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: November, 2016

Abstract

The synthetic control (SC) method has been recently proposed as an alternative to estimate treatment effects in comparative case studies. In this paper, we revisit the SC method in a linear factor model setting and consider the asymptotic properties of the SC estimator when the number of pre-treatment periods (T_0) goes to infinity. Differently from Abadie et al. (2010), we do not condition the analysis on a close-to-perfect pre-treatment fit, as the probability that this happens goes to zero when T_0 is large. We show that, even when a close-to-perfect fit is not achieved, the SC method can substantially improve relative to the difference-in-differences (DID) estimator, both in terms of bias and variance. However, we show that, in our setting, the SC estimator is asymptotically biased if treatment assignment is correlated with the unobserved heterogeneity. If common factors are stationary, then the asymptotic bias of the SC estimator goes to zero when the variance of the transitory shocks is small, which is also the case in which it is more likely that the pre-treatment fit will be good. If a subset of the common factors is non-stationary, then the SC estimator can be asymptotically biased even conditional on a close-to-perfect fit. In this case, the identification assumption relies on orthogonality between treatment assignment and the stationary common factors. Finally, we also consider the statistical properties of the permutation tests suggested in Abadie et al. (2010).

Keywords: synthetic control, difference-in-differences; linear factor model, inference, permutation test
JEL Codes: C12; 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 2016 California Econometrics Conference, 2016 Bristol Econometric Study Group, USP, George Washington University, University of Miami, and PUC-Chile 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 (2016), *"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"*.

Abadie et al. (2010) show that, conditional on a perfect matching in the pre-treatment periods, the bias of the SC estimator is bounded by a term that goes to zero with the number of pre-treatment periods (T_0), even if treatment assignment is correlated with the unobserved heterogeneity.[1] In this paper, we revisit the SC method in a linear factor model setting, and consider the asymptotic properties of the SC estimator when T_0 goes to infinity. Differently from Abadie et al. (2010), we do not condition the analysis on a close-to-perfect pre-treatment match, as the probability that this happens goes to zero when T_0 is large. Assuming a model with stationary common factors, we show that the SC weights, in our setting, converge in probability to weights that do *not* reconstruct the factor loadings of the treated unit.[2] As a consequence, the SC estimator is asymptotically biased if treatment assignment is correlated with the unobserved heterogeneity.[3] This result is 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 goes to zero when the transitory shocks are small, which is also the case in which it is more likely that the pre-treatment match will be close to perfect for a moderate T_0.

We 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. We show that, if selection into treatment is only correlated with time-invariant common factors (which is essentially the identification assumption of the difference-in-differences (DID) model), then this demeaned SC estimator is unbiased. In this case, we can guarantee that the asymptotic variance of this demeaned SC estimator is lower than the

[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 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 economic predictors.

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

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. However, 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. Therefore, while we argue that the SC method is asymptotically biased if treatment assignment is correlated with time-varying confounders, it still provides important improvement over standard methods, even if a close-to-perfect pre-treatment match is not achieved. We also provide 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.

We consider, in Monte Carlo (MC) simulations, the capacity of the SC method to allocate weights that correctly reconstruct the factor loadings of the treated unit conditional on a close-to-perfect pre-treatment match. With stationary errors, we show that, with moderate T_0, the pre-treatment fit will only be good if the variance of the transitory shocks is very low. In this case, the discrepancy between the factor loadings of the treated and the SC units would be relatively low. If the variance of the transitory shocks is higher, then the pre-treatment fit will only be good if we have very few pre-treatment periods. In this case, the discrepancy between the factor loadings of the treated and the SC units can be large even conditional on having a good pre-treatment fit, as suggested in Abadie et al. (2010) and Abadie et al. (2015). Therefore, we provide evidence that, with stationary data and a moderate T_0, the SC estimator would only provide a close-to-perfect match, and therefore be close to unbiased even if treatment assignment is correlated with unobserved characteristics that vary with time, when the variance of the transitory shocks converges to zero.

We also consider the case with non-stationary errors, which is more consistent with the applications presented in Abadie and Gardeazabal (2003), Abadie et al. (2010), and Abadie et al. (2015). In MC simulations with both stationary and non-stationary common factors, we show that the probability of having a good pre-intervention fit is substantially higher than in the stationary case, and that the SC weights is highly successful in reconstructing the factor loadings associated with the non-stationary common factors. However, the SC weights may not reconstruct well the factor loadings associated with the stationary common factors, *even conditional on having a good pre-intervention fit.* This happens because, even with moderate T_0, the non-stationary common factors dominate the stationary common factors and the transitory shocks. These results suggest that this is a scenario where the SC method significantly improves relative to DID, as it is extremely successful in selecting a SC unit that follows the same non-stationary trend as the treated unit. However, an important qualification is that the identification assumption in this case relies on orthogonality between treatment assignment and the stationary common factors, as the SC method fails to reconstruct the factor loadings associated with the stationary common factors, *even conditional on a close-to-perfect match*.

Finally, we consider the statistical properties of the permutation test proposed in Abadie et al. (2010).[4] In the absence of random assignment, Abadie et al. (2010) and Abadie et al. (2015) interpret the p-value from their permutation test as *"the probability of obtaining an estimate at least as large as the one obtained for the unit representing the case of interest when the intervention is reassigned at random in the data set"* (Abadie et al. (2015), page 500). While we agree this is a useful measure, it is important to evaluate the statistical properties of the permutation test. We first show that the graphical analysis proposed in Abadie et al. (2010) or a permutation test using the post-treatment mean squared prediction error (MSPE) as test statistic might lead to important size distortions, as the distribution of the post-treatment prediction errors for a given permutation might depend on, for example, the variance of the transitory shocks or the concentration of the SC weights. We show that a permutation test using the ratio of post/pre-treatment MSPE as test statistic, also suggested in Abadie et al. (2010), ameliorates this problem as, under some conditions, it generates test statistics with the same asymptotic (marginal) distribution for all permutations.[5] However, even under such conditions, we cannot guarantee that the test is asymptotically valid, as the test statistics are generally not based on functions of the data that exhibit approximate symmetry.[6] We provide examples in which we can have size distortions even when the test statistics for all permutations have the same marginal distribution. We also show that heteroskedasticity might generate important size distortions if the number of pre-treatment periods is small, even in situations in which the asymptotic size distortions would be negligible. This might happen because, with small T_0, the model might overfit the pre-treatment mean squared prediction error (MSPE), so it might not provide a proper correction for the post-treatment

[4]There are two recent papers that analyze in detail the permutation test proposed in Abadie et al. (2010). Firpo and Possebom (2016) formalize the permutation test for the case where treatment is randomly assigned. In this case, the inference method suggested in Abadie et al. (2010) would provide valid inference for *unconditional* tests. Our paper considers the asymptotic properties of the permutation test when we relax the hypothesis of random assignment. Also, even under random assignment, we can consider hypothesis testing conditional on the data on hand. See Ferman and Pinto (2016) for details on why conditional tests should be preferable when there are few treated units. In another recent paper, Ando and Sävje (2013) argue that the permutation test proposed by Abadie et al. (2010) is generally not valid and derive an alternative inference method. Differently from Ando and Sävje (2013), we consider the asymptotic properties of Abadie et al. (2010) permutation test when the number of pre-intervention is large.

[5]This will be the case if linear combinations of the transitory shocks and common factors are stationary, serially uncorrelated, and i.i.d. across units up to a scale parameter. We derive an alternative test statistic that guarantees the same asymptotic expected value and variance for all permutations under weaker conditions.

[6]Canay et al. (2014) develop a theory of randomization tests under an approximate symmetry assumption. They provide conditions under which it is possible to construct tests that asymptotically control the probability of a false rejection whenever the distribution of the observed data exhibits approximate symmetry in the sense that the limiting distribution of a function of the data exhibits symmetry under the null hypothesis.

MSPE.

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 (2013) derive conditions under which the assumption 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. 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. Xu (2016) 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 (2013) and Xu (2016), we consider the case with a finite number of control units and let the number of pre-intervention periods go to infinity. We also provide an example in which the SC weights do not reconstruct the factor loadings of the treated unit even when the number of control units goes to infinity. This happens because the number of common factors grow with the number of control units, breaking one of the assumptions in Gobillon and Magnac (2013) and Xu (2016). Wong (2015) and Powell (2016) also consider the asymptotic properties of the SC estimator (or a generalization of the SC estimator) when T_0 goes to infinity while holding the number of control units constant. They argue that the estimators would be asymptotically unbiased. However, it is possible to show that, in their setting, the estimators will also be asymptotically biased under the same conditions we find in our paper.[7]

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 than Abadie et al. (2010). Then we show that the SC estimator that uses all pre-treatment outcome lags as economic predictors is, in our setting, asymptotically biased. We also consider the asymptotic properties of alternative specifications of the SC estimator. In Section 3 we propose two alternatives to the original SC estimator. In Section 4 we present a particular class of linear factor models in which we consider the asymptotic and finite T_0 properties of the SC estimator. We start with a stationary model, and then we consider a model in which a subset of the common factors is non-stationary. In Section 5 we consider the statistical properties of the permutation test proposed in Abadie et al. (2010) in our setting. We conclude in Section 6.

[7]Details in Appendix A.5.

2 Revisiting the Synthetic Control Model

2.1 The Synthetic Control 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. We assume ϵ_{it} independent across units and in time. Note that the unobserved error $u_{it} = \lambda_t \mu_i + \epsilon_{it}$ might be correlated across unit and in time due to the presence of $\lambda_t \mu_i$. As in Abadie et al. (2010), Gobillon and Magnac (2013) and Powell (2016), we allow for correlation between $\lambda_t \mu_i$ and the treatment assignment. 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 until Section 2.3.2.

