2 = 0.67$ (Figure 1.D).[29]

We consider in Figure 2 the application in Abadie et al. (2010), who estimate the effects of California's tobacco control program. This empirical application also presents a seemingly perfect pre-treatment fit, with an $\tilde{R}^2 = 0.96$, but with a highly non-stationary trend. Our first strategy to de-trend the data by subtracting the controls' average outcomes still leads to a non-stationary series, suggesting that the non-

[29]We used a polynomial of order 5 to fit the entire time series of the synthetic control unit (including both pre- and post-periods. Then we consider the de-trended series $\tilde{y}_{jt} = y_{jt} - \hat{a}(t)$.

stationary common factors cannot be resumed into a simple time effect δ_t. When we consider a polynomial $a(t)$, then the pre-treatment fit for the de-trended series is very low. However, note also that there is not much variation in the de-trended series in the pre treatment relative to the difference in the treated and synthetic control units in the post treatment, which suggests that most of the common variation that the SC estimator aims to control for comes from these non-stationary trends. Therefore, such low $\tilde{R}^2$ should not necessarily be interpreted as relevant possibilities for asymptotic bias in the SC estimator. Finally, we consider in Figure 3 the study on the economic impact of the 1990 German reunification on West Germany, by Abadie et al. (2015). Again, this application displays a seemingly perfect pre-treatment fit ($\tilde{R}^2 = 0.99$), but a more modest pre-treatment fit when we de-trend the data using a time polynomial ($\tilde{R}^2 = 0.70$).

Overall, these results suggest that, in these applications, the SC estimator probably worked reasonably well in constructing a counterfactual for the treated unit, as either the pre-treatment fit is reasonably good even after we de-trend the series (although not as good as when we consider the original series), or there is not much variation left in the de-trended series. However, our results point out that the diagnosis based on the pre-treatment fit for non-stationary series should be considered with caution, as they may hide discrepancies in common factors beyond these non-stationary trends that may lead to asymptotic bias in the SC estimator. Indeed, in two of the three applications we considered, there is still some significant variation beyond the non-stationary trends for the treated unit that is only partially captured by the SC unit.

Importantly, note that 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, in these three applications, 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. The only caveat is that measures of pre-intervention fit could be misleading as diagnostic tests, as they may hide important discrepancies in the factor loadings associated to the stationary common factors.

6 Particular Class of Linear Factor Models & Monte Carlo Simulations

We consider now in detail the implications of our results for a particular class of linear factor models in which all units are divided into groups that follow different times trends.[30] We present both theoretical and MC simulation results for these models. In Section 6.1 we consider the case with stationary common factors, while in Section 6.2 we consider the case in which there are both $I(1)$ and $I(0)$ common factors.

6.1 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:

$$y_{jt}(0) = \delta_t + \lambda_t^k + \epsilon_{jt} \tag{13}$$

for some $k = 1, ..., K$. We start considering the case in which $\frac{1}{T_0}\sum_{t=1}^{T_0} \lambda_t^k \overset{p}{\to} 0$, $\frac{1}{T_0}\sum_{t=1}^{T_0} (\lambda_t^k)^2 \overset{p}{\to} 1$, $\frac{1}{T_0}\sum_{t=1}^{T_0} \epsilon_{jt} \overset{p}{\to} 0$, and $\frac{1}{T_0}\sum_{t=1}^{T_0} \epsilon_{jt}^2 \overset{p}{\to} \sigma_\epsilon^2$.

6.1.1 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, the SC estimator should 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 economic predictors. From the minimization problem 3, we have that, when $T_0 \to \infty$, the proportion of misallocated weights converges to:

$$\gamma_2(\sigma_\epsilon^2, J) = \sum_{j=\frac{J+1}{2}+1}^{J+1} \bar{w}_j = \frac{J+1}{J^2 + 2 \times J \times \sigma_\epsilon^2 - 1} \times \sigma_\epsilon^2 \tag{14}$$

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

[30] Monte Carlo simulations using this model was also studied in detail in Ferman et al. (2016) and in Ferman and Pinto (2017).

We present in Figure 4.A the relationship between asymptotic misallocation of weights, variance of the transitory shocks, and number of control units. Note that, for a fixed J, the proportion of misallocated weights converges to zero when $\sigma^2_\epsilon \to 0$, while this proportion converges to $\frac{J+1}{2J}$ (the proportion of misallocated weights of DID) when $\sigma^2_\epsilon \to \infty$. This is consistent with the results we have in Section 3. Moreover, note that, for a given σ^2_ϵ, 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 in Abadie et al. (2010) for unbiasedness are satisfied.

Note that, in this example, the SC estimator converges to:

$$\hat{\alpha}_{1t} \xrightarrow{d} \alpha_{1t} + \left(\epsilon_{1t} - \sum_{j \neq 1} \bar{w}_j \epsilon_{jt} \right) + \lambda^1_t \times \gamma_2(\sigma^2_\epsilon, J) - \lambda^2_t \times \gamma_2(\sigma^2_\epsilon, J) \tag{15}$$

Therefore, if $E[\lambda^1_t | D(1, T_0) = 1] = 1$ for $t > T_0$ (that is, the expected value of the common factor associated to the treated unit is one standard deviation higher), then the bias of the SC estimators in terms of the standard deviation of y_{1t} would be given by $\frac{\gamma_2(\sigma^2_\epsilon, J)}{\sqrt{1+\sigma^2_\epsilon}}$. Therefore, while a higher σ^2_ϵ increases the misallocation of weights, the importance of this misallocation in terms of bias of the SC estimator is limited by the fact that the common factor (which we allow to be correlated with treatment assignment) becomes less relevant. We present the relative asymptotic bias of the SC estimator as a function of σ^2_ϵ and J in Figure 4.B. Note that, if $J + 1 \geq 20$, then the bias of the SC estimator will always be lower than 0.1 standard deviations of y_{1t} when treatment assignment is associated with a one standard deviation increase in λ^1_t. This happens because, in this model, the misallocation of weights diminishes when the number of control groups increases.

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 other words, 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:

$$\gamma_{\frac{J+1}{2}}(\sigma^2_\epsilon, J) = \sum_{j=3}^{J+1} \bar{w}_j = \frac{J - 1}{2 + \sigma^2_\epsilon + (1 + \sigma^2_\epsilon)(J - 1)} \times \sigma^2_\epsilon \tag{16}$$

We present the relationship between misallocation of weights, variance of the transitory shocks, and number of control units in Figure 4.C. Note that, again, the proportion of misallocated weights converges to zero when $\sigma^2_\epsilon \rightarrow 0$ and to the proportion of misallocated weights of DID when $\sigma^2_\epsilon \rightarrow \infty$ (in this case, $\frac{J-1}{J}$). Differently from the previous case, however, for a given σ^2_ϵ, the proportion of misallocated weights converges to $\frac{\sigma^2_\epsilon}{1+\sigma^2_\epsilon}$ when $J \rightarrow \infty$. 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 in Gobillon and Magnac (2016) are not satisfied. As presented in Figure 4.D, in this case, the asymptotic bias can be substantially higher, and it does not vanishes when the number of control units increases. Therefore, the asymptotic bias of the SC estimator can be relevant even when the number of control units increases.

Finally, note that, 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. However, this is not a general result, as we show in Appendix A.2.

6.1.2 Monte Carlo Simulations

The results presented in Section 6.1.1 are based on large-T_0 asymptotics. We now consider, in MC simulations, the finite T_0 properties of the SC estimator, both unconditional and conditional on a good pre-treatment fit. We present Monte Carlo (MC) simulation results using a data generating process (DGP) based on equation 13. We consider in our MC simulations $J+1=20$, λ^k_t normally distributed following an AR(1) process with 0.5 serial correlation parameter, $\epsilon_{jt} \sim N(0, \sigma^2_\epsilon)$, 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 $t \in \{1, ..., T\}$. 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: $\sigma^2_\epsilon \in \{0.1, 0.5, 1\}$.
- The number of groups with different λ^k_t: $K = 2$ (2 groups of 10) or $K = 10$ (10 groups of 2)

For each simulation, we calculate the SC estimator that uses all pre-treatment outcome lags as economic predictors, and calculate the proportion of misallocated weights. We also evaluate whether the SC method provides a good pre-intervention fit and calculate the proportion of misallocated weights conditional on a good pre-intervention fit. In order to determine that the SC estimator provided a good fit, we consider a pre-treatment normalized mean squared error index, presented in equation 12. For each scenario, we generate 20,000 simulations.

In columns 1 to 3 of Table 1, we present the proportion of misallocated weights when $K = 10$ for different values of T_0 and σ^2_ϵ. Consistent with our analytical results from Section 6.1.1, the misallocation of weights is increasing with the variance of the transitory shocks. With $T_0 = 100$, the proportion of misallocated weights is close to the theoretical values, while the proportion of misallocated weights is substantially higher when T_0 is small. We present in columns 4 to 6 of Table 1 the probability that the SC method provides a good fit when we define good fit as $\widetilde{R}^2 > 0.8$. As expected, with a large T_0 the SC method only provides a good pre-intervention fit if the variance of the transitory shock is low. If the variance of the transitory shocks is higher, then the probability that the SC method provides a good match is approximately zero, unless the number of pre-treatment periods is rather low. These results suggest that, in a model with stationary factors, the SC estimator would only provide a close-to-perfect pre-treatment fit with a moderate number of pre-treatment periods if the variance of the transitory shocks is low, in which case the bias of the SC estimator would be relatively small. With $T_0 = 20$ and $\sigma^2_\epsilon = 0.5$ or $\sigma^2_\epsilon = 1$, the probability of having a good fit is, respectively, equal to 1.3% and 0.1%. Interestingly, when we condition on having a good pre-treatment fit the proportion of misallocated weights reduces but still remains quite high (goes from 50% to 33% when $\sigma^2_\epsilon = 0.5$ and from 65% to 45% when $\sigma^2_\epsilon = 1$). These results are presented in Table 1, columns 7 to 9. In Appendix Table A.1 we replicate Table 1 using a more stringent definition of good fit, which is equal to one if $\widetilde{R}^2 > 0.9$. In this case, conditioning has a larger effect in reducing the discrepancy of factor loadings between the treated and the SC units, but at the expense of having a lower probability of accepting that the pre-treatment fit is good. These results suggest that, with stationary data, the SC estimator would only provide a close-to-perfect match with a moderate T_0, and therefore be close to unbiased, when the variance of the transitory shocks converges to zero. In Appendix Table A.2 we also consider the case with 2 groups of 10 units each ($K = 2$). All results are qualitatively the same.

Note that, 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. We present in Table 2 the DID/SC ratio of standard errors. With $T_0 = 100$, the DID standard error is 2.6 times higher than the SC standard errors when $\sigma_\epsilon^2 = 0.1$. When σ_ϵ^2 is higher, the advantage of the SC estimator is reduced, although the DID standard error is still 1.4 (1.2) 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 $\sigma_\epsilon^2 \to \infty$. 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. These results suggest 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 of this result 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.