We assume that there is a stable linear combination of the control units that absorbs all time correlated shocks $\lambda_t \mu_i$.

Assumption 1 (existence of weights):

$$\exists \ \mathbf{w}^* = \{w_1^{*j}\}_{j \neq 1} \mid \mu_1 = \sum_{j \neq 1} w_1^{*j} \mu_j, \ \sum_{j \neq 1} w_1^{*j} = 1, \text{ and } w_1^{*j} \geq 0$$

Note that we consider the existence of weights that reconstruct the unobserved factors loadings μ_1, following the structure of Ando and Sävje (2013) and Powell (2016).[8] There is no guarantee that there is a unique set of weights that satisfies assumption 1, so we define $\Phi_1 = (\mathbf{w} = \{w_1^j\}_{j \neq 1} \mid \mu_1 = \sum_{j \neq 1} w_1^j \mu_j, \ \sum_{j \neq 1} w_1^j = 1, \text{ and } w_1^j \geq 0)$ as the set of weights that satisfy this condition. We treat $\mathbf{w}^*$ as a nuisance parameter that

[8] Powell (2016) treats μ_i as random variables, so he considers that assumption 1 is valid in expectation. Wong (2015) considers weights that reconstruct the expected value of the potential outcome if the observation is not treated, without imposing a linear factor model structure. As we show in Appendix A.5, our main results remain valid in the setting considered in Wong (2015).

need to be estimated to construct the SC 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 $y_{1t} = \sum_{j \neq 1} w_1^{*j} y_{jt}$ 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"*. They show that the condition $y_{1t} = \sum_{j \neq 1} w_1^{*j} y_{jt}$ for all $t \leq T_0$ (for large T_0) can only be satisfied as long as our assumption 1 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 1 is valid, 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. We analyze the properties of the SC weights conditional on a good pre-intervention fit in Monte Carlo simulations in Section 4.

The main idea of the SC method consists of estimating the SC weights $\widehat{\mathbf{w}}_1 = \{\hat{w}_1^j\}_{j \neq 1}$ using information on the pre-treatment period. Then we construct the SC estimator $\hat{\alpha}_{1t} = y_{1t} - \sum_{j \neq 1} \hat{w}_1^j y_{jt}$ for $t > T_0$. Abadie et al. (2010) suggest a minimization problem to estimate these weights using the pre-intervention data. They define a set of K economic predictors where X_1 is a $(K \times 1)$ vector containing the economic predictors for the treated unit and X_0 is a $(K \times J)$ matrix of economic predictors for the control units.[9] The SC weights are estimated by minimizing $||X_1 - X_0\mathbf{w}||_V$ subject to $\sum_{i=2}^{J+1} w_1^j = 1$ and $w_1^j \geq 0$, 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 $||\mathbf{Y}_1^P - \mathbf{Y}_0^P \widehat{\mathbf{w}}(V)||$.

As argued in Ferman et al. (2016), the SC method does not provide a clear guidance on how one should choose the economic predictors in matrices X_1 and X_0. This reflects in a wide range of different specification

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

choices in SC applications. We consider here 3 common specifications: (1) the use of all pre-intervention outcome values, (2) the use of the average of the pre-intervention outcomes, and (3) the use of other time invariant covariates in addition to the average of the pre-intervention outcomes.[10]

2.2 The asymptotic bias of the SC estimator

We focus first 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 an M-estimator given by:

$$\begin{aligned} \{\hat{w}_1^j\}_{j\neq 1} &= \text{argmin}_{\mathbf{w}\in W} \frac{1}{T_0}\sum_{t=1}^{T_0}\left[y_{1t} - \sum_{j\neq 1} w_1^j y_{jt}\right]^2 \\ &= \text{argmin}_{\mathbf{w}\in W} \frac{1}{T_0}\sum_{t=1}^{T_0}\left[\epsilon_{1t} - \sum_{j\neq 1} w_1^j \epsilon_{jt} + \lambda_t\left(\mu_1 - \sum_{j\neq 1} w_1^j \mu_j\right)\right]^2 \end{aligned} \tag{2}$$

where $W = \{\{w_1^j\}_{j\neq 1} \in \mathbb{R}^J | w_1^j \geq 0 \text{ and } \sum_{j\neq 1} w_1^j = 1\}$.

We impose conditions such that this objective function converges uniformly in probability to its population average. We relax this assumption in Section 4.3.

Assumption 2 (stationary process): $(\epsilon_{jt}, \lambda_t)'$ is weakly stationary and second moment ergodic.

Under assumption 2, we have that:

$$\frac{1}{T_0}\sum_{t=1}^{T_0}\left[\epsilon_{1t} - \sum_{j\neq 1} w_1^j \epsilon_{jt} + \lambda_t\left(\mu_1 - \sum_{j\neq 1} w_1^j \mu_j\right)\right]^2 \xrightarrow{p} E\left[\epsilon_{1t} - \sum_{j\neq 1} w_1^j \epsilon_{jt} + \lambda_t\left(\mu_1 - \sum_{j\neq 1} w_1^j \mu_j\right)\right]^2 \tag{3}$$

Let $\bar{\mathbf{w}} = \{\bar{w}_1^j\}_{j\neq 1}$ be the weights that minimize this expectation and treat $\widehat{\mathbf{w}} = \{\hat{w}_1^j\}_{j\neq 1}$ as an M-estimator. We show in Appendix A.1 that $\widehat{\mathbf{w}} \rightarrow_p \bar{\mathbf{w}}$. We show now that $\bar{\mathbf{w}} \notin \Phi_1$, which implies that the SC weights will converge in probability to weights that do not satisfy the condition stated in assumption 1, even under the assumption of existence of such weights. We consider a simple case where $var(\epsilon_{it}) = \sigma_\epsilon^2$ for all i and ϵ_{it} is uncorrelated with λ_t. Let $E[\lambda_t'\lambda_t] = \Omega$ be the matrix of second moments of λ_t. Therefore, the

[10]Kaul et al. (2015) show that the weights allocated to time-invariant covariates would be zero if one uses all pre-treatment intervention outcome values as economic predictors. Therefore, we do not consider this case.

objective function simplifies to:

$$\Gamma(\{w_1^j\}_{j\neq 1}) = \sigma_\epsilon^2 \left(1 + \sum_{j\neq 1}(w_1^j)^2\right) + \left(\mu_1 - \sum_{j\neq 1} w_1^j \mu_j\right)' \Omega \left(\mu_1 - \sum_{j\neq 1} w_1^j \mu_j\right) \tag{4}$$

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. If assumption 1 is satisfied, then we can set this part equal to zero by choosing $\mathbf{w}^* \in \Phi_1$.

Consider that we start at $\{w_1^{*j}\}_{j\neq 1} \in \Phi_1$ and move in the direction of $w_1^j = \frac{1}{J}$ for all $j = 2, .., J+1$, with $w_1^j = w_1^{*j} + \Delta(\frac{1}{J} - w_1^{*j})$. 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 4 with respect to Δ at $\Delta = 0$, we have that:

$$\Gamma'(\{w_1^{*j}\}_{j\neq 1}) = 2\sigma_\epsilon^2 \left(\frac{1}{J} - \sum_{j=2}^{J+1}(w_1^{*j})^2\right) < 0 \text{ unless } w_1^{*j} = \frac{1}{J} \tag{5}$$

Therefore, $\mathbf{w}^* \in \Phi_1$ 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 satisfy assumption 1, unless it turns out that $\mathbf{w}^*$ also minimizes the variance of the idiosyncratic errors. The SC estimator will be given by:

$$\hat{\alpha}_{1t} = y_{1t} - \sum_{j\neq 1} \hat{w}_1^j y_{it} \xrightarrow{d} \alpha_{1t} + \left(\epsilon_{1t} - \sum_{j\neq 1} \bar{w}_1^j \epsilon_{jt}\right) + \lambda_t \left(\mu_1 - \sum_{j\neq 1} \bar{w}_1^j \mu_j\right) \tag{6}$$

The SC estimator will only be asymptotically unbiased if we have that $E\left[\epsilon_{1t} - \sum_{j\neq 1} \bar{w}_1^j \epsilon_{jt} | d_{1t}\right] = 0$ and $E\left[\lambda_t \left(\mu_1 - \sum_{j\neq 1} \bar{w}_1^j \mu_j\right) | d_{1t}\right] = 0$.[11] Since $\left(\mu_1 - \sum_{j\neq 1} \bar{w}_1^j \mu_j\right) \neq 0$, this implies that we cannot have selection on unobservables, even if selection is based on the common factors. 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

[11] 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.2 that these two definitions are equivalent in our setting under standard assumptions.

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 show that the bias of the SC estimator is bounded by a term that goes to zero when T_0 increases, *if the pre-treatment fit is close to perfect.* Differently from Abadie et al. (2010), we consider the asymptotic distribution of the SC estimator when $T_0 \to \infty$. Therefore, we cannot condition on a close-to-perfect pre-intervention fit, as the probability of having a close-to-perfect fit converges to zero when T_0 is large. We show that, in our setting, the SC estimator is asymptotically biased, and the bias is increasing with the variance of the transitory shocks. Note that our results are not as conflicting as they may appear at first glance. In a model with stationary 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 bias of the SC estimator should be small in situations where one would expect to have a close-to-perfect fit. We consider in MC simulations the properties of the SC estimator conditional on finding a good pre-treatment match in Section 4.