6.2 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:

$$y_{jt}(0) = \delta_t + \lambda_t^k + \gamma_t^r + \epsilon_{jt} \tag{17}$$

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.

6.2.1 Asymptotic results

Based on our results from Section 5 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.

6.2.2 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. For the same reason, the probability of having a dataset with a close-to-perfect pre-treatment fit is also very high if a subset of the common factors is $I(1)$ (columns 7 to 9 of Table 3). Finally, we show in columns 10 to 12 of Table 3 that conditioning on a close-to-perfect match makes virtually no difference in the proportion of misallocated weights for the stationary factor.

These results suggest that the SC method works remarkably well to control for $I(1)$ common factors. In this scenario, one would usually have a close-to-perfect fit, and there would be virtually no bias associated to the $I(1)$ factors. However, we might have a substantial misallocation of weights for the $I(0)$ common factors *even conditional on a close-to-perfect pre-treatment match.* 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 Table 4. 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.

7 Conclusion

In this paper, we revisit the theory behind the SC method. We consider the asymptotic properties of the SC estimator when the number of pre-treatment periods goes to infinity in a linear factors model. If the model has "stationary" common factors, in the sense that pre-treatment averages of the first and second moments of the common factors converge, then we show that the SC estimator is asymptotically biased if treatment assignment is correlated with unobserved confounders, even when weights that reconstruct the factor loadings of the treated unit exist and when $T_0 \to \infty$. The asymptotic bias goes to zero when the variance of the transitory shocks goes to zero, which is exactly the case in which one would expect to find a good pre-treatment fit. Therefore, our results, under these conditions on the common factors, are consistent with the results in Abadie et al. (2010). However, if pre-treatment averages of a subset of the common factors diverge, then we show that the SC estimator can be asymptotically biased even conditional on a close-to-perfect pre-treatment match.

Despite these caveats, we show that a demeaned SC estimator can substantially improve relative to the DID estimator, even if the pre-treatment fit is not close to perfect and if T_0 is not large. This is particularly true when a subset of the common factors is non-stationary, as it allows treatment assignment to be correlated with common factors that diverge. However, our results show that researchers should be more careful in interpreting the identification assumptions required for the SC method. 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 assess whether there might be relevant possibilities for bias arising due to a correlation between treatment assignment and common factors beyond non-stationary trends. Finally, our results also have implications for the placebo test suggested in Abadie et al. (2010), as Ferman and Pinto (2017) show building on our results.

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.

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

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

Borwein, Jonathan and Jon Vanderwerff, "Epigraphical and uniform convergence of convex functions," *Transactions of the American Mathematical Society*, 1996, *348* (4), 1617–1631.

Carvalho, Carlos V., Ricardo Mansini, and Marcelo C. Medeiros, "ArCo: An Artificial Counterfactual Approach for Aggregate Data," February 2015. Working Paper.

_ , _ , and _ , "The Perils of Counterfactual Analysis with Integrated Processes," December 2016. Working Paper.

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

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

Ferman, Bruno and Cristine Pinto, "Placebo Tests for Synthetic Controls," MPRA Paper 78079, University Library of Munich, Germany April 2017.

_ , _ , and Vitor Possebom, "Cherry Picking with Synthetic Controls," 2016. Working Paper.

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.

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

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.

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

Phillips, P. C. B., "Optimal Inference in Cointegrated Systems," *Econometrica*, 1991, *59* (2), 283–306.

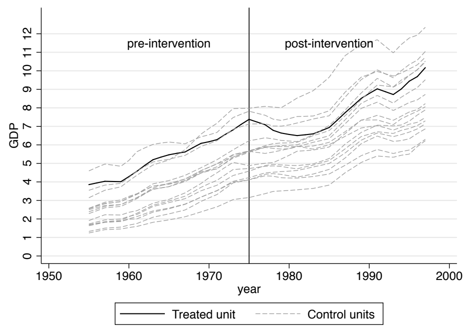
Powell, David, "Imperfect Synthetic Controls: Did the Massachusetts Health Care Reform Save Lives?," 2017.

Wong, Laurence, "Three Essays in Causal Inference." PhD dissertation, Stanford University March 2015.

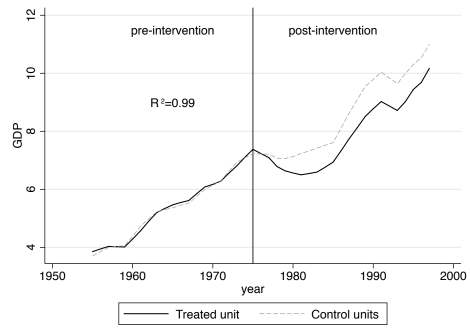
Xu, Yiqing, "Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models," *Political Analysis*, 2017, *25* (1), 57?76.

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

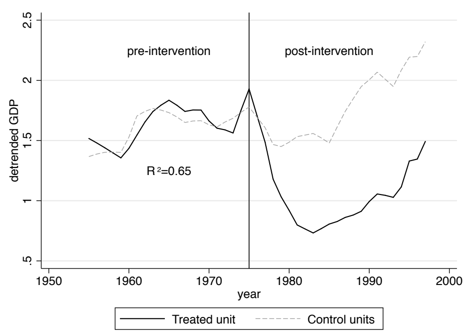
1.A: Raw data



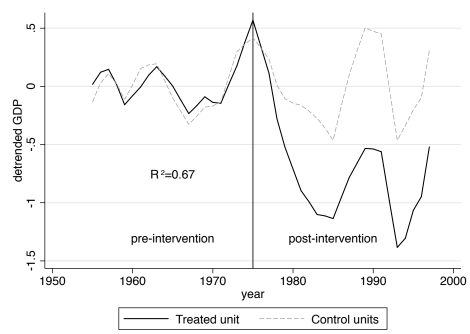
1.B: GDP in level



1.C: de-trended with controls' average



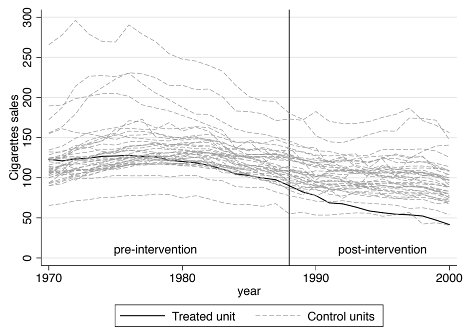
1.D: de-trended with time polynomial



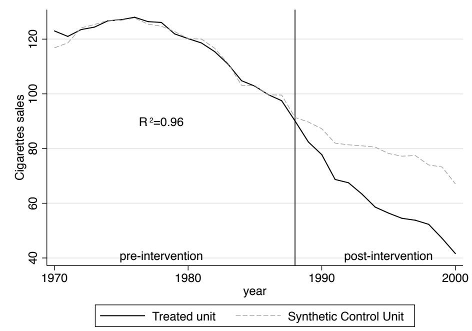
Notes: Figure A presents time series for the treated and for the control units used in the empirical application in Abadie and Gardeazabal (2003). In Figure B we present the time series for the treated and for the SC units. In Figure C we present the same information as in Figure B after subtracting the control groups' averages for each time period. In Figure C we present the same information as in Figure B after subtracting a time trend estimated by fitting a 5th order polynomial on the SC series. Figures B to D we also report the measure of pre-treatment fit defined in equation 12.

Figure 2: **Abadie et al. (2010) application**

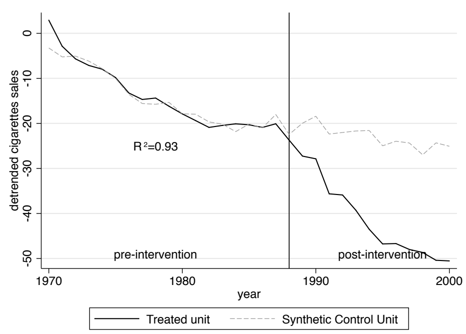
2.A: Raw data



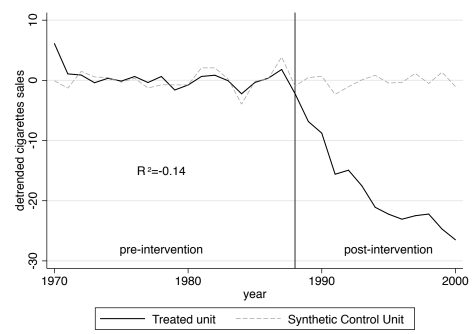
2.B: Outcome in level



2.C: de-trended with controls' average

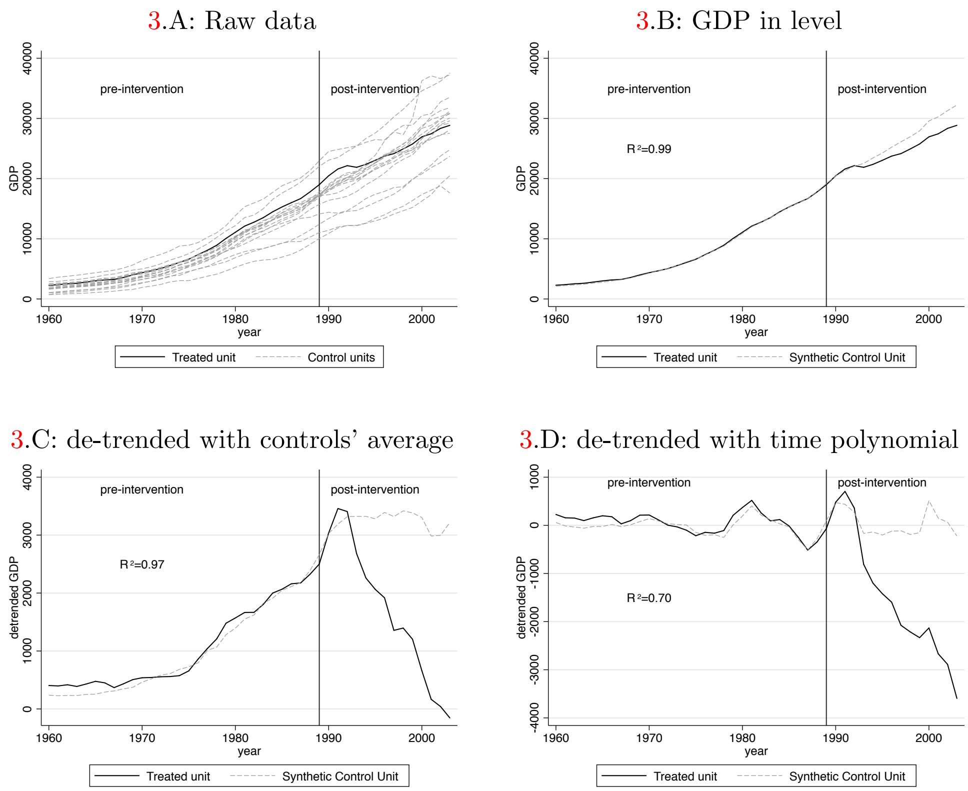


2.D: de-trended with time polynomial



Notes: Figure A presents time series for the treated and for the control units used in the empirical application in Abadie et al. (2010). 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. In Figure C we present the same information as in Figure B after subtracting a time trend estimated by fitting a 5th order polynomial on the SC series. Figures B to D we also report the measure of pre-treatment fit defined in equation 12.

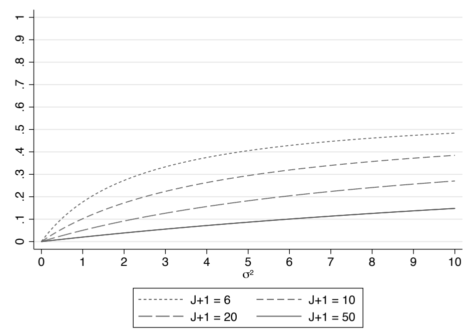
Figure 3: **Abadie et al. (2015) application**



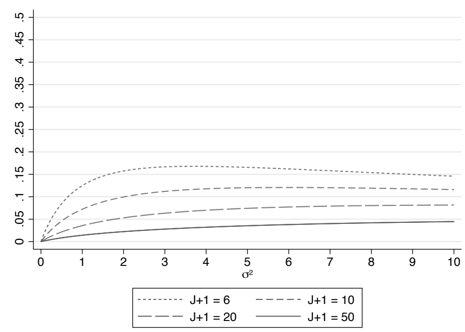
Notes: Figure A presents time series for the treated and for the control units used in the empirical application in Abadie et al. (2015). In Figure B we present the time series for the treated and for the SC units. In Figure C we present the same information as in Figure B after subtracting the control groups' averages for each time period. In Figure C we present the same information as in Figure B after subtracting a time trend estimated by fitting a 5th order polynomial on the SC series. Figures B to D we also report the measure of pre-treatment fit defined in equation 12.

Figure 4: **Asymptotic Misallocation of Weights and Bias**

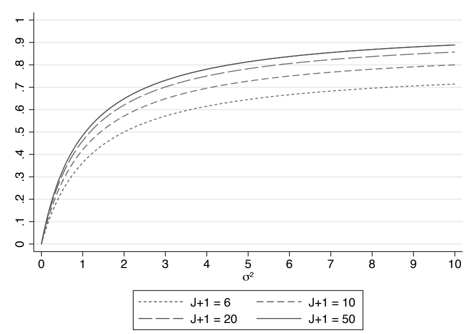
4.A: Misallocation of weights - 2 groups



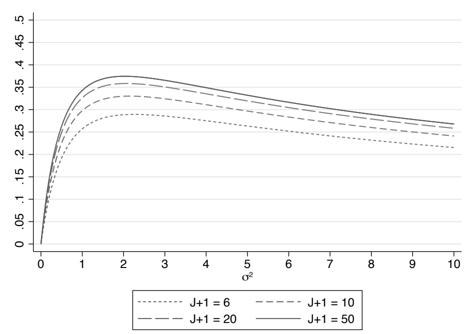
4.B: Bias - 2 groups



4.C: Misallocation of weights - $\frac{J+1}{2}$ groups



4.D: Bias - $\frac{J+1}{2}$ groups



Notes: these figures present the asymptotic misallocation of weights and bias of the SC estimator as a function of the variance of the transitory shocks for different numbers of control units. Figures 4.A and 4.B report results when there are 2 groups of $\frac{J+1}{2}$ units each, while figures 4.C and 4.D report 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. The bias of the SC estimator is reported in terms of standard deviations of y_{jt} (which is equal to $\sqrt{1+\sigma_\epsilon^2}$) when the expected value of the common factor associated to the treated unit, conditional on this unit being treated, is equal to one standard deviation of the common factor.

Table 1: **Misallocation of weights and probability of perfect match - stationary model**

	Misallocation of weights			Probability of perfect match ($\widetilde{R}^2 > 0.8$)			Misallocation conditional on perfect match		
	$\sigma_\epsilon^2 = 0.1$	$\sigma_\epsilon^2 = 0.5$	$\sigma_\epsilon^2 = 1$	$\sigma_\epsilon^2 = 0.1$	$\sigma_\epsilon^2 = 0.5$	$\sigma_\epsilon^2 = 1$	$\sigma_\epsilon^2 = 0.1$	$\sigma_\epsilon^2 = 0.5$	$\sigma_\epsilon^2 = 1$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$T_0 = 5$	0.418	0.714	0.807	0.729	0.510	0.469	0.425	0.743	0.833
	[0.002]	[0.002]	[0.002]	[0.003]	[0.004]	[0.004]	[0.003]	[0.003]	[0.002]
$T_0 = 20$	0.197	0.495	0.653	0.639	0.013	0.001	0.174	0.331	0.445
	[0.001]	[0.001]	[0.001]	[0.003]	[0.001]	[0.000]	[0.001]	[0.008]	[0.040]
$T_0 = 50$	0.150	0.415	0.573	0.701	0.000	0.000	0.137	-	-
	[0.000]	[0.001]	[0.001]	[0.003]	[0.000]	[0.000]	[0.000]	-	-
$T_0 = 100$	0.130	0.384	0.539	0.766	0.000	0.000	0.122	-	-
	[0.000]	[0.001]	[0.001]	[0.003]	[0.000]	[0.000]	[0.000]	-	-

Notes: this table presents MC simulations results from a stationary model. We consider the SC estimator that uses all pre-treatment outcome lags as economic predictors for a given (T_0, σ_ϵ^2). 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_t^k. 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 probability that the pre-treatment match is close to perfect, defined as a $\widetilde{R}^2 > 0.8$. Columns 7 to 9 present the proportion of misallocated weights conditional on a perfect match.

Table 2: **DID/SC ratio of standard errors - stationary model**

	$\sigma_\epsilon^2 = 0.1$	$\sigma_\epsilon^2 = 0.5$	$\sigma_\epsilon^2 = 1$
	(1)	(2)	(3)
$T_0 = 5$	1.719	1.150	1.049
	[0.012]	[0.007]	[0.006]
$T_0 = 20$	2.425	1.306	1.125
	[0.014]	[0.007]	[0.005]
$T_0 = 50$	2.548	1.382	1.158
	[0.017]	[0.008]	[0.005]
$T_0 = 100$	2.607	1.404	1.175
	[0.018]	[0.008]	[0.006]

Notes: this table presents MC simulations results from a stationary model as in Table 1. We present the ratio of standard errors of the DID estimator vs. the SC estimator for different (T_0, σ_ϵ^2) scenarios.

Table 3: **Misallocation of weights and probability of perfect match - non-stationary model**

	Misallocation of weights			Misallocation of weights (non-stationary factors)		
	$\sigma^2_\epsilon = 0.1$	$\sigma^2_\epsilon = 0.5$	$\sigma^2_\epsilon = 1$	$\sigma^2_\epsilon = 0.1$	$\sigma^2_\epsilon = 0.5$	$\sigma^2_\epsilon = 1$
	(1)	(2)	(3)	(4)	(5)	(6)
$T_0 = 5$	0.372	0.661	0.762	0.107	0.192	0.232
	[0.002]	[0.002]	[0.002]	[0.001]	[0.002]	[0.002]
$T_0 = 20$	0.176	0.441	0.589	0.029	0.069	0.095
	[0.001]	[0.001]	[0.001]	[0.000]	[0.001]	[0.001]
$T_0 = 50$	0.136	0.373	0.518	0.015	0.036	0.050
	[0.001]	[0.001]	[0.001]	[0.000]	[0.000]	[0.000]
$T_0 = 100$	0.120	0.346	0.489	0.009	0.022	0.030
	[0.000]	[0.001]	[0.001]	[0.000]	[0.000]	[0.000]

	Probability of perfect match ($\widetilde{R}^2 > 0.8$)			Misallocation conditional on perfect match		
	$\sigma^2_\epsilon = 0.1$	$\sigma^2_\epsilon = 0.5$	$\sigma^2_\epsilon = 1$	$\sigma^2_\epsilon = 0.1$	$\sigma^2_\epsilon = 0.5$	$\sigma^2_\epsilon = 1$
	(7)	(8)	(9)	(10)	(11)	(12)
$T_0 = 5$	0.846	0.618	0.542	0.377	0.683	0.784
	[0.003]	[0.003]	[0.004]	[0.002]	[0.003]	[0.003]
$T_0 = 20$	0.984	0.556	0.296	0.175	0.427	0.571
	[0.001]	[0.004]	[0.003]	[0.001]	[0.002]	[0.003]
$T_0 = 50$	1.000	0.835	0.550	0.136	0.371	0.515
	[0.000]	[0.003]	[0.004]	[0.001]	[0.001]	[0.001]
$T_0 = 100$	1.000	0.973	0.822	0.120	0.346	0.487
	[0.000]	[0.001]	[0.003]	[0.000]	[0.001]	[0.001]

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 $(T_0, \sigma^2_\epsilon, K)$. 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^k_t) 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^r_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 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 probability that the pre-treatment match is close to perfect, defined as a $\widetilde{R}^2 > 0.8$. Columns 10 to 12 present the proportion of misallocated weights conditional on a perfect match. Standard errors in brackets.

Table 4: **DID/SC ratio of standard errors - non-stationary model**

	$\sigma_\epsilon^2 = 0.1$	$\sigma_\epsilon^2 = 0.5$	$\sigma_\epsilon^2 = 1$
	(1)	(2)	(3)
$T_0 = 5$	3.469	1.992	1.662
	[0.032]	[0.014]	[0.011]
$T_0 = 20$	8.370	4.004	3.021
	[0.057]	[0.028]	[0.022]
$T_0 = 50$	13.490	6.372	4.747
	[0.086]	[0.045]	[0.026]
$T_0 = 100$	19.595	9.239	6.862
	[0.145]	[0.066]	[0.049]

Notes: this table presents MC simulations results from a non-stationary model as in Table 3. We present the ratio of standard errors of the DID estimator vs. the SC estimator for different (T_0, σ_ϵ^2) scenarios. Standard errors in brackets.

Table A.1: **Misallocation of weights and probability of perfect match - alternative definition of perfect match**

	Misallocation of weights			Probability of perfect match ($\widetilde{R}^2 > 0.9$)			Misallocation conditional on perfect match		
	$\sigma^2_\epsilon = 0.1$	$\sigma^2_\epsilon = 0.5$	$\sigma^2_\epsilon = 1$	$\sigma^2_\epsilon = 0.1$	$\sigma^2_\epsilon = 0.5$	$\sigma^2_\epsilon = 1$	$\sigma^2_\epsilon = 0.1$	$\sigma^2_\epsilon = 0.5$	$\sigma^2_\epsilon = 1$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$T_0 = 5$	0.418	0.714	0.807	0.490	0.319	0.296	0.448	0.771	0.848
	[0.002]	[0.002]	[0.002]	[0.004]	[0.003]	[0.003]	[0.003]	[0.003]	[0.003]
$T_0 = 20$	0.197	0.495	0.653	0.128	0.000	0.000	0.143	-	-
	[0.001]	[0.001]	[0.001]	[0.002]	[0.000]	[0.000]	[0.002]	-	-
$T_0 = 50$	0.150	0.415	0.573	0.032	0.000	0.000	0.102	-	-
	[0.000]	[0.001]	[0.001]	[0.001]	[0.000]	[0.000]	[0.002]	-	-
$T_0 = 100$	0.130	0.384	0.539	0.005	0.000	0.000	0.088	-	-
	[0.000]	[0.001]	[0.001]	[0.000]	[0.000]	[0.000]	[0.003]	-	-

Notes: this table replicates the results from Table 1 using a more stringent definition of perfect match.