2.3 Alternative SC specifications

2.3.1 Average of pre-intervention outcome as 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:

$$\begin{aligned}
\{\hat{w}_1^j\}_{j\neq 1} &= \text{argmin}_{w\in W}\left[\frac{1}{T_0}\sum_{t=1}^{T_0}\left(y_{1t}-\sum_{j\neq 1} w_1^j y_{jt}\right)\right]^2 \\
&= \text{argmin}_{w\in W}\left[\frac{1}{T_0}\sum_{t=1}^{T_0}\left(\epsilon_{1t}-\sum_{j\neq 1} w_1^j \epsilon_{jt} + \lambda_t\left(\mu_1-\sum_{j\neq 1} w_1^j\mu_j\right)\right)\right]^2 \qquad (7)
\end{aligned}$$

where $W = \{\{w_1^j\}_{j\neq 1} \in \mathbb{R}^J | w_1^j \geq 0 \text{ and } \sum_{j\neq 1} w_1^j = 1\}$.

Therefore, assuming weakly dependence of λ_t, the objective function converges in probability to:

$$\Gamma(\{w_1^j\}_{j\neq 1}) = \left[E\left[\lambda_t\right] \left(\mu_1 - \sum_{j\neq 1} w_1^j \mu_j \right) \right]^2 \tag{8}$$

Assuming that there is a time-invariant common factor (that is, $\lambda_t^1 = 1$ for all t) and that λ_t is weakly stationary, we have that, without loss of generality, $E[\lambda_t^k] = 0$ for $k > 1$. In this case, the objective function collapses to:

$$\Gamma(\{w_1^j\}_{j\neq 1}) = \left[\left(\mu_1^1 - \sum_{j\neq 1} w_1^j \mu_j^1 \right) \right]^2 \tag{9}$$

Therefore, while we assume that there exists at least one set of weights that reproduces all factor loadings (assumption 1), the objective function will only look for weights that approximate the first factor loading. This is problematic because it might be that assumption 1 is satisfied, but there are weights $\{\tilde{w}_1^j\}_{j\neq 1} \notin \Phi_1$ that satisfy $\mu_1^1 = \sum_{j\neq 1} \tilde{w}_1^j \mu_j^1$. 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.

2.3.2 Adding other covariates as predictors

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

$$\begin{cases} y_{it}(0) = \delta_t + \theta_t Z_i + \lambda_t \mu_i + \epsilon_{it} \\ y_{it}(1) = \alpha_{it} + y_{it}(0) \end{cases} \tag{10}$$

We also modify assumption 1 so that the weights reproduce both μ_1 and Z_1.

Assumption 1′ (existence of weights):

$$\exists\ \{w_1^{*j}\}_{j\neq 1} \mid \mu_1 = \sum_{j\neq 1} w_1^{*j} \mu_j,\ Z_1 = \sum_{j\neq 1} w_1^{*j} Z_j,\ \sum_{j\neq 1} w_1^{*j} = 1, \text{ and } w_1^{*j} \geq 0$$

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:

$$\widehat{\mathbf{w}}(V) \in \text{argmin}_{\mathbf{w} \in W} ||X_1 - X_0 \mathbf{w}||_V \tag{11}$$

where $W = \{\{w_1^j\}_{j \neq 1} \in \mathbb{R}^J | w_1^j \geq 0 \text{ and } \sum_{j \neq 1} w_1^j = 1\}$. Note that the objective function of this minimization problem converges to $||\bar{X}_1 - \bar{X}_0 \mathbf{w}||_V$, where:

$$\bar{X}_1 - \bar{X}_0 \mathbf{w} = \begin{bmatrix} E[\theta_t] \left(Z_1 - \sum_{j \neq 1} w_1^j Z_j \right) + \left(\mu_1^1 - \sum_{j \neq 1} w_1^j \mu_j^1 \right) \\ \left(Z_1^1 - \sum_{j \neq 1} w_1^j Z_j^1 \right) \\ \vdots \\ \left(Z_1^R - \sum_{j \neq 1} w_1^j Z_j^R \right) \end{bmatrix} \tag{12}$$

Similarly to the case with only the average pre-intervention outcome value as economic predictor, it might be that assumption $1'$ is satisfied, but there are weights $\{\tilde{w}_1^j\}_{j \neq 1}$ that satisfy $\mu_1^1 = \sum_{j \neq 1} \tilde{w}_1^j \mu_j^1$ and $Z_1 = \sum_{j \neq 1} \tilde{w}_1^j Z_j$, although $\mu_1^k \neq \sum_{j \neq 1} \tilde{w}_1^j \mu_j^k$ for some $k > 1$. Therefore, there is no guarantee that an estimator based on this minimization problem would converge to weights that satisfy assumption $1'$ 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:

$$\widehat{\mathbf{w}} \quad = \quad \text{argmin}_{w \in \widetilde{W}} \left[\frac{1}{T_0} \sum_{t=1}^{T_0} \left(y_{1t} - \sum_{j \neq 1} w_1^j y_{jt} \right) \right]^2 \tag{13}$$

where $\widetilde{W} \subseteq W$ is the set of $\mathbf{w}$ such that $\mathbf{w}$ is the solution to problem 11 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 $1'$ 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 $1'$. Moreover, since we might have multiple solutions to problem 11, there might be no V such that $\widehat{\mathbf{w}}(V)$ converges in probability to weights in Φ_1.

Therefore, it is not possible to guarantee that this SC estimator would be asymptotically unbiased.

3 Alternatives

3.1 Demeaned SC Estimator

In contrast to the SC estimator, the DID estimator for the treatment effect in a given post-intervention period $t > T_0$, under assumption 2, would be given by:

$$
\begin{aligned}
\hat{\alpha}_{1t}^{DID} &= y_{1t} - \frac{1}{J}\sum_{j\neq 1} y_{jt} - \frac{1}{T_0}\sum_{\tau=1}^{T_0}\left[y_{1\tau} - \frac{1}{J}\sum_{j\neq 1} y_{j\tau}\right] \\
&= \epsilon_{1t} - \frac{1}{J}\sum_{j\neq 1}\epsilon_{jt} + \lambda_t\left(\mu_1 - \frac{1}{J}\sum_{j\neq 1}\mu_j\right) - \frac{1}{T_0}\sum_{\tau=1}^{T_0}\left[\epsilon_{1\tau} - \frac{1}{J}\sum_{j\neq 1}\epsilon_{j\tau} + \lambda_\tau\left(\mu_1 - \frac{1}{J}\sum_{j\neq 1}\mu_j\right)\right] \\
&\overset{d}{\to} \epsilon_{1t} - \frac{1}{J}\sum_{j\neq 1}\epsilon_{jt} + \left(\lambda_t - E\left[\lambda_\tau\right]\right)\left(\mu_1 - \frac{1}{J}\sum_{j\neq 1}\mu_j\right)
\end{aligned} \tag{14}
$$

where we assumed that the pre-intervention average for the common factors converges in probability to their unconditional means.[12]

Therefore, the DID estimator would only be asymptotically unbiased if common factors that are not constant over time are uncorrelated with treatment assignment. In this case, these common factors would enter the error term and would not cause bias because their expectation conditional on treatment status would be equal to zero. The DID model allows for selection on common factors that are constant over time. In this case, the characteristics that are correlated with treatment assignment would be captured by the unit fixed effects. Therefore, if the DID assumptions are satisfied, then the DID estimator would be unbiased while the SC estimator would be, in general, asymptotically biased.

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.[13] If common factors are stationary, this implies a model with no time-invariant common factor. We show in Appendix A.3 that the only difference relative to the original model is that the common factors $\tilde{\lambda}_t$ and factor loadings $\tilde{\mu}_i$ would not include the time-invariant common factor. Also, we can assume, without loss of generality, that $E[\tilde{\lambda}_t] = 0$. In this case, we guarantee

[12] This is guaranteed with our assumption 2 even if λ_t is correlated with the treatment assignment for t close to T_0.

[13] Doudchenko and Imbens (2016) propose a generalization of the SC method which includes an intercept parameter in the minimization problem to estimate the SC weights. As in our demeaned SC estimator, this also allows for the possibility of a systematic additive difference between treated and control units.

that the SC estimator is asymptotically unbiased when the DID assumptions are satisfied. Note that we also make assumption 1 weaker, since there might be weights that reconstruct all common factors $\tilde{\lambda}_t$ that are not constant over time, but does not match the level of the treated unit.[14] We can show that, if the DID assumption is valid, then both this demeaned SC estimator and the DID estimator will be asymptotically unbiased, but the variance of the asymptotic distribution of the demeaned SC estimator will always be weakly lower relative to the DID estimator. Let $\hat{\alpha}_{1t}^{\text{SC}'}$ be the demeaned SC estimator. Under the DID assumption, $\tilde{\lambda}_t$ and ϵ_{ij} will be independent of the fact that unit 1 was treated after T_0. Therefore, for a given for $t > T_0$, under assumption 2, the variance of the asymptotic distribution of the SC estimator would be given by:

$$a.var(\hat{\alpha}_{1t}^{\text{SC}'} - \alpha_{1t}) = E\left[\left(\tilde{\epsilon}_{1t} - \sum_{j\neq 1}\bar{w}_1^j \tilde{\epsilon}_{jt}\right) + \tilde{\lambda}_t\left(\tilde{\mu}_1 - \sum_{j\neq 1}\bar{w}_1^j\tilde{\mu}_j\right)\right]^2 \tag{15}$$

while:

$$a.var(\hat{\alpha}_{1t}^{\text{DID}} - \alpha_{1t}) = E\left[\left(\tilde{\epsilon}_{1t} - \sum_{j\neq 1}\frac{1}{J}\tilde{\epsilon}_{jt}\right) + \tilde{\lambda}_t\left(\tilde{\mu}_1 - \sum_{j\neq 1}\frac{1}{J}\tilde{\mu}_j\right)\right]^2 \tag{16}$$

Since the DID weights belong to W and the demeaned SC weights converge in probability to weights that minimize the function $E\left[\left(\epsilon_{1t} - \sum_{j\neq 1} w_1^j \epsilon_{jt}\right) + \tilde{\lambda}_t\left(\mu_1 - \sum_{j\neq 1} w_1^j\tilde{\mu}_j\right)\right]^2$, it must be that $a.var(\hat{\alpha}_{1t}^{\text{SC}'} - \alpha_{1t}) \leq a.var(\hat{\alpha}_{1t}^{\text{DID}} - \alpha_{1t})$. Note that this result is valid even if assumption 1 does not hold.

If the correlation comes from common factors that are not constant over time and assumption 1 is satisfied, then the bias of the SC estimator would usually be lower than the bias of the DID estimator. We show in Section 4 a particular class of linear factor models in which the asymptotic bias of the demeaned SC estimator will always be lower. However, we provide in Appendix A.4 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.

3.2 IV-Like SC Estimator

We also propose an alternative way of estimating the SC weights that provide consistent estimators if we impose additional assumptions on the common factors and transitory shocks. Note that the asymptotic bias of the SC estimator derived in Section 2.2 comes from the first step of the SC method in which one estimates the SC weights using the pre-treatment information. As noted by Doudchenko and Imbens (2016),

[14]Note that if assumption 1 is valid for the original model, then it will also be valid for the demeaned model.

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 $\{w_1^{*j}\}_{j\neq 1} \in \Phi_1$, we can write:

$$y_{1t} = \sum_{j=1}^{J+1} w_1^{*j} y_{jt} + \eta_t, \text{ for } t \leq T_0$$

where:

$$\eta_t = \epsilon_{1t} - \sum_{j=1}^{J+1} w_1^{*j} \epsilon_{jt}$$

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, $\{w_1{}^j\}_{j\neq 1} \in \Phi_1$.

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:

$$\begin{aligned} \mathbf{y}_{t-1}(y_{1t} - \mathbf{y}_t'\mathbf{w}) &= (\mu_0'\lambda_{t-1}' + \epsilon_{t-1})\lambda_t\,(\mu_1 - \mu_\mathbf{0}\mathbf{w}) + (\mu_0'\lambda_{t-1}' + \epsilon_{t-1})(\epsilon_{1t} - \epsilon_t'\mathbf{w}) \qquad (17) \\ &= \mu_0'\lambda_{t-1}'\lambda_t\,(\mu_1 - \mu_\mathbf{0}\mathbf{w}) + \epsilon_{t-1}\lambda_t\,(\mu_1 - \mu_\mathbf{0}\mathbf{w}) + \mu_0'\lambda_{t-1}'(\epsilon_{1t} - \epsilon_t'\mathbf{w}) + \epsilon_{t-1}(\epsilon_{1t} - \epsilon_t'\mathbf{w}) \end{aligned}$$

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

$$E\left[\mathbf{y}_{t-1}(y_{1t} - \mathbf{y}_t'\mathbf{w})\right] = \mu_0' E\left[\lambda_{t-1}'\lambda_t\right](\mu_1 - \mu_\mathbf{0}\mathbf{w}) \qquad (18)$$

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, $\mathbf{w} \in \Phi_1$. 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.[15] 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

[15] 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.

to generate an IV-like SC estimator that is asymptotically unbiased. A possible limitation of this method is that it might rely on a very large number of pre-treatment periods so that weights are close to weights that satisfy assumption 1. Results from MC simulations available upon request.

4 Asymptotic vs. Finite T_0 Properties of the SC Estimator

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.[16] More specifically, we consider that the $J+1$ units are divided into K groups, where for each j we have that:

$$y_{jt}(0) = \delta_t + \lambda_t^k + \epsilon_{jt} \tag{19}$$

for some $k = 1, ..., K$. We start considering the asymptotic properties of the SC estimator in this particular class of linear factor models when $T_0 \rightarrow \infty$. Then we consider the properties of the SC estimator in the setting proposed in Abadie et al. (2010), where T_0 is fixed and the analysis is conditional on a good pre-intervention fit.

4.1 Asymptotic Results

We start considering the case in which the model is stationary. 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. Assume that $var(\lambda_t^k) = 1$ and $var(\epsilon_{jt}) = \sigma_\epsilon^2$. 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 4, we have that, when $T_0 \rightarrow \infty$, the proportion of misallocated weights converges to:

$$\gamma_2(\sigma_\epsilon^2, J) = \sum_{j=\frac{J+1}{2}+1}^{J+1} \bar{w}_1^j = \frac{J+1}{J^2 + 2 \times J \times \sigma_\epsilon^2 - 1} \times \sigma_\epsilon^2 \tag{20}$$

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

We present in Figure 1.A the relationship between asymptotic misallocation of weights, variance of the

[16] Monte Carlo simulations using this model was studied in detail in Ferman et al. (2016).

transitory shocks, and number of control units. Note that, for a fixed J, the proportion of misallocated weights converges to zero when $\sigma^2_\epsilon \rightarrow 0$, while this proportion converges to $\frac{J+1}{2J}$ (the proportion of misallocated weights of DID) when $\sigma^2_\epsilon \rightarrow \infty$. This is consistent with the results we have in Section 2.2. 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 (2013), 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:

$$\hat{\alpha}_{1t} \xrightarrow{d} \alpha_{1t} + \left(\epsilon_{1t} - \sum_{j \neq 1} \bar{w}_1^j \epsilon_{jt} \right) + \lambda_t^1 \times \gamma_2(\sigma^2_\epsilon, J) - \lambda_t^2 \times \gamma_2(\sigma^2_\epsilon, J) \tag{21}$$

Therefore, if $E[\lambda_t^1 | d_{1t} = 1] = 1$ (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 asymptotic bias of the SC estimator as a function of σ^2_ϵ and J in Figure 1.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_t^1. 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:

$$\gamma_{\frac{J+1}{2}}(\sigma^2_\epsilon, J) = \sum_{j=2}^{J+1} \bar{w}_1^j = \frac{J-1}{2 + \sigma^2_\epsilon + (1+\sigma^2_\epsilon)(J-1)} \times \sigma^2_\epsilon \tag{22}$$

We present the relationship between misallocation of weights, variance of the transitory shocks, and number of control units in Figure 1.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 (2013) are not satisfied. As presented in Figure 1.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.4.

4.2 Finite T_0 results - stationary process

The results presented in Section 4.1 are based on the setting studied in this paper in which we consider $T_0 \rightarrow \infty$. 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 19. We consider in our MC simulations $J+1 = 20$, λ_t^k normally distributed following an AR(1) process with 0.5 serial correlation parameter, $\epsilon_{jt} \sim N(0, \sigma_\epsilon^2)$, and $T - T_0 = 10$. We also impose that there is no treatment effect, i.e., $y_{jt} = y_{jt}(0) = y_{jt}(1)$ for each time period $t \in \{1, ..., T\}$. 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: $\sigma_\epsilon^2 \in \{0.1, 0.5, 1\}$.
- The number of groups with different λ_t^k: $K = 2$ (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. While Abadie et al. (2015) recommend that the SC method should not be used if the pre-treatment if is poor, they do not provide an objective rule to determine whether one should consider that the pre-treatment fit is good. We determine that the SC estimator provided a good fit if the R^2 of a regression of the pre-treatment outcomes of the treated unit on the pre-treatment outcomes of the SC unit is greater than $\bar{r}$. 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 4.1, 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 asymptotic values. The proportion of misallocated weights is substantially higher when T_0 is very 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 an $R^2 > 0.8$. 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, which implies that 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 3.4% and 0.7%. 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 38% when $\sigma^2_\epsilon = 0.5$ and from 66% to 54% 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 $R^2 > 0.9$. 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 3.1, we also know that, if the DID identification assumption is satisfied, 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 this model as well. We also 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 $\sigma^2_\epsilon = 0.1$. When σ^2_ϵ 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^2_ϵ 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 in situations where 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.

4.3 Finite T_0 results - non-stationary process

So far we have considered stationary models. However, the applications in Abadie and Gardeazabal (2003), Abadie et al. (2010), and Abadie et al. (2015) are clearly non-stationary. In this Section, we consider a model in which a subset of the common factors is non-stationary. We consider the following DGP:

$$y_{jt}(0) = \delta_t + \lambda_t^k + \phi_t^r + \epsilon_{jt} \tag{23}$$

for some $k = 1, ..., K$ and $r = 1, ..., R$. We maintain that λ_t^k is normally distributed following an AR(1) process with 0.5 serial correlation parameter, while ϕ_t^r follows a random walk. We consider in our simulations $K = 10$ and $R = 2$. Therefore, units $j = 2, ..., 10$ follow the same non-stationary path ϕ_t^1 as the treated unit, although only unit $j = 2$ also follows the same stationary path λ_t^1 as the treated unit.

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 non-stationary 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 non-stationary trend dominates 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 non-stationary (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 non-stationary factors. In this scenario, one would usually have a close-to-perfect fit, and there would be virtually no bias associated to the non-stationary factors. However, these results also suggest that, in a non-stationary model, we might have a substantial misallocation of weights for the stationary 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 scenarios with non-stationary common factors, as the SC estimator is extremely efficient in capturing the non-stationary factors. Also, if the DID and SC estimators are unbiased, then the variance of the DID 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 non-stationary factors. Still, this potential bias of the SC estimator due to a correlation between treatment assignment and the stationary common shocks would be lower than the bias of DID.