Table A.2: **Misallocation of weights and probability of perfect match - stationary model ($K = 2$)**

	Misallocation of weights			Probability of perfect match ($\widetilde{R}^2 > 0.8$)			Misallocation conditional on perfect match		
	$\sigma^2_\epsilon = 0.1$	$\sigma^2_\epsilon = 0.5$	$\sigma^2_\epsilon = 1$	$\sigma^2_\epsilon = 0.1$	$\sigma^2_\epsilon = 0.5$	$\sigma^2_\epsilon = 1$	$\sigma^2_\epsilon = 0.1$	$\sigma^2_\epsilon = 0.5$	$\sigma^2_\epsilon = 1$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$T_0 = 5$	0.092	0.199	0.266	0.842	0.631	0.555	0.086	0.198	0.268
	[0.001]	[0.001]	[0.002]	[0.003]	[0.003]	[0.004]	[0.001]	[0.002]	[0.002]
$T_0 = 20$	0.066	0.140	0.191	0.921	0.167	0.030	0.063	0.100	0.121
	[0.000]	[0.001]	[0.001]	[0.002]	[0.003]	[0.001]	[0.000]	[0.002]	[0.004]
$T_0 = 50$	0.053	0.110	0.155	0.987	0.024	0.000	0.052	0.066	-
	[0.000]	[0.000]	[0.001]	[0.001]	[0.001]	[0.000]	[0.000]	[0.003]	-
$T_0 = 100$	0.044	0.095	0.134	0.999	0.001	0.000	0.044	-	-
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	-	-

Notes: this table replicates the results from Table 1 using a DGP with $K = 2$.

A Supplemental Appendix: Revisiting the Synthetic Control Estimator

A.1 Proof of the Main Results

A.1.1 Proposition 1

Proof. Let $\mathbf{w} \equiv \{w_j\}_{j\neq 1}$ be the vector Jx1 of unknown weights, we consider the vector Jx1 $\widehat{\mathbf{w}} \equiv \{\hat{w}_j\}_{j\neq 1}$ as as the M-estimator that solves the following optimization problem:

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}\in W} \frac{1}{T_0}\sum_{t_0=1}^{T}\left(y_{1t} - \sum_{j\neq 1} w_j y_{jt}\right)^2 = \arg\min_{\mathbf{w}\in W} \widehat{Q}_{T_0}(\mathbf{w})$$

subject to $\mathbf{w} \in W = \{\mathbf{w} \in \mathbb{R}^J | w_j \geq 0 \text{ and } \sum_{j\neq 1} w_j = 1\}$.[31]

Under assumptions 1 and 4′, the objective function converges in probability to:

$$\widehat{Q}_{T_0}(\mathbf{w}) \xrightarrow{p} Q_0(\mathbf{w}) = \sigma_\epsilon^2 + \sigma_\epsilon^2 \sum_{j\neq 1}(w_j)^2 + \left(\mu_1 - \sum_{j\neq 1} w_j \mu_j\right)' \Omega_0 \left(\mu_1 - \sum_{j\neq 1} w_j \mu_j\right) \tag{18}$$

Note that the first element of this expression is a constant, and it does not matter for the optimization problem. Except for the constant, we can represent this objective function using matrices. Define $\mathbf{w}$ as a vector $(J \times 1)$ of the weights, $\{w_j\}_{j\neq 1}$, μ_1 is a vector $(K \times 1)$ with the factor loadings for the treated units and μ_0 is a matrix $(K \times J)$ that contains the factor loadings for the all the control units, we can write this optimization problem as:

$$\arg\min_{\mathbf{w}\in W} \sigma_\epsilon^2 \mathbf{w}'\mathbf{w} + (\mu_1 - \mu_0\mathbf{w})' \Omega_0 (\mu_1 - \mu_0\mathbf{w})$$

where $\mathbf{w} \in W = \{\mathbf{w} \in \mathbb{R}^J | w_j \geq 0 \text{ and } \sum_{j\neq 1} w_j = 1\}$. This is a minimization of a quadratic function in a compact space, and has a unique solution $\mathbf{w}^*$.

Note that $\widehat{Q}_{T_0}$ is a convex function. In addition, $\sup_{\mathbf{w}\in W}\left\|\widehat{Q}_{T_0}(w)\right\| \leq C$.

By Lemma 1.6 of Borwein and Vanderwerff (1996), if $\widehat{Q}_{T_0}$ and Q_0 are continuous convex functions, uniformly bounded on a compact space, and $\widehat{Q}_{T_0}$ converges pointwise to Q_0, then $\widehat{Q}_{T_0}$ converges uniformly to Q_0 on W.

At the end, $\mathbf{w}^*$ is the unique minimum of Q_0, W is a compact space, Q_0 is continuous and $\widehat{Q}_T$ converges uniformly to Q_0. By Theorem 2.1 of Newey and McFadden (1994), $\widehat{\mathbf{w}}$ exists with probability approaching one and $\widehat{\mathbf{w}} \xrightarrow{p} \mathbf{w}^*$.

Now, we need to show that $\mathbf{w}^*$ does not necessary reconstruct the factor loadings. Note that the objective function 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 $\frac{1}{J}$. The second part reflects the presence of common factors λ_t that would remain after we choose the weights to construct the SC unit. Suppose that we start at $\left\{w_j^*\right\}_{j\neq 1}$ such that $\mu_1 = \sum_{j\neq 1} w_j^{*\prime} \mu_j$ and move in the direction of $w_j = \frac{1}{J}$ for all $j = 2, .., J+1$, with $w_j = w_j^* + \Delta(\frac{1}{J} - w_j^*)$. 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:

$$\Gamma'(\{w_j^*\}_{j\neq 1}) = 2\sigma_\epsilon^2\left(\frac{1}{J} - \sum_{j=2}^{J+1}(w_j^*)^2\right) < 0 \text{ unless } w_j^* = \frac{1}{J} \text{ or } \sigma_\epsilon^2 = 0$$

Therefore, $\mathbf{w}^*$ cannot be, in general, a solution of the objective function of the M-estimator. 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, unless it turns

[31] 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.

out that $\mathbf{w}^*$ also minimizes the variance of this linear combination of the idiosyncratic errors or if $\sigma^2 = 0_\epsilon$. ■

A.1.2 Proposition 2

Proof. Note first that the minimization problem 7 is equivalent to:

$$\widehat{\mathbf{w}}^{\mathrm{SC}'} = \underset{a \in \mathbb{R}, \mathbf{w} \in W}{\operatorname{argmin}} \frac{1}{T_0} \sum_{t=1}^{T_0} \left[y_{1t} - \sum_{j \neq 1} w_j y_{jt} - a \right]^2 \tag{19}$$

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

Under assumptions 1, 2 and 4′:

$$\begin{aligned} \frac{1}{T_0} \sum_{t=1}^{T_0} \left[y_{1t} - \sum_{j \neq 1} w_j y_{jt} - a \right]^2 &= \frac{1}{T_0} \sum_{t=1}^{T_0} \left[\left(\epsilon_{1t} - \sum_{j \neq 1} w_j \epsilon_{jt} \right) + \lambda_t \left(\mu_1 - \sum_{j \neq 1} w_j \mu_j \right) - a \right]^2 \\ &\overset{p}{\rightarrow} \sigma_\epsilon^2 \left(1 + \sum_{j \neq 1} (w_j)^2 \right) + \left(\mu_1 - \sum_{j \neq 1} w_j \mu_j \right)' \Omega_0 \left(\mu_1 - \sum_{j \neq 1} w_j \mu_j \right) + \\ &\quad + a^2 - 2 \times \omega_0 \left(\mu_1 - \sum_{j \neq 1} w_j \mu_j \right) \equiv Q(a, \mathbf{w}) \end{aligned} \tag{20}$$

For any $\mathbf{w}$, this objective function is minimized at $a(\mathbf{w}) = \bar{y}_1 - \sum_{j \neq 1} w_j \bar{y}_j$. Since $\mathbf{w} \in W$, where W is a compact space, we can restrict the parameter space $a \in [-K, K]$. Therefore, by Lemma 1.6 of Borwein and Vanderwerff (1996), we have that this convergence is uniform. By Theorem 2.1 of Newey and McFadden (1994), $(\hat{a}^{\mathrm{SC}'}, \widehat{\mathbf{w}}^{\mathrm{SC}'})$ exists with probability approaching one and $(\hat{a}^{\mathrm{SC}'}, \widehat{\mathbf{w}}^{\mathrm{SC}'}) \overset{p}{\rightarrow} (\bar{a}^{\mathrm{SC}'}, \bar{\mathbf{w}}^{\mathrm{SC}'})$ that minimize $Q(a, \mathbf{w})$.

Note that $Q(a, \mathbf{w})$ is minimized at $\bar{a}^{\mathrm{SC}'} = \omega_0 \left(\mu_1 - \sum_{j \neq 1} \bar{w}_j^{\mathrm{SC}'} \mu_j \right)$, where $\bar{\mathbf{w}}^{\mathrm{SC}'} \notin \Phi$ unless $\sigma_\epsilon^2 = 0$ or $\exists \mathbf{w} \in \Phi | \mathbf{w} \in \underset{\mathbf{w} \in W}{\operatorname{argmin}} \left\{ \sigma_\epsilon^2 \left(1 + \sum_{j \neq 1} (w_j)^2 \right) \right\}$, following the same steps as in Proposition 1.