5 Permutation Tests

Abadie et al. (2010) argue that large sample inferential techniques are not well suited to comparative case studies when the number of units in the comparison group is small.[17] They propose a permutation test where they apply the SC method to every potential control in the sample. First, they consider a graphical analysis where they compare the post-treatment prediction error of the SC estimator with the prediction error for each of SC placebo estimator. Then they consider whether the prediction error when one considers the actual treated unit is "unusually" large relative to the distribution of prediction errors for the units in the donor pool. They also suggest a permutation test using the post/pre-treatment mean squared prediction error (MSPE) as test statistic. Abadie et al. (2010) and Abadie et al. (2015) recognize that the assumptions required in the classical randomization inference setting (in particular, random treatment assignment) are rather restrictive in the SC setting. Still, they argue that it is possible to interpret the p-value from their permutation test as *"the probability of obtaining an estimate at least as large as the one obtained for the unit representing the case of interest when the intervention is reassigned at random in the data set"* (Abadie et al. (2015), page 500). While we agree that this interpretation of the permutation test p-value is useful, it is important to consider the statistical properties of the test. We start considering the asymptotic properties of the permutation test, and then we provide evidence from MC simulations.

[17]Carvalho et al. (2015) and Powell (2016) rely on large sample inferential techniques. Instead of testing the null hypothesis of no effect for all post-treatment periods, they test whether the average effect across time is equal to zero. If both the number of pre- and post-intervention periods is large, then they are able to derive the asymptotic distribution of the estimator. This method would not work if one wants to test the null of no effect for all post-treatment periods or if the number of post-intervention periods is finite.

5.1 Asymptotic Properties

Note that the graphical analysis suggested in Abadie et al. (2010) does not provide a clear decision rule on whether the null hypothesis should be rejected. Still, this analysis would implicitly reject the null when the post-intervention mean squared prediction error (MSPE) for the SC estimate is greater than the post-intervention MSPE for the placebo estimates. We consider the post-intervention MSPE as the test statistic:

$$t_i^{\text{post}} = \frac{1}{T-T_0} \sum_{t=T_0+1}^{T} \left[y_{it} - \sum_{j \neq i} \hat{w}_i^j y_{jt} \right]^2 \tag{24}$$

We start assuming that assumption 1 is valid for all i.[18] In particular, we assume that:

$$\exists \; \{w_i^{*j}\}_{j \neq i} \mid \mu_i = \sum_{j \neq i} w_i^{*j} \mu_j, \; \sum_{j \neq 1} w_i^{*j} = 1, \text{ and } {w^*}_i^j \geq 0 \; \forall i = 1, ..., J+1 \tag{25}$$

Therefore, if we assume that the estimator of the SC weights $\hat{\mathbf{w}}_i \xrightarrow{p} \mathbf{w}_i^* \in \Phi_i$, then for all i we will have that:

$$t_i^{\text{post}} \xrightarrow{d} \frac{1}{T-T_0} \sum_{t=T_0+1}^{T} \left[\epsilon_{it} - \sum_{j \neq i} {w^*}_i^j \epsilon_{jt} \right]^2 \tag{26}$$

where $\Phi_i = (\{w_i^j\}_{j \neq i} \mid \mu_i = \sum_{j \neq i} w_1^j \mu_j, \; \sum_{j \neq i} {w^*}_i^j = 1, \text{ and } {w^*}_i^j \geq 0)$.

There are at least three reasons why this test statistic might not have the same (marginal) asymptotic distribution for all permutations. First, the transitory shock might be heteroskedastic. Ferman and Pinto (2016) show that this would usually be true if we have unit x time aggregate values when there is variation in the number of observations per unit. This would be the case, for example, if one uses the Current Population Survey (CPS). Note that, in this case, t_i^{post} would tend to attain higher values when the treated unit is small relative to the units in the donor pool. Second, even if the transitory shock is homoskedastic, the variance of $\epsilon_{it} - \sum_{j \neq i} {w^*}_i^j \epsilon_{jt}$ will depend on the weights $\{{w^*}_i^j\}_{j \neq i}$. If the weights for unit i are more concentrated around a few units in the donor pool, then the variance of t_i^{post} should be higher than if the weights were more evenly distributed. Finally, t_i^{post} would not have the same distribution as t_1^{post} if assumption 1 is not

[18] Ando and Sävje (2013) argue that in most applications it would not be reasonable to assume that this assumption is valid for all i. We believe that this condition might be reasonable in some applications. For example, this condition is satisfied if we have different groups of units where time trends are different across groups but parallel within groups, as considered in Ferman et al. (2016). We analyze this case in detail in Section 4. In this case, the main idea of the SC estimator would be to select the control units that follow the same time trend as the treated unit. We consider below the implications in case assumption 1 is not valid for all i.

valid for unit i or if the SC weights converge in probability to weights that do not satisfy assumption 1. In this case, we would have that $y_{it} - \sum_{j \neq i} \hat{w}_i^j y_{jt} \xrightarrow{d} \epsilon_{it} - \sum_{j \neq i} \bar{w}_i^j \epsilon_{jt} + \lambda_t \left(\mu_i - \sum_{j \neq i} \bar{w}_i^j \mu_j \right)$. Hahn and Shi (2016) provide MC simulations showing that a permutation test using t_i^{post} as test statistic may severely over-reject under the null, even if one uses an infeasible SC estimator that relies on weights that correctly reconstruct the factor loadings of the treated unit.

Abadie et al. (2010) correctly noticed that the outcome variable may not be well reproduced for some units by a convex combination of the other units for the pre-intervention periods, and that the post-intervention MSPE for these units should be high as well. For this reason, they exclude permutations in which the pre-intervention MSPE is 20 times (or 5 times) larger than the pre-intervention MSPE for the treated unit. Note that, if assumption 1 is satisfied and $\hat{\mathbf{w}}_i \xrightarrow{p} \mathbf{w}_i^* \in \Phi_i$ for all i, then the prediction error would converge to $\epsilon_{it} - \sum_{j \neq i} {w^*}_i^j \epsilon_{jt}$ as $T_0 \to \infty$ whether time t is pre- or post-intervention. Therefore, assuming that ϵ_{it} is stationary, then it would be likely that, in our setting, t_i^{post} has the same asymptotic marginal distribution as t_1^{post} if the pre-intervention MSPE for unit i and unit 1 are similar. Note, however, that Abadie et al. (2010) procedure only excludes permutations with pre-intervention MSPE *higher* than the pre-intervention MSPE for the treated unit. Therefore, if there are many permutations with lower pre-intervention MSPE, then the test would over-reject the null since t_1^{post} would tend to attain larger values. In this case, Abadie et al. (2010) graphical analysis could be misleading, even if the SC weights converge in probability to weights that satisfy assumption 1 for all units.

A second inference procedure suggested by Abadie et al. (2010) is a permutation test using the ratio of post/pre-intervention MSPE (t_i^{ratio}). According to them, *"the main advantage of looking at ratios is that it obviates choosing a cut-off for the exclusion of ill-fitting placebo runs"*. Ando and Sävje (2013) argue that the distribution of this test statistic would not have the same distribution for all permutations. However, they do not consider the asymptotic distribution when $T_0 \to \infty$. Assuming that $T_0 \to \infty$ and that $T - T_0$ is fixed, then:

$$
\begin{aligned}
t_i^{\text{ratio}} &= \frac{\frac{1}{T-T_0} \sum_{t=T_0+1}^{T} \left[y_{it} - \sum_{j \neq i} \hat{w}_i^j y_{jt} \right]^2}{\frac{1}{T_0} \sum_{t=1}^{T_0} \left[y_{it} - \sum_{j \neq i} \hat{w}_i^j y_{jt} \right]^2} \\
&\xrightarrow{d} \frac{1}{T-T_0} \sum_{t=T_0+1}^{T} \left[\frac{\epsilon_{it} - \sum_{j \neq i} \bar{w}_i^j \epsilon_{jt} + \lambda_t (\mu_i - \sum_{j \neq i} \bar{w}_i^j \mu_j)}{\sqrt{var(\epsilon_{it} - \sum_{j \neq i} \bar{w}_i^j \epsilon_{jt} + \lambda_t (\mu_i - \sum_{j \neq i} \bar{w}_i^j \mu_j))}} \right]^2 \qquad (27)
\end{aligned}
$$

where $\widehat{w}_i^j \xrightarrow{p} \bar{w}_i^j$.

Equation 27 makes it clear that, if the SC estimator is asymptotically biased (that is, $\mu_i \neq \sum_{j \neq i} \bar{w}_i^j \mu_j$ and treatment assignment is correlated with the unobserved heterogeneity), then the expected value of the asymptotic distribution of t_1^{ratio} should be higher than the expected value of the asymptotic distribution of t_i^{ratio} for $i > 1$. This would lead to over-rejection.

If the SC estimator is unbiased, then t_i^{ratio} will have the same asymptotic (marginal) distribution for all i if $Q_{it} = \epsilon_{it} - \sum_{j \neq i} \bar{w}_i^j \epsilon_{jt} + \lambda_t(\mu_i - \sum_{j \neq i} \bar{w}_i^j \mu_j)$ is stationary, serially uncorrelated, and i.i.d. across i up to a scale parameter. If we assume that $\frac{E[Q_{it}^4]}{(E[Q_{it}^2])^2}$ is constant, and still maintain that errors are serially uncorrelated and stationary, then the test statistic has, asymptotically, the same expected value and variance for all permutations. If we also relax the assumption that errors are serially uncorrelated, then we can construct a test statistic $\tilde{t}_i$ that has asymptotically the same expected value and variance for all permutations. Define $S_i = \frac{1}{T-T_0} \sum_{t=T_0+1}^{T} \left[\epsilon_{it} - \sum_{j \neq i} \bar{w}_i^j \epsilon_{jt} + \lambda_t(\mu_i - \sum_{j \neq i} \bar{w}_i^j \mu_j)\right]^2$. We can use:

$$\tilde{t}_i = \frac{\frac{1}{T-T_0} \sum_{t=T_0+1}^{T} \left[y_{it} - \sum_{j \neq i} \hat{w}_i^j y_{jt}\right]^2 - \widehat{E[S_i]}}{\sqrt{\widehat{var(S_i)}}} \tag{28}$$

where $\widehat{E[S_i]}$ is an estimator for $E[S_i]$ and $\widehat{var(S_i)}$ is an estimator for $var[S_i]$. With large T_0, we can construct a new time serie $S_{it} = \frac{1}{T-T_0} \sum_{t'=t}^{t+T-T_0} \left[y_{it'} - \sum_{j \neq i} \hat{w}_i^j y_{jt'}\right]^2$ using the pre-treatment periods and calculate $\widehat{E[S_i]}$ and $\widehat{var(S_i)}$. In Appendix A.6, we provide conditions such that these are consistent estimators, and show that, in this case, the asymptotic distribution of $\tilde{t}_i$ has expected value equal to zero and variance equal to 1 for all permutations.

The test statistics t_i^{ratio} and $\tilde{t}_i$ help prevent that test statistics for different permutations have wildly different asymptotic (marginal) distributions, which could generate severe size distortions. However, it is important to note that, even if the test statistics for all permutations have the same asymptotic (marginal) distributions, it is not possible to guarantee that the permutation test is asymptotically valid. Following Canay et al. (2014), a permutation test would be asymptotically valid if the test statistics are based on a function of the data that exhibits approximate symmetry. In the SC setting, this will not generally be the case, because the SC estimator is a function of transitory shocks of the treated and control units, which induces correlation between test statistics in different permutations. With fixed J, this correlation will not vanish, even when $T_0 \to \infty$, as noticed in Powell (2016). We show in Section 5.2 examples in which the test statistics can have the same asymptotic distribution for all permutations, but we still can have size distortions. We also consider in Section 5.2 the finite T_0 properties of the permutation test.

5.2 Examples and Monte Carlo Simulations

We consider first examples in which the permutation test might have size distortions even though the test statistics have the same asymptotic marginal distributions for all permutations. We assume first that we know $\mathbf{w}_i \in \Phi_1$ and that we know the variance of $y_{it} - \sum_{j \neq i} w_i^j y_{jt}$. Therefore, the size distortions we present in these examples are not related to our result in Section 2.2 that the SC weights converge to weights that do not reconstruct the factor loadings of the treated unit, nor by the fact that weights are estimated with a finite T_0.[19] Then we show that the permutation test using the actual SC estimator can have important size distortions when T_0 is small and/or $\widehat{\mathbf{w}}_i$ is estimated, even when the asymptotic size distortions of the infeasible SC estimator are negligible.

Consider first a model with two common factors, $\lambda_t = (\lambda_t^1, \lambda_t^2)$, where $\mu_i = (1,0)$ for $i = 1,2,3$ and $\mu_i = (0,1)$ for $i = 4,...,20$. Assume also that $\epsilon_{it} \overset{i.i.d.}{\sim} N(0,1)$ for all i and t. An (infeasible) SC estimator for the treatment effect at time t in this model for units $i = 1,2,3$ uses the average of the other 2 units that have the same factor loadings to construct the SC estimator, while for units $i = 4,...,20$ it uses the average of the the other 16 units that have the same factor loadings. Now consider the vector $\left(\frac{\hat{\alpha}_1}{\sqrt{var(\hat{\alpha}_1)}}, \frac{\hat{\alpha}_2}{\sqrt{var(\hat{\alpha}_2)}}, ..., \frac{\hat{\alpha}_{20}}{\sqrt{var(\hat{\alpha}_{20})}} \right)'$, where $\hat{\alpha}_j$ is the SC estimator using unit j as treated. For all i, $\frac{\hat{\alpha}_i}{\sqrt{var(\hat{\alpha}_i)}} \sim N(0,1)$. However:

$$cov\left(\frac{\hat{\alpha}_i}{\sqrt{var(\hat{\alpha}_i)}}, \frac{\hat{\alpha}_k}{\sqrt{var(\hat{\alpha}_k)}} \right) = \begin{cases} -0.5 \text{ if } i \in \{1,2,3\} \text{ and } k \in \{1,2,3\}/\{i\} \\ 0 \text{ if } i \in \{1,2,3\} \text{ and } k \in \{4,...,20\} \\ -0.06 \text{ if } i \in \{4,...,20\} \text{ and } k \in \{4,...,20\}/\{i\} \end{cases} \tag{29}$$

Therefore, while all elements in this vector have the same marginal distribution, the conditional distributions are not the same for all permutations. This implies a mild under-rejection of 4.3% for a 5% test when we consider unit 1 as treated.[20] Intuitively, this happens because the high correlation between $\frac{\hat{\alpha}_1}{var(\hat{\alpha}_1)}$ and $\frac{\hat{\alpha}_2}{var(\hat{\alpha}_2)}$ implies that, when $\frac{\hat{\alpha}_1}{var(\hat{\alpha}_1)}$ is extreme, the realization of $\frac{\hat{\alpha}_2}{var(\hat{\alpha}_2)}$ is likely to be extreme as well. On the contrary, when a realization of $\frac{\hat{\alpha}_i}{var(\hat{\alpha}_i)}$ for $i > 3$ is extreme, it does not imply that the realizations of other $\frac{\hat{\alpha}_k}{var(\hat{\alpha}_k)}$ for $k \neq i$ are likely to be extreme as well. Of course, if we do have random assignment, then the permutation test would still have the correct size for *unconditional* tests. However, if the probability

[19] Hahn and Shi (2016) also consider size distortions with an infeasible SC estimator. However, they consider in their simulations a test statistic that does not have the same distribution for all permutations.

[20] This rejection rate was calculated based on 10.000.000 MC simulations.

that a unit with the same characteristics as unit 1 is more likely to receive treatment, then we would have under-rejection.

We now show another example in which heteroskedasticity can also generate size distortions, even if the linear factor structure is symmetric. Assume now that we have 20 units in total. We have 5 common factors $\lambda_t = (\lambda_t^1, ..., \lambda_t^5)$, and $\mu_i = (1, 0, 0, 0, 0)$ for units $i = 1, ..., 4$, $\mu_i = (0, 1, 0, 0, 0)$ for units $i = 5, ..., 8$, and so on. Consider that $var(\epsilon_{1t}) = \sigma^2$ and $var(\epsilon_{it}) = 1$ for all $i > 1$. We calculate the infeasible SC estimator $\hat{\alpha}_i$ as the minimum variance estimator such that $\mathbf{w}_i \in \Phi_i$.[21] In this case, a higher σ^2 implies a lower correlation between $\hat{\alpha}_1$ and $\hat{\alpha}_i$ for $i \in \{2, 3, 4\}$. This happens because, when σ^2 is higher, then the SC estimator $\hat{\alpha}_i$ for $i \in \{2, 3, 4\}$ will assign lower weights for y_{1t}. If $\sigma^2 = 2$, then rejection rate is 5.3%, while rejection rate is 5.5% if $\sigma^2 = 5$. If $\sigma^2 < 1$, then we increase the correlation between $\hat{\alpha}_1$ and $\hat{\alpha}_i$ for $i \in \{2, 3, 4\}$. If $\sigma^2 = 0.5$, then rejection rate is 4.6%, while if $\sigma^2 = 0.1$, then rejection rate is 4%. These results suggest that heteroskedasticity can generate size distortions in the permutation test even when the marginal distributions of the test statistics are the same for all permutations. However, based on our examples, size distortions are relatively mild even if we consider a highly heteroskedastic model.

The examples so far provided evidence that we can have some size distortions even if we consider an infeasible test statistic. We now consider MC simulations in case the SC weights are estimated using a fixed T_0. We explore a model where $J + 1 = 20$ units are divided into 10 groups of 2 units each. Note that the permutation test would work in this case if we use $\mathbf{w}_i \in \Phi_1$ and the variance of $y_{it} - \sum_{j \neq i} w_i^j y_{jt}$ is known, even if the transitory shocks are heteroskedastic.[22]

We consider first the size of a permutation test when treatment assignment is correlated with the unobserved heterogeneity. In this case, the SC estimator is asymptotically biased, as the SC weights do not completely reconstruct the factor loadings of the treated unit. In columns 1 to 3 of Table 5 we present rejection rates in a stationary model when $E[\lambda_t^1 | d_{1t} = 1] = 1$, while in columns 4 to 6 we present rejection rates when $E[\lambda_t^1 | d_{1t} = 1] = 2$. As expected, the permutation test over-rejects the null, as the expected value of the test statistic is higher for the treated unit. Interestingly, we find the largest over-rejection when $\sigma_\epsilon^2 = 0.1$, in which case we found that the misallocation of weights (and, therefore, the asymptotic bias) should be relatively lower. This happens because, while the bias is lower in this case, the variance of the SC estimator is also lower. We present in columns 7 to 12 of Table 5 the same results for the non-stationary

[21] In this case, if $i = 1$ or $i > 4$, then we construct the SC unit as the simple average of the other units that have the same factor loading as the treated unit. If $i \in \{2, 3, 4\}$, then we construct the SC unit assigning weight equal to $\frac{1}{2\sigma^2+1}$ for unit 1 and $\frac{\sigma^2}{2\sigma^2+1}$ for the other two units.