Therefore:

$$\hat{\alpha}_{1t}^{\mathrm{SC}'} = y_{1t} - \sum_{j \neq 1} \hat{w}_j^{\mathrm{SC}'} y_{it} - \left[\bar{y}_1 - \sum_{j \neq 1} \hat{w}_j^{\mathrm{SC}'} \bar{y}_i \right] \overset{d}{\rightarrow} \alpha_{1t} + \left(\epsilon_{1t} - \sum_{j \neq 1} \bar{w}_j^{\mathrm{SC}'} \epsilon_{jt} \right) + (\lambda_t - \omega_0) \left(\mu_1 - \sum_{j \neq 1} \bar{w}_j^{\mathrm{SC}'} \mu_j \right) \tag{21}$$

■

A.1.3 Proposition 3

Proof. From Proposition 2:

$$\hat{\alpha}_{1t}^{\mathrm{SC}'} \overset{d}{\rightarrow} \alpha_{1t} + \left(\epsilon_{1t} - \sum_{j \neq 1} \bar{w}_j^{\mathrm{SC}'} \epsilon_{jt} \right) + (\lambda_t - \omega_0) \left(\mu_1 - \sum_{j \neq 1} \bar{w}_j^{\mathrm{SC}'} \mu_j \right) \tag{22}$$

Under assumption 4″, we have that λ_t conditional on $D(1, T_0) = 1$ is stationary. Therefore, without loss of generality we can assume that the first common factor is time invariant while the other common factors are such that $E[\lambda_t | D(1, T_0) = 1] = 0$ for all t. Therefore:

$$\hat{\alpha}_{1t}^{\mathrm{SC}'} \overset{d}{\rightarrow} \alpha_{1t} + \left(\epsilon_{1t} - \sum_{j \neq 1} \bar{w}_j^{\mathrm{SC}'} \epsilon_{jt} \right) + \tilde{\lambda}_t \left(\tilde{\mu}_1 - \sum_{j \neq 1} \bar{w}_j^{\mathrm{SC}'} \tilde{\mu}_j \right) \tag{23}$$

where $\tilde{\lambda}_t$ and $\tilde{\mu}_j$ exclude the time-invariant common factor. Therefore:

$$a.var(\hat{\alpha}_{1t}^{\text{SC}'} - \alpha_{1t}) = E\left[\left(\epsilon_{1t} - \sum_{j\neq 1} \bar{w}_j^{\text{SC}'} \epsilon_{jt}\right) + \tilde{\lambda}_t \left(\tilde{\mu}_1 - \sum_{j\neq 1} \bar{w}_j^{\text{SC}'} \tilde{\mu}_j\right) |D(1,T_0)=1\right]^2 \tag{24}$$

Similarly:

$$\hat{\alpha}_{1t}^{\text{DID}} \xrightarrow{d} \alpha_{1t} + \left(\epsilon_{1t} - \sum_{j\neq 1} \frac{1}{J}\epsilon_{jt}\right) + \tilde{\lambda}_t \left(\tilde{\mu}_1 - \sum_{j\neq 1} \frac{1}{J}\tilde{\mu}_j\right) \tag{25}$$

which implies that:

$$a.var(\hat{\alpha}_{1t}^{\text{DID}} - \alpha_{1t}) = E\left[\left(\epsilon_{1t} - \sum_{j\neq 1} \frac{1}{J}\epsilon_{jt}\right) + \tilde{\lambda}_t \left(\tilde{\mu}_1 - \sum_{j\neq 1} \frac{1}{J}\tilde{\mu}_j\right) |D(1,T_0)=1\right]^2 \tag{26}$$

Now note that, under assumptions 1, 2 and $4''$:

$$\begin{aligned}
\frac{1}{T_0}\sum_{t=1}^{T_0}\left[y_{1t} - \sum_{j\neq 1} w_j y_{jt} - a\right]^2 &= \frac{1}{T_0}\sum_{t=1}^{T_0}\left[\left(\epsilon_{1t} - \sum_{j\neq 1} w_j \epsilon_{jt}\right) + \lambda_t\left(\mu_1 - \sum_{j\neq 1} w_j\mu_j\right) - a\right]^2 \\
&\xrightarrow{p} E\left[\left(\epsilon_{1t} - \sum_{j\neq 1} w_j \epsilon_{jt}\right) + \lambda_t\left(\mu_1 - \sum_{j\neq 1} w_j\mu_j\right) - a|D(1,T_0)=1\right]^2
\end{aligned}$$

Note that, for a given $\mathbf{w}$, $a^*(\mathbf{w}) = E\left[\lambda_t\left(\mu_1 - \sum_{j\neq 1} w_j\mu_j\right)|D(1,T_0)=1\right]$. Using the assumption that λ_t is stationary conditional on $D(1,T_0)=1$, we have that $a^*(\mathbf{w}) = \mu_1^1 - \sum_{j\neq 1} w_j\mu_j^1$. Therefore, from Proposition 2, we know that the demeaned SC weights converge to $\bar{\mathbf{w}}^{\text{SC}'}$ that minimize:

$$\Gamma(\mathbf{w}) = E\left[\left(\epsilon_{1t} - \sum_{j\neq 1} w_j\epsilon_{jt}\right) + \tilde{\lambda}_t\left(\tilde{\mu}_1 - \sum_{j\neq 1} w_j\tilde{\mu}_j\right)|D(1,T_0)=1\right]^2 \tag{27}$$

The fact that $\Gamma(w)$ is such that $\Gamma(w_j^{\text{SC}'}) = a.var(\hat{\alpha}_{1t}^{\text{SC}'} - \alpha_{1t})$ and $\Gamma(\{\frac{1}{J},...,\frac{1}{J}\}) = a.var(\hat{\alpha}_{1t}^{\text{DID}} - \alpha_{1t})$ concludes the proof.
■

A.1.4 Proposition 4

Consider the OLS estimator of $y_{1t} = \beta + w_2 y_{2t} + ... + w_{J+1}y_{J+1,t} + u_t$. We consider first the case with no restrictions on the coefficients (which is Hsiao et al. (2012) estimator) and then imposing combinations of the no-intercept, adding-up and non-negativity constraints. Let W be the set of possible weights $\mathbf{w} = (w_2,...,w_{J+1})'$ given the restrictions imposed in the minimization problem and let $\mathbf{w}^* \in \Phi_1 \cap W$ be the cointegration weights that minimize $E[u_t^2]$ subject to $\mathbf{w} \in W$.

The case $W = \mathbb{R}^J$ with intercept follows directly from Proposition 19.3 in Hamilton (1994). We now expand this proposition for the other cases. We first show that this result is valid for the case with no intercept.

Lemma 1 Under assumptions 1, 2, $3'$, and $4'''$, we have that the OLS estimator of $y_{1t} = w_2y_{2t} + ... + w_{J+1}y_{J+1,t} + u_t$ (with no intercept) converges in probability to weights in Φ_1 that minimize the $E[u_t^2]$

Proof.

The proof is a trivial extension of proof of proposition 19.3 in Hamilton (1994). Suppose there is a basis of dimension h for the space of cointegration vectors. We can represent the cointegration relationships by:

$$\mathbf{y}_{1t} = \Gamma'\mathbf{y}_{2t} + \mathbf{z}_t$$
$$\Delta\mathbf{y}_{2t} = \mu_2 + \mathbf{u}_{2t}$$

where $\mathbf{y}_{1t}$ is a vector of dimension hx1 and $\mathbf{z}_t$ represents the error associated with cointegration relation. Note that Assumption 3′ guarantees that $h > 0$ and that we can include unit 1 in the vector $\mathbf{y}_{1t}$. By definition, $\mathbf{z}_t$ is stationary and let $\mu_1 \equiv E[\mathbf{z}_t]$. In addition, μ_2 is the vector with the expected values of $\Delta \mathbf{y}_{2t}$.

Define β_2, β_3,...,β_h as the population coefficients associated with the linear projection of z_{1t} on $\mathbf{z}_{2t} \equiv (z_{2t}, z_{3t}, ..., z_{ht})$,

$$z_{1t} = \beta_2 z_{2t} + ... + \beta_h z_{ht} + u_t$$

where u_t is error with $E[u_t] = \mu^*$, and it is uncorrelated with $\mathbf{z}_{2t}$. Define $u_t \equiv \varpi_t + \mu^*$, where ϖ_t is an unobservable component that has mean zero and is uncorrelated with $\mathbf{z}_{2t}$. First consider the regression of z_{1t} on $\mathbf{z}_{2t}$ and $\mathbf{y}_{2t}$.

$$z_{1t} = \beta' \mathbf{z}_{2t} + \mathbf{\Psi}' \mathbf{y}_{2t} + u_t$$

Note that the true value of $\mathbf{\Psi}$ is zero, β are the coefficients of the linear projection, and u_t is uncorrelated with $\mathbf{z}_{2t}$. The OLS estimator for this model is:

$$\begin{bmatrix} \widehat{\beta} - \beta \\ T^{1/2} \widehat{\mathbf{\Psi}} \end{bmatrix} = \begin{bmatrix} T^{-1} \sum \mathbf{z}_{2t} \mathbf{z}_{2t}^{'} & T^{-3/2} \sum \mathbf{z}_{2t} \mathbf{y}_{2t}^{'} \\ T^{-3/2} \sum \mathbf{y}_{2t}^{'} \mathbf{z}_{2t} & T^{-2} \sum \mathbf{y}_{2t} \mathbf{y}_{2t}^{'} \end{bmatrix}^{-1} \begin{bmatrix} T^{-1} \sum \mathbf{z}_{2t} u_t \\ T^{-3/2} \sum \mathbf{y}_{2t} u_t \end{bmatrix}$$

Since $\mathbf{z}_{2t}$ and u_t are stationary processes, we have:

$$T^{-1} \sum \mathbf{z}_{2t} \mathbf{z}_{2t}^{'} \xrightarrow{p} E\left[\mathbf{z}_{2t} \mathbf{z}_{2t}^{'}\right]$$

$$T^{-1} \sum \mathbf{z}_{2t} u_t \xrightarrow{p} E[\mathbf{z}_{2t} u_t] = \mathbf{0}$$

Using the results in proposition 9.3 in Hamilton (1994):

$$T^{-2} \sum \mathbf{y}_{2t} \mathbf{y}_{2t}^{'} \rightarrow_L \Lambda_2 \left\{ \int [W(r)] [W(r)]^{'} dr \right\} \Lambda_2^{'}$$

Let $\mathbf{z}_{2t} = \mu_{12} + \mathbf{z}_{2t}^*$, where $\mathbb{E}[\mathbf{z}_{2t}^*] = 0$.

$$T^{-3/2} \sum \mathbf{z}_{2t} \mathbf{y}_{2t}^{'} = T^{-3/2} \sum \mathbf{z}_{2t}^* \mathbf{y}_{2t}^{'} + T^{-3/2} \sum \mu_{\mathbf{12}} \mathbf{y}_{2t}^{'} \rightarrow_L 0 + \mu_{\mathbf{12}} \cdot \left\{ \int [W(r)]^{'} dr \right\} \Lambda_2^{'}$$

$$T^{-3/2} \sum \mathbf{y}_{2t} u_t = T^{-3/2} \sum \mathbf{y}_{2t} \varpi_t + \mu^* \cdot T^{-3/2} \sum \mathbf{y}_{2t} \rightarrow_L 0 + \mu^* \cdot \left\{ \int [W(r)]^{'} dr \right\} \Lambda_2^{'}$$

Using these results,

$$\begin{bmatrix} \widehat{\beta} - \beta \\ \widehat{\mathbf{\Psi}} \end{bmatrix} \xrightarrow{p} \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}$$

Note that $\widehat{\mathbf{\Psi}}$ converges in probability to zero since $T^{1/2} \widehat{\mathbf{\Psi}}$ converges to a combination of Wiener processes with finite variance.

At the end, the OLS estimators are consistent for the parameters of the linear projection of z_{1t} on $\mathbf{z}_{2t}$ which minimizes $E[u_t^2]$. Now we need to show the equivalence between these estimators and the coefficients of the OLS regression of y_{1t} on $\mathbf{y_{2t}} \equiv (y_{2t}, ..., y_{J+1t})$. Note that:

$$\begin{bmatrix} 1 & -\beta^{'} \end{bmatrix} \mathbf{z}_t = \mathbf{\Psi}' \mathbf{y}_{2t} + u_t$$

Recall that:

$$\mathbf{z}_t = \mathbf{y}_{1t} - \Gamma^{'} \mathbf{y}_{2t}$$

Using this expression, we have

$$y_{1t} = \beta_2 y_{12t} + \beta_3 y_{13t} + ... + \beta_h y_{1ht} + \left(\mathbf{\Psi}' + \begin{bmatrix} 1 & -\beta^{'} \end{bmatrix} \Gamma^{'} \right) \mathbf{y}_{2t} + u_t \tag{28}$$

Since the OLS coefficients of the linear projection can be consistently estimated by the regression of z_{1t} on a constant, $\mathbf{z}_{2t}$ and $\mathbf{y}_{2t}$, the OLS coefficients of model 28 will give consistent estimators of the transformed coefficients. ■

We show now that this result is also valid for the case with adding-up constraint, whether or not we include an intercept.

Lemma 2 Under assumptions 1, 2, 3′, and 4′′, we have that the OLS estimator of $y_{1t} = w_2 y_{2t} + ... + w_{J+1} y_{J+1,t} + u_t$ (or $y_{1t} = \beta + w_2 y_{2t} + ... + w_{J+1} y_{J+1,t} + u_t$) subject to $W = \{\mathbf{w} \in \mathbb{R}^J \mid \sum_{j=2}^{J+1} w_j = 1\}$ converges in probability to weights in $\Phi_1 \cap W$ that minimize the $E[u_t^2]$

Proof. Just consider the OLS regression of $y_{1t} - y_{2t}$ on $y_{3t} - y_{2t}, ..., y_{J+1,t} - y_{2t}$ (and an intercept for the case with intercept). Under assumption 3′′, this transformed model is cointegrated, so we can apply again Proposition 19.3 from Hamilton (1994) or Lemma 1. ■

We show now that this result is valid for the case with the non-negative constraint.

Lemma 3 Under assumptions 1, 2, 3′, and 4′′, we have that the OLS estimator of $y_{1t} = w_2 y_{2t} + ... + w_{J+1} y_{J+1,t} + u_t$ (or $y_{1t} = \beta + w_2 y_{2t} + ... + w_{J+1} y_{J+1,t} + u_t$) subject to $W = \{\mathbf{w} \in \mathbb{R}^J \mid \sum_{j=2}^{J+1} w_j = 1 \text{ and } w_j \geq 0\}$ (or $W = \{\mathbf{w} \in \mathbb{R}^J \mid w_j \geq 0\}$) converges in probability to weights in $\Phi_1 \cap W$ that minimize the $E[u_t^2]$

Proof. Consider the case $W = \{\mathbf{w} \in \mathbb{R}^J \mid w_j \geq 0\}$.

Suppose first that $\mathbf{w}^* \in int(W)$. This implies that $\mathbf{w}^* \in int(\Phi \cap W)$ relative to Φ. By convexity of $E[u_t^2]$, $\mathbf{w}^*$ also minimizes $E[u_t^2]$ subject to Φ. We know that OLS without the non-negativity constraints converges in probability to $\mathbf{w}^*$. 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 $\mathbf{w}^* \in int(W)$, then it must be that, for all $\epsilon > 0$, $Pr(|\widehat{\mathbf{w}}_u - \mathbf{w}^*| > \epsilon) = 0$ 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 $\mathbf{w}^*$ is on the boundary of W. This means that $w_j^* = 0$ for at least one j. Let $A = \{j | w_j^* = 0\}$. Note first that $\mathbf{w}^*$ 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 $w_j^* = 0$, then we would have the same minimizer, even if we ignore the other non-negative constraints. Suppose there is an $\bar{\mathbf{w}} \neq \mathbf{w}^*$ 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 $\mathbf{w}^*$ 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 $\mathbf{w}^*$. However, this contradicts the fact that $\mathbf{w}^* \in \Phi \cap W$ 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 $\mathbf{w}^*$ (Lemma 2). 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 $\epsilon > 0$ such that $LimPr(|\widehat{\mathbf{w}}' - \widehat{\mathbf{w}}''| > \epsilon) \neq 0$. There are two possible cases.

First, suppose that $\text{Lim}Pr\left(|\widehat{w}_j''| > \epsilon'\right) = 0$ for all $\epsilon' > 0$ (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, $|\widehat{\mathbf{w}}' - \widehat{\mathbf{w}}''| > \epsilon$. 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 $Q_{T_0}(\widehat{\mathbf{w}}'' - e_j \widehat{w}_j'') < Q_{T_0}(\widehat{\mathbf{w}}')$. Since $\widehat{\mathbf{w}}'' - e_j \widehat{w}_j'' \in \{\mathbf{w} | w_j = 0 \; \forall j \in A\}$, this contradicts $\widehat{\mathbf{w}}'$ being OLS subject to $\{\mathbf{w} | w_j = 0 \; \forall j \in A\}$.

Alternatively, suppose that there exists $\epsilon' > 0$ such that $\text{Lim}Pr\left(|\widehat{w}_j''| > \epsilon'\right) \neq 0$. 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 $\widehat{w}_j'' > \epsilon' > 0$. 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 (Proposition 19.3 in Hamilton (1994) and Lemma 2).

Finally, we show that $\widehat{\mathbf{w}}''$ and $\widehat{\mathbf{w}}_r$ are asymptotically equivalent. Note that $\mathbf{w}^*$ 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 $\widehat{\mathbf{w}}'' = \widehat{\mathbf{w}}_r$.

The case $W = \{\mathbf{w} \in \mathbb{R}^J \mid \sum_{j=2}^{J+1} w_j = 1 \text{ and } w_j \geq 0\}$ is essentially the same since this set is convex. ■

Now we can prove Proposition 4.

Proof. Given that OLS estimator of the weights (regardless of which constraints we consider) minimize $E[u_t^2]$ subject to $\mathbf{w} \in \Phi$ (Proposition 19.3 in Hamilton (1994) and Lemmas 1, 2, and 3), the rest of the proof is essentially the same as the proof of Proposition 1. ■

A.1.5 Proposition 5

Proof. We consider now a modification of model 11 in which $\gamma_t = (t, t^2, ..., t^{F_1})$ is a deterministic polynomial time trend, while we maintain that λ_t is a vector of $I(0)$ variables. Consider first the case without the no-intercept, adding-up, and non-negativity constraints. Let h be the number of linearly independent vectors $v \in \mathbb{R}^{J+1}$ such that $v'\mathbf{y}_t$ is stationary, where $\mathbf{y}_t = (y_{1,t}, ..., y_{J+1,t})'$. We consider a triangular representation as the one Phillips (1991) introduced for the cointegration case. Note that assumption $3'$ guarantees that $h > 0$ and that we can find a vector v with first element (the one associated to the treated unit) different from zero such that $v'\mathbf{y}_t$ is stationary. So we consider:

$$\begin{bmatrix} 1 & 0 & \dots & 0 & g_{1,h+1} & \dots & g_{1,J+1} \\ 0 & 1 & \dots & 0 & g_{2,h+1} & \dots & g_{2,J+1} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 & g_{h,h+1} & \dots & g_{h,J+1} \end{bmatrix} \begin{bmatrix} y_{1t} \\ \vdots \\ y_{J+1,t} \end{bmatrix} = \begin{bmatrix} z_{1t} \\ \vdots \\ z_{ht} \end{bmatrix} = \mu_z + \begin{bmatrix} z^*_{1t} \\ \vdots \\ z^*_{ht} \end{bmatrix} \tag{29}$$

where z^*_{jt} is a stationary process with mean zero (note that we can define $\mu_z = E[\mathbf{z}_t]$ because $\mathbf{z}_t$ is stationary), which is a linear combination of the stationary common factors λ_t and the transitory shocks $\{\epsilon_{it}\}_{i=1}^{J+1}$. Define $\mathbf{y}_{1t} \equiv (y_{1t}, ..., y_{ht})'$ and $\mathbf{y}_{2t} \equiv (y_{h+1,t}, ..., y_{J+1,t})'$. Let $\tilde{a} \equiv J + 1 - h$ be the dimension of vector $\mathbf{y}_{2t}$.

Define $\mathbf{z}^*_{2t} \equiv (z^*_{2t}, ..., z^*_{ht})'$ and consider the population regression:

$$z^*_{1t} = \boldsymbol{\beta}\mathbf{z}^*_{2t} + u_t$$

where β is defined such that $E[\mathbf{z}^*_{2t}u_t] = 0$.

Consider now the OLS regression:

$$z^*_{1t} = \alpha + \boldsymbol{\beta}\mathbf{z}^*_{2t} + \phi\mathbf{y}_{2t} + u_t \tag{30}$$

Note that, evaluated at $\boldsymbol{\beta}$ (as defined above) and $\alpha = \phi = 0$, u_t is stationary with mean zero and uncorrelated with $\mathbf{z}^*_{2t}$. We just need to show that the OLS estimator of z^*_{1t} on $\mathbf{z}^*_{2t}$ and $\mathbf{y}_{2t}$ is consistent, and then the rest of the proof is identical to the proof in proposition 19.3 of Hamilton (1994).

Define the $\tilde{a} \times F_1$ matrix $\Theta \equiv [\theta_{h+1}\ \dots\ \theta_{J+1}]'$ which contains the factor loadings associated to the deterministic common factors γ_t for the elements in $\mathbf{y}_{2t}$.[32] From the definition of $\mathbf{y}_{2t}$, we have that $rank(\Theta) = \tilde{a}$. 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 $\tilde{\mathbf{y}}_{2t} \equiv \mathbf{A}\mathbf{y}_{2t}$ for some invertible $\tilde{a} \times \tilde{a}$ matrix $\mathbf{A}$ such that the matrix $\tilde{\Theta} \equiv A\Theta$ with elements $\tilde{\theta}_{j,f}$ has the following property: there exist integers $f_1 > ... > f_{\tilde{a}} \geq 1$ such that $\tilde{\theta}_{j,f_j} \neq 0$ and $\tilde{\theta}_{j,f} = 0$ if $f > f_j$. We show that it is possible to construct such matrix given the definition of $\mathbf{y}_{2t}$.

Let first f_1 be the largest $f \in \{1, ..., F_1\}$ so that $\theta_{j,f_1} \neq 0$ for some $j \in \{h+1, .., J+1\}$. We set $\tilde{y}_{1,t} = y_{j,t}$ for a $j \in \{h+1, .., J+1\}$ such that $\theta_{j,f_1} \neq 0$. For the second row, consider linear combinations $b'\mathbf{y}_{2t}$ for some $b \in \mathbb{R}^{\tilde{a}}$ and let $\tilde{\theta}_f(b)$ be the f-component of the $(1 \times F_1)$ row vector $b'\Theta$. Consider now the set of all linear combinations $b'\mathbf{y}_{2t}$ such that $\tilde{\theta}_{f_1}(b) = 0$, 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 $\tilde{\theta}_{f_1}(b) = 0$ and

[32] Note that, by definition, $\mathbf{y}_{2t} - \Theta\gamma_t'$ is a $\tilde{a} \times 1$ vector of stationary variables.