[22] In this case, the infeasible SC estimator is equal to $y_{it} - y_{i't}$, where i' is the pair that follows the same parallel trend as i. Therefore, for all i, the correlation between i and j will be equal to one if j is the pair of j, and zero otherwise.

model. The permutation test still over-rejects the null, but not as much as in the stationary model. The reason is that the variance of the SC estimator is higher in the non-stationary model, due to the small discrepancy in the factor loadings of the treated and SC units associated with the non-stationary common factor. Overall, these results suggest that, when the SC estimator is biased, then the permutation test can over-reject the null even when the bias of the SC estimator is relatively small.

We consider next whether heteroskedasticity can generate size distortions in this model. In this case, we consider the same model where $J+1=20$ units are divided into 10 groups of 2 units each, but we set the variance of the transitory shocks of the treated unit equal to 0.1, while the variance of the transitory shocks of the control units is equal to one. We present in column 1 of Table 6 rejection rates when transitory shocks and common factors are serially uncorrelated, using the test statistic proposed in Abadie et al. (2010) (t_i^{ratio}). With $T_0 = 1000$, rejection rate is around 5%. This was expected given that, with serially uncorrelated transitory shocks and common factors, t_i^{ratio} would have the same asymptotic marginal distribution for all permutations.[23] With finite T_0, however, our simulation results suggest that the size distortion can actually be relevant even if the common factors are serially uncorrelated. We over-reject the null when the treated unit has a lower variance. With a finite T_0, the distribution of t_i^{ratio} is given by:

$$t_i^{\text{ratio}} = \frac{\frac{1}{T-T_0}\sum_{t=T_0+1}^{T}\left[\epsilon_{it} - \sum_{j\neq i}\widehat{w}_i^j\epsilon_{jt} + \lambda_t(\mu_i - \sum_{j\neq i}\widehat{w}_i^j\mu_j)\right]^2}{\frac{1}{T_0}\sum_{t=1}^{T_0}\left[\epsilon_{it} - \sum_{j\neq i}\widehat{w}_i^j\epsilon_{jt} + \lambda_t(\mu_i - \sum_{j\neq i}\widehat{w}_i^j\mu_j)\right]^2} \tag{30}$$

While both numerator and denominator of the test statistic depend on a linear combination of common and transitory shocks, the weights $\widehat{w}_i^j$ are chosen as to minimize the denominator. If T_0 is not large enough relative to J, we might "over-fit" the model. As a consequence, the denominator (in-sample prediction error) would not provide an adequate correction for the variance of the numerator (out-of-sample prediction error), so the marginal distribution of the test statistic would depend on the variance of the treated unit. One possible solution to this problem is to use pre-treatment periods not used in the estimation of the SC weights in the denominator. However, this implies not using all pre-treatment outcome lags as economic predictors exactly when T_0 is small. Also, the variance of the denominator should be large if one leaves out only a few pre-treatment lags, which would imply in a test with low power. Another possible solution might be to avoid over-fitting using a different method to estimate the SC weights that takes into account the fact that the number of parameters might be large relative to the number of pre-treatment periods. Doudchenko

[23] Differently from the infeasible SC estimator, the actual SC estimator will not assign 100% of the weight to the pair of the treated unit, even when $T_0 \to \infty$. Therefore, there is no guarantee the the permutation test is asymptotically valid even in this case. Still, our MC simulations suggest that asymptotic size distortions are negligible *for this particular DGP*.

and Imbens (2016) consider the use of regularization methods such as best subset regression or LASSO to estimate the SC weights.

We present in column 3 of Table 6 rejection rates when common factors follow an AR(1) process with serial correlation equal to 0.9. In this case, the test statistic t_i^{ratio} does not have the same asymptotic marginal distribution for all permutations. This implies an over-rejection even when T_0 is large. In this case, an alternative test statistic, $\tilde{t}_i$, that properly corrects the *marginal* distributions of the test statistics when $T_0 \to \infty$ provides rejection rates close to 5% when T_0 is large (column 4 of Table 6). With finite T_0, however, we have over-rejection when the treated unit has a lower variance, whether we use t_i^{ratio} or $\tilde{t}_i$, as in the case with serially uncorrelated common factors. The results using the non-stationary DGP are qualitatively similar (columns 5 to 8 of Table 6).

Finally, note that Abadie et al. (2010) and Abadie et al. (2015) suggest that the SC estimator should not be used if the pre-treatment fit is poor. However, when they recommend the permutation test using the t_i^{ratio} test statistic, they suggest that all permutations should be considered. In other words, t_1^{ratio} is conditional on a good pre-treatment fit, while t_i^{ratio} for $i > 1$ is unconditional. We evaluate now whether this might generate size distortions. We consider an homoskedastic model in which the SC estimator is asymptotically unbiased (that is, treatment assignment is uncorrelated with common factors). Note that this model is consistent with random assignment of the treated unit. The only difference is that we will only consider simulations in which the pre-treatment fit for the actual SC estimator is good. We present in Table 7 rejection rates conditional on a good pre-treatment fit for the treated unit. We also present in this table the probability of having a good pre-treatment match. The results suggest that the test may over-reject when the probability of finding a good match is not high. As an extreme example, if we set a threshold for good fit as $R^2 > 0.9$ and look at the $(T_0, \sigma_\epsilon^2) = (20, 0.1)$ case, then we would have a probability of 20% of having a good pre-treatment fit, and we would have a rejection rate of 9% for a 5% test if we consider only SC estimators that provided a good pre-treatment fit. If the probability of having a good fit is close to one (which is usually the case in the non-stationary model), then over-rejection is very mild.

6 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. This is different from the setting analyzed in Abadie et al. (2010) and Abadie et al. (2015), as they consider properties of the SC estimator with T_0 fixed and conditional on a good pre-treatment fit. If the model is stationary, we show that, in our setting, the SC estimator is asymptotically biased 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 are consistent with the results in Abadie et al. (2010). Still, we find that the SC estimator can improve relative to the DID estimator even if the pre-treatment fit is not close to perfect. However, in this case the method would rely on stronger identification assumptions. If the model has a non-stationary component, then we show that the SC method is extremely efficient in reconstructing the factor loadings associated with the non-stationary factors, and that the method will usually provide a good pre-treatment match if T_0 is moderate. However, the method might fail to reconstruct the factor loadings associated with the stationary factors for the treated unit, even if we condition on a close-to-perfect pre-treatment match. Therefore, the method would be approximately unbiased if treatment assignment is correlated with non-stationary factors, although it may be severely biased if it is correlated with stationary factors. We also show that a slight modification in the permutation test proposed in Abadie et al. (2010) can generally improve relative to DID both in terms of bias and variance.

Overall, our results suggest that the SC method substantially improves relative to the DID estimator. However, researchers should be more careful in interpreting the identification assumptions required for this method. If the pre-treatment fit is poor in a stationary model, then the identification assumption in a demeaned SC estimator is essentially the same as in the DID estimator. If the model has both non-stationary and stationary factors, then the method allows only for correlation between treatment assignment and the non-stationary factors, even if we conditional on a close-to-perfect pre-treatment fit.

Finally, we consider the statistical properties of the permutation test proposed in Abadie et al. (2010). We show that the permutation test might exhibit size distortions even if the test statistics in all permutations have the same asymptotic marginal distribution. We provide examples in which we can have size distortions even when we consider an infeasible SC estimator that correctly reconstructs the factor loadings of the treated unit. While the size distortions we find in these examples are relatively small, further research is necessary to determine whether there might be examples in which size distortions could be more severe, or whether there is a bound to the size distortions we might have in the SC permutation test. We also show that heteroskedasticity can generate important size distortions when the number of pre-treatment periods is not large.

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 Imbens, "The State of Applied Econometrics-Causality and Policy Evaluation," *arXiv preprint arXiv:1607.00699*, 2016.

Canay, Ivan A., Joseph P. Romano, and Azeem M. Shaikh, "Randomization Tests under an Approximate Symmetry Assumption?," 2014.

Carvalho, Carlos V., Ricardo Mansini, and Marcelo C. Medeiros, "ArCo: An Artificial Counterfactual Approach for Aggregate Data," February 2015. 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, "Inference in Differences-in-Differences with Different Group Sizes," 2016. Working Paper.

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

Firpo, Sergio and Vitor Possebom, "Synthetic Control Estimator: A Generalized Inference Procedure and Confidence Sets," April 2016.

Gobillon, Laurent and Thierry Magnac, "Regional Policy Evaluation: Interactive Fixed Effects and Synthetic Controls," PSE Working Papers halshs-00849071, HAL July 2013.

Hahn, Jinyong and Ruoyao Shi, "Synthetic Control and Inference," 2016.

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.