$\tilde{\theta}_{f_2}(b) \neq 0$ and set $\tilde{y}_{2,t} = b'\mathbf{y}_{2t}$. Since, $rank(\Theta) = \tilde{a}$, we can continue this construction until we get $\tilde{y}_{\tilde{a},t} = b'\mathbf{y}_{2t}$ for a linear combination b such that $\tilde{\theta}_f(b) = 0$ for all $f \geq f_{\tilde{a}-1}$ and $\tilde{\theta}_f(b) \neq 0$ for only one $f = f_{\tilde{a}}$ such that $1 \leq f_1 < f_{\tilde{a}-1}$.

We consider now the OLS regression:

$$z^*_{1t} = \alpha + \boldsymbol{\beta}\mathbf{z}^*_{2t} + \tilde{\phi}\tilde{\mathbf{y}}_{2t} + u_t \tag{31}$$

The OLS estimator for this model is:

$$\begin{bmatrix} \widehat{\alpha} \\ \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ T^{f_1}\widehat{\tilde{\phi}}_{h+1} \\ \vdots \\ T^{f_{\tilde{a}}}\widehat{\tilde{\phi}}_{J+1} \end{bmatrix} = \begin{bmatrix} 1 & \frac{\sum \mathbf{z}^{*\prime}_{2t}}{T} & \frac{\sum \tilde{y}_{1,t}}{T^{f_1+1}} & \cdots & \frac{\sum \tilde{y}_{1,t}}{T^{f_{\tilde{a}}+1}} \\ \frac{\sum \mathbf{z}^*_{2t}}{T} & \frac{\sum \mathbf{z}^*_{2t}\mathbf{z}^{*\prime}_{2t}}{T} & \frac{\sum \mathbf{z}^*_{2t}\tilde{y}_{1,t}}{T^{f_1+1}} & \cdots & \frac{\sum \mathbf{z}^*_{2t}\tilde{y}_{\tilde{a},t}}{T^{f_{\tilde{a}}+1}} \\ \frac{\sum \tilde{y}_{1,t}}{T^{f_1+1}} & \frac{\sum \tilde{y}_{1,t}\mathbf{z}^{*\prime}_{2t}}{T^{f_1+1}} & \frac{\sum \tilde{y}^2_{1,t}}{T^{2f_1+1}} & \cdots & \frac{\sum \tilde{y}_{1,t}\tilde{y}_{\tilde{a},t}}{T^{f_1+f_{\tilde{a}}+1}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\sum \tilde{y}_{\tilde{a},t}}{T^{f_{\tilde{a}}+1}} & \frac{\sum \tilde{y}_{\tilde{a},t}\mathbf{z}^{*\prime}_{2t}}{T^{f_{\tilde{a}}+1}} & \frac{\sum \tilde{y}_{\tilde{a},t}\tilde{y}_{1,t}}{T^{f_1+f_{\tilde{a}}+1}} & \cdots & \frac{\sum \tilde{y}^2_{\tilde{a},t}}{T^{2f_{\tilde{a}}+1}} \end{bmatrix}^{-1} \times \begin{bmatrix} T^{-1}\sum u_t \\ T^{-1}\sum \mathbf{z}^*_{2t}u_t \\ T^{-(1+f_1)}\sum \tilde{y}_{1,t}u_t \\ \vdots \\ T^{-(1+f_{\tilde{a}})}\sum \tilde{y}_{\tilde{a},t}u_t \end{bmatrix} \tag{32}$$

Note that $\sum \tilde{y}^2_{j,t}$ for $j \in \{1,..,\tilde{a}\}$ involves terms of the order t^q with $q \leq 2f_j$, $t^q v_t$ with $q \leq f_j$, v_t^2, and v_t, where v_t is a stationary process. From Proposition 17.1.(h) in Hamilton (1994), we know that terms $T^{-(2f_j+1)}\sum t^{2f_j}$ converge in probability to a positive constant, while terms $T^{-(2f_j+1)}\sum t^q$ with $q < 2f_j$ converge in probability to zero. From Lemma 4.(a) in Carvalho et al. (2016), we also have that terms $T^{-(2f_j+1)}\sum t^q v_t$ with $q \leq f_j$ converge in probability to zero, while the same will be true for $T^{-(2f_j+1)}\sum v_t$ and $T^{-(2f_j+1)}\sum v_t^2$. Therefore, $T^{-(2f_j+1)}\sum y^2_{j,t}$ converge in probability to a positive constant. Following similar arguments, the off-diagonal elements $T^{-(f_j+f_{j'}+1)}\sum \tilde{y}_{j,t}\tilde{y}_{j',t}$ with $j \neq j'$ and $T^{-(f_j+1)}\sum \tilde{y}_{j,t}\mathbf{z}^{*\prime}_{2t}$ converge in probability to zero. We also have that $T^{-1}\mathbf{z}^{*\prime}_{2t}$ converge in probability to zero and $T^{f_j+1}\sum \tilde{y}_{j,t}$ converge to finite constant. Finally, stationarity of $\mathbf{z}_{2t}$ guarantees that the second element in the diagonal will converge in probability to a positive definite matrix. Therefore, the first matrix in the right hand side of equation 32 converges in probability to an invertible matrix.

Now, from the definition of u_t, we have that $T^{-1}\sum u_t$ and $T^{-1}\sum \mathbf{z}^*_{2t}u_t$ converge in probability to zero. Also, $\sum \tilde{y}_{j,t}u_t$ involves terms $t^q v_t$ with $q \leq f_j$, v_t^2, and v_t, so that $T^{-(1+f_j)}\sum \tilde{y}_{j,t}u_t$ converges in probability to zero. Therefore, $\hat{\alpha} \xrightarrow{p} 0$, $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$, and $\widehat{\tilde{\phi}} \xrightarrow{p} 0$. Since we have that $\tilde{\mathbf{y}}_{2t} = \mathbf{A}\mathbf{y}_{2t}$ where $\mathbf{A}$ has full rank, then we also have that $\widehat{\phi} \xrightarrow{p} 0$, where $\widehat{\phi}$ is the OLS estimator of ϕ from equation 30.

Following the same steps as in Proposition 19.3 in Hamilton (1994), this implies that OLS of y_{1t} on $y_{2t}, ..., y_{J+1,t}$ yields consistent estimators for the parameters minimize the variance of u_t conditional on reconstructing the polynomial trend of the treated unit. Following the same steps as in Lemmas 1, 2 and 3, we can show that this result remains valid for OLS with combinations of the no-intercept, adding-up, and non-negativity constraints. ■

A.2 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 $E[\lambda_t] = 0$. We have that the factor loadings are given by:

$$\mu_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \mu_2 = \begin{pmatrix} 0.5 \\ 1 \end{pmatrix}, \mu_3 = \begin{pmatrix} 1.5 \\ 1 \end{pmatrix}, \mu_4 = \begin{pmatrix} 0.5 \\ 0 \end{pmatrix}, \mu_5 = \begin{pmatrix} 1.5 \\ 1 \end{pmatrix} \tag{33}$$

Note that the linear combination $0.5\mu_2 + w_1^3\mu_3 + w_1^5\mu_5 = \mu_1$ with $w_1^3 + w_1^5 = 0.5$ satisfy assumption 3. Note also that DID equal weights would set the first factor loading to 1, which is equal to μ_1^1, but the second factor loading would be equal to $0.75 \neq \mu_1^2$. 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 $\sigma^2_\epsilon = E[(\lambda^1_t)^2] = E[(\lambda^2_t)^2] = 1$, 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^1_t, then the SC estimator would be more asymptotically biased than the DID estimator.

A.3 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. Let γ be the expected value of the asymptotic distribution of $\hat{\alpha}_{1t} - \alpha_{1t}$. Therefore, we have that:

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

Given that $\hat{w}_j$ is a consistent estimator for $\bar{w}_j$, if we have that ϵ_{it} has finite variance, then:

$$|E\left[(\bar{w}_j - \hat{w}_j)\epsilon_{jt}\right]| \leq E\left[|(\bar{w}_j - \hat{w}_j)\epsilon_{jt}|\right] \leq \sqrt{E\left[(\bar{w}_j - \hat{w}_j)^2\right] E\left[(\epsilon_{jt})^2\right]} \rightarrow 0$$

Similarly, if λ_t^f has finite variance for all $f = 1, ..., F$, then $E\left[\lambda_t(\bar{w}_j - \hat{w}_j)\right]\mu_j \rightarrow 0$.

A.4 Alternatives specifications and alternative estimators

A.4.1 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:

$$\begin{aligned} \{\hat{w}_j\}_{j \neq 1} &= \text{argmin}_{w \in W} \left[\frac{1}{T_0}\sum_{t=1}^{T_0}\left(y_{1t} - \sum_{j \neq 1} w_j y_{jt}\right)\right]^2 \\ &= \text{argmin}_{w \in W} \left[\frac{1}{T_0}\sum_{t=1}^{T_0}\left(\epsilon_{1t} - \sum_{j \neq 1} w_j \epsilon_{jt} + \lambda_t\left(\mu_1 - \sum_{j \neq 1} w_j \mu_j\right)\right)\right]^2 \end{aligned} \tag{34}$$

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 4′, the objective function converges in probability to:

$$\Gamma(\mathbf{w}) = \left[E\left[\lambda_t | D(1, T_0) = 1\right]\left(\mu_1 - \sum_{j \neq 1} w_j \mu_j\right)\right]^2 \tag{35}$$

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 $E[\lambda_t^k] = 0$ for $k > 1$, the objective function collapses to:

$$\Gamma(\mathbf{w}) = \left[\left(\mu_1^1 - \sum_{j \neq 1} w_j \mu_j^1\right)\right]^2 \tag{36}$$

Therefore, even if we assume that there exists at least one set of weights that reproduces all factor loadings (Assumption 3), the objective function will only look for weights that approximate the first factor loading. This is problematic because it might be that assumption 3 is satisfied, but there are weights $\{\tilde{w}_j\}_{j \neq 1} \notin \Phi$ 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 Monte Carlo simulations in Ferman et al. (2016), who show that this specification performs particularly bad in allocating the weights correctly.

A.4.2 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 41 changes to:

$$\begin{cases} y_{it}(0) = \delta_t + \theta_t Z_i + \lambda_t \mu_i + \epsilon_{it} \\ y_{it}(1) = \alpha_{it} + y_{it}(0) \end{cases} \tag{37}$$

We also modify assumption 3 so that the weights reproduce both μ_1 and Z_1.

Assumption 3′′ (existence of weights)

$$\exists\ \mathbf{w} \in W \mid \mu_1 = \sum_{j \neq 1} w_j^* \mu_j,\ Z_1 = \sum_{j \neq 1} w_j^* Z_j$$

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:

$$\widehat{\mathbf{w}}(V) \in \text{argmin}_{\mathbf{w} \in W} ||X_1 - X_0 \mathbf{w}||_V \tag{38}$$

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 $E[\lambda_t^k] = 0$ for $k > 1$, objective function of this minimization problem converges to $||\bar{X}_1 - \bar{X}_0 \mathbf{w}||_V$, where:

$$\bar{X}_1 - \bar{X}_0 \mathbf{w} = \begin{bmatrix} E[\theta_t | D(1, T_0) = 1] \left(Z_1 - \sum_{j \neq 1} w_j Z_j \right) + \left(\mu_1^1 - \sum_{j \neq 1} w_j \mu_j^1 \right) \\ \left(Z_1^1 - \sum_{j \neq 1} w_j Z_j^1 \right) \\ \vdots \\ \left(Z_1^R - \sum_{j \neq 1} w_j Z_j^R \right) \end{bmatrix} \tag{39}$$

Similarly to the case with only the average pre-intervention outcome value as economic predictor, it might be that assumption 3′′ is satisfied, 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 that satisfy assumption 3′′ for any given matrix V.

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:

$$\widehat{\mathbf{w}} \quad = \quad \text{argmin}_{w \in \widetilde{W}} \left[\frac{1}{T_0} \sum_{t=1}^{T_0} \left(y_{1t} - \sum_{j \neq 1} w_j y_{jt} \right) \right]^2 \tag{40}$$

where $\widetilde{W} \subseteq W$ is the set of $\mathbf{w}$ such that $\mathbf{w}$ is the solution to problem 38 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 that satisfy assumption 3′′ even when $T_0 \to \infty$. More specifically, if the variance of ϵ_{it} is large, then the SC estimator would tend to choose weights that are uniform across the control units in detriment of weights that satisfy assumption 3′′. Therefore, it is not possible to guarantee that this SC estimator would be asymptotically unbiased. MC simulation results in Ferman et al. (2016) confirm that this SC specification systematically misallocates more weight than alternatives that use a large number of pre-treatment outcome lags as predictors.

A.4.3 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:

$$y_{it}(0) = \dot{\lambda}_t \dot{\mu}_i + \epsilon_{it} \tag{41}$$

where $\dot{\lambda}_t = (\delta_t, \lambda_t)$ and $\dot{\mu}_i = (1, \mu_i)'$. We modify assumption $4'$ to include assumptions on the convergence of δ_t.

Assumption $4'''''$ (convergence of pre-treatment averages) $\frac{1}{T_0} \sum_{t=1}^{T_0} \dot{\lambda}_t \xrightarrow{p} \omega_0$, $\frac{1}{T_0} \sum_{t=1}^{T_0} \dot{\lambda}_t' \dot{\lambda}_t \xrightarrow{p} \Omega_0$, $\frac{1}{T_0} \sum_{t=1}^{T_0} \epsilon_{jt} \xrightarrow{p} 0$, $\frac{1}{T_0} \sum_{t=1}^{T_0} \epsilon_{jt}^2 \xrightarrow{p} \sigma_\epsilon^2$, and that $\epsilon_{jt} \perp \dot{\lambda}_s$ for all s, t and for all j.

Under assumptions 1 and $4'''''$, the objective function converges in probability to:

$$\widehat{Q}_{T_0}(\mathbf{w}) \xrightarrow{p} Q_0(\mathbf{w}) = \sigma_\epsilon^2 + \sigma_\epsilon^2 \sum_{j \neq 1} (w_j)^2 + \left(\mu_1 - \sum_{j \neq 1} w_j \mu_j \right)' \Omega_0 \left(\mu_1 - \sum_{j \neq 1} w_j \mu_j \right)$$

Note that the first element of this expression is a constant, and it does not matter for the optimization problem. Except for the constant, we can represent this objective function using matrices. Define $\mathbf{w}$ as a vector $(J \times 1)$ of the weights, $\{w_j\}_{j \neq 1}$, μ_1 is a vector $(K \times 1)$ with the factor loadings for the treated units and μ_0 is a matrix $(K \times J)$ that contains the factor loadings for the all the control units, we can write this optimization problem as:

$$\arg \min_{\mathbf{w} \in W} \sigma_\epsilon^2 \mathbf{w}'\mathbf{w} + (\mu_1 - \mu_0 \mathbf{w})' \Omega_0 (\mu_1 - \mu_0 \mathbf{w})$$

where W is a convex set. This is a minimization of a quadratic function in a convex space, and has a unique interior solution $\mathbf{w}_0$.

By assumptions 1 and $4'''''$, $\widehat{Q}_{T_0} \xrightarrow{p} Q_0$. In addition, $\widehat{Q}_{T_0}$ is convex and $\mathbf{w}_0$ is the unique maximum of Q_0 and belongs to the interior of the convex set W. 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 $y_{1t} - y_{2t}$ 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 relax the non-negativity constraint follows from the same arguments as in the proof of Lemma 3.

Given that we assure convergence of $\widehat{\mathbf{w}}$ to $\arg\min_{\mathbf{w} \in W} Q_0(\mathbf{w})$, the fact that $\widehat{\mathbf{w}}$ does not 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.4.4 IV-Like SC Estimator

As noted by Doudchenko and Imbens (2016), the minimization problem when one includes all pre-intervention lags is equivalente to a restricted OLS estimator of y_{1t} on $y_{2,t}, ..., y_{J+1,t}$. For weights $\{w_j^*\}_{j \neq 1} \in \Phi$, we can write:

$$y_{1t} = \sum_{j=1}^{J+1} w_j^* y_{jt} + \eta_t, \text{ for } t \leq T_0$$

where:

$$\eta_t = \epsilon_{1t} - \sum_{j=1}^{J+1} w_j^* \epsilon_{jt}$$

The key problem is that η_t is correlated with y_{jt}, which implies that the restricted OLS estimators are 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 $\mathbf{y}_t = (y_{2,t}, ..., y_{J+1,t})'$, $\mu_\mathbf{0}$ be a $(F \times J)$ matrix with columns μ_j, $\epsilon_t = (\epsilon_{2,t}, ..., \epsilon_{J+1,t})$, and $\mathbf{w} = (w_1^2, ..., w_1^{J+1})'$. In this case, we can look at:

$$\begin{aligned} \mathbf{y}_{t-1}(y_{1t} - \mathbf{y}_t'\mathbf{w}) &= (\mu_0'\lambda_{t-1}' + \epsilon_{t-1})\lambda_t (\mu_1 - \mu_\mathbf{0}\mathbf{w}) + (\mu_0'\lambda_{t-1}' + \epsilon_{t-1})(\epsilon_{1t} - \epsilon_t'\mathbf{w}) \\ &= \mu_0'\lambda_{t-1}'\lambda_t (\mu_1 - \mu_\mathbf{0}\mathbf{w}) + \epsilon_{t-1}\lambda_t (\mu_1 - \mu_\mathbf{0}\mathbf{w}) + \mu_0'\lambda_{t-1}'(\epsilon_{1t} - \epsilon_t'\mathbf{w}) + \epsilon_{t-1}(\epsilon_{1t} - \epsilon_t'\mathbf{w}) \end{aligned} \tag{42}$$

If we assume that ϵ_{it} is independent across t and independent of λ_t, then, for $t < T_0$:

$$E\left[\mathbf{y}_{t-1}(y_{1t} - \mathbf{y}_t'\mathbf{w})\right] = \mu_0' E\left[\lambda_{t-1}'\lambda_t\right](\mu_1 - \mu_\mathbf{0}\mathbf{w}) \tag{43}$$

Therefore, if the $(J \times F)$ matrix $\mu_0' E\left[\lambda_{t-1}'\lambda_t\right]$ 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.[33] Assumption $4'$ guarantees that the pre-treatment averages of the moment conditions, which are based on the conditional process z_{jt} converge in probability to the unconditional moment conditions. One challenge to analyze this method is that there might be multiple solutions to the moment condition. Based on the results in 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.5 Asymptotic bias in Wong (2015)

In the third chapter of his thesis, Wong (2015) shows in Section 3.9 that the SC weights is given by:

$$\widehat{\mathbf{w}} - \mathbf{w} = \left((Y'Y)^{-1} - (Y'Y)^{-1}\mathbf{j}(\mathbf{j}'(Y'Y)^{-1}\mathbf{j})^{-1}\mathbf{j}'(Y'Y)^{-1}\right) Y'(\zeta - Y'\mathbf{w}) \tag{44}$$

where ζ is a $(T_0 \times 1)$ vector with the pre-intervention outcomes for the treated group (with elements y_{1t}), while Y is a $(T_0 \times J)$ matrix with the pre-intervention outcomes for the control units (with rows $\mathbf{y}_t'$). Also, let $\mathbf{j}$ be a $(J \times 1)$ vector of ones.[34]

Let $E[y_{1t}] = y_{1t}^*$ and $E[\mathbf{y}_t] = \mathbf{y}_t^*$, so that $y_{1t} = y_{1t}^* + \epsilon_{1t}$ and $\mathbf{y}_t = \mathbf{y}_t^* + \epsilon_t$. The main assumption in his model states that there exists weights $\mathbf{w}$ such that $y_{1t}^* = \mathbf{y}_t^{*\prime}\mathbf{w}$. Assuming $(y_{1t}, \mathbf{y}_t')$ stationary and ergodic, they show that $\frac{1}{T_0}Y'Y \to E[\mathbf{y}_t\mathbf{y}_t']$ and $\frac{1}{T_0}Y'(\zeta - Y\mathbf{w}) \to E[\mathbf{y}_t(y_{1t} - \mathbf{y}_t'\mathbf{w})]$. Wong (2015) argues that $E[\mathbf{y}_t(y_{1t} - \mathbf{y}_t'\mathbf{w})] = 0$. However, we have that:

$$\begin{aligned} E[\mathbf{y}_t(y_{1t} - \mathbf{y}_t'\mathbf{w})] &= E[\mathbf{y}_t y_{1t}] - E[\mathbf{y}_t\mathbf{y}_t'\mathbf{w}] = E[(\mathbf{y}_t^* + \epsilon_t)(y_{1t}^* + \epsilon_{1t})] - E[(\mathbf{y}_t^* + \epsilon_t)(\mathbf{y}_t^* + \epsilon_t)'\mathbf{w}] \\ &= \mathbf{y}_t^* y_{1t}^* - \mathbf{y}_t^*\mathbf{y}_t^{*\prime}\mathbf{w} - E[\epsilon_t\epsilon_t']\mathbf{w} = -E[\epsilon_t\epsilon_t']\mathbf{w} \end{aligned} \tag{45}$$

Therefore, this term will only be equal to zero if $var(\epsilon_t) = 0$, which is exactly the condition we find so that the SC weights would be consistent.

[33] 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.

[34] We use here the same notation as in Wong (2015), which is slightly different from the notation used in our paper.