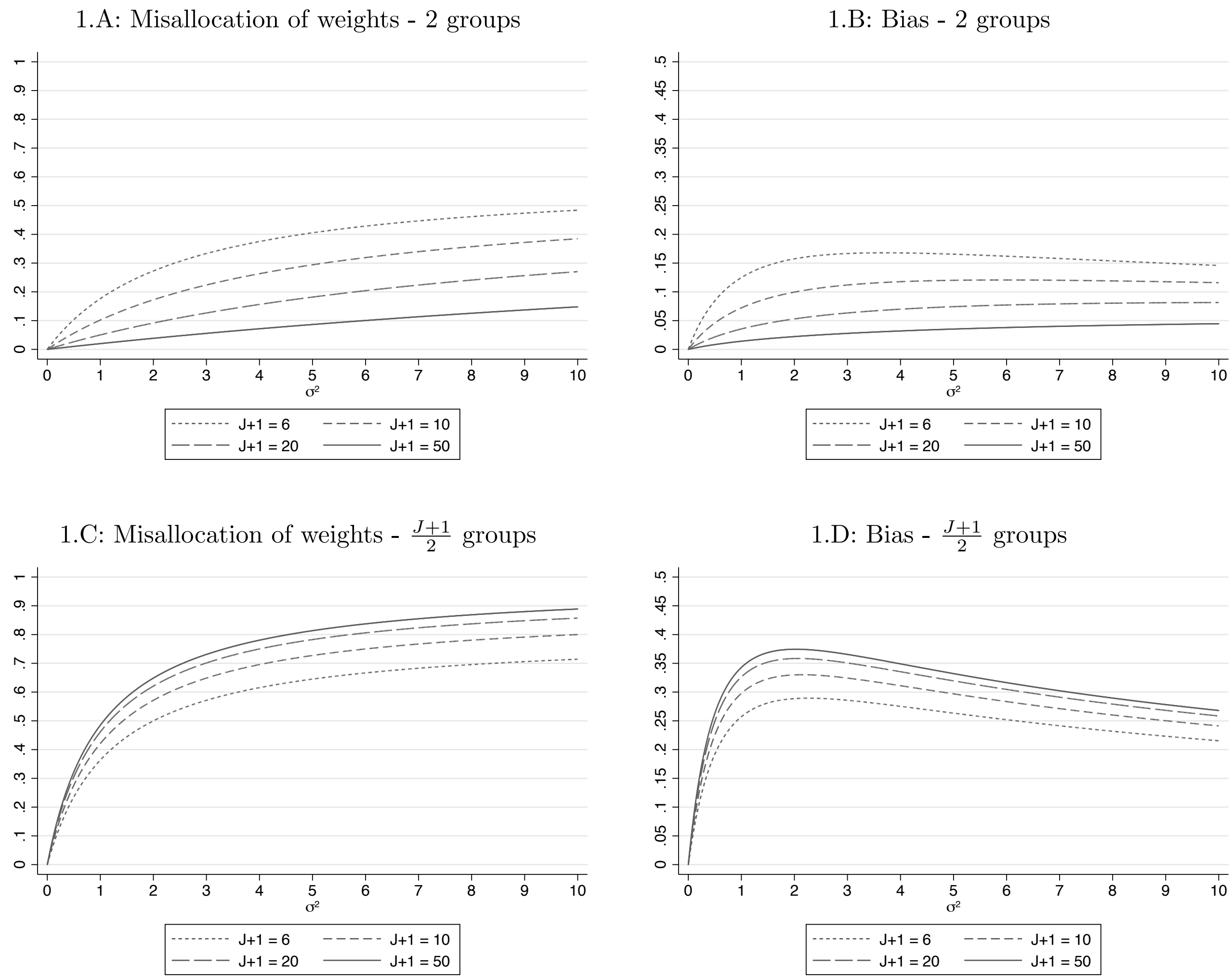
Newey, Whitney K. and Daniel McFadden, "Chapter 36 Large sample estimation and hypothesis testing," in "in," Vol. 4 of *Handbook of Econometrics*, Elsevier, 1994, pp. 2111 – 2245.

Powell, David, "Synthetic Control Estimation Beyond Case Studies: Does the Minimum Wage Reduce Employment?," 2016. RAND Corporation, WR-1142.

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

Xu, Yiqing, "Regional Policy Evaluation: Interative Fixed Effects and Synthetic Controls," *Political Analysis*, 2016. Forthcoming.

Figure 1: **Asymptotic Misallocation of Weights and Bias**



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 1.A and 1.B report results when there are 2 groups of $\frac{J+1}{2}$ units each, while figures 1.C and 1.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 ($R^2 > 08$)			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)
$T_0 = 5$	0.415	0.708	0.804	0.869	0.752	0.750	0.415	0.722	0.817
	[0.002]	[0.002]	[0.002]	[0.002]	[0.003]	[0.003]	[0.002]	[0.002]	[0.002]
$T_0 = 20$	0.196	0.493	0.653	0.728	0.036	0.006	0.178	0.369	0.558
	[0.001]	[0.001]	[0.001]	[0.003]	[0.001]	[0.001]	[0.001]	[0.006]	[0.020]
$T_0 = 50$	0.150	0.415	0.573	0.747	0.000	0.000	0.139	-	-
	[0.000]	[0.001]	[0.001]	[0.003]	[0.000]	[0.000]	[0.000]	-	-
$T_0 = 100$	0.130	0.384	0.540	0.793	0.000	0.000	0.124	-	-
	[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^2_ϵ). In all simulations, we set $J + 1 = 20$ and $K = 10$, which means that the 20 units are divided into 10 groups of 2 units that follow the same common factor λ^k_t. Columns 1 to 3 present the proportion of misallocated weights, which is given by the sum of weights allocated to units 3 to 20. Columns 4 to 6 present the probability that the pre-treatment match is close to perfect, defined as a R^2 greater than 0.8 in a regression of the treated unit pre-treatment outcomes on the SC unit pre-treatment outcomes. 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)
$T_0 = 5$	1.758	1.152	1.045
	[0.014]	[0.008]	[0.006]
$T_0 = 20$	2.412	1.321	1.121
	[0.013]	[0.007]	[0.005]
$T_0 = 50$	2.557	1.388	1.163
	[0.014]	[0.008]	[0.006]
$T_0 = 100$	2.569	1.403	1.187
	[0.015]	[0.008]	[0.005]

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)
$T_0 = 5$	0.367	0.655	0.758	0.103	0.191	0.235
	[0.002]	[0.002]	[0.002]	[0.001]	[0.002]	[0.002]
$T_0 = 20$	0.174	0.440	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	0.014	0.035	0.049
	[0.001]	[0.001]	[0.001]	[0.000]	[0.000]	[0.000]
$T_0 = 100$	0.120	0.346	0.488	0.009	0.021	0.030
	[0.000]	[0.001]	[0.001]	[0.000]	[0.000]	[0.000]

	Probability of perfect match ($R^2 > 08$)			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)
$T_0 = 5$	0.911	0.764	0.726	0.367	0.663	0.769
	[0.002]	[0.003]	[0.003]	[0.002]	[0.002]	[0.002]
$T_0 = 20$	0.988	0.618	0.366	0.173	0.428	0.576
	[0.001]	[0.003]	[0.003]	[0.001]	[0.002]	[0.002]
$T_0 = 50$	1.000	0.856	0.586	0.136	0.370	0.512
	[0.000]	[0.002]	[0.003]	[0.001]	[0.001]	[0.001]
$T_0 = 100$	1.000	0.975	0.836	0.120	0.346	0.488
	[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 ϕ^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 R^2 greater than 0.8 in a regression of the treated unit pre-treatment outcomes on the SC unit pre-treatment outcomes. 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)
$T_0 = 5$	3.516	2.003	1.655
	[0.029]	[0.014]	[0.010]
$T_0 = 20$	8.430	4.006	3.017
	[0.061]	[0.026]	[0.021]
$T_0 = 50$	13.526	6.371	4.739
	[0.116]	[0.033]	[0.037]
$T_0 = 100$	19.193	9.019	6.695
	[0.138]	[0.053]	[0.046]

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 5: **Permutation test with asymptotically biased estimator**

	Stationary model					
	$E[\lambda_t^1 \mid d_{1t}=1]=1$			$E[\lambda_t^1 \mid d_{1t}=1]=2$		
	$\sigma_\epsilon^2=0.1$	$\sigma_\epsilon^2=0.5$	$\sigma_\epsilon^2=1$	$\sigma_\epsilon^2=0.1$	$\sigma_\epsilon^2=0.5$	$\sigma_\epsilon^2=1$
	(1)	(2)	(3)	(4)	(5)	(6)
$T_0=5$	0.070	0.064	0.062	0.128	0.107	0.100
	[0.001]	[0.001]	[0.001]	[0.002]	[0.002]	[0.001]
$T_0=20$	0.128	0.094	0.082	0.326	0.225	0.185
	[0.002]	[0.001]	[0.001]	[0.002]	[0.002]	[0.002]
$T_0=50$	0.159	0.116	0.100	0.397	0.295	0.244
	[0.002]	[0.002]	[0.001]	[0.002]	[0.002]	[0.002]
$T_0=100$	0.172	0.125	0.109	0.416	0.323	0.271
	[0.002]	[0.002]	[0.002]	[0.002]	[0.002]	[0.002]

	Non-stationary model					
	$E[\lambda_t^1 \mid d_{1t}=1]=1$			$E[\lambda_t^1 \mid d_{1t}=1]=2$		
	$\sigma_\epsilon^2=0.1$	$\sigma_\epsilon^2=0.5$	$\sigma_\epsilon^2=1$	$\sigma_\epsilon^2=0.1$	$\sigma_\epsilon^2=0.5$	$\sigma_\epsilon^2=1$
	(7)	(8)	(9)	(10)	(11)	(12)
$T_0=5$	0.058	0.056	0.053	0.079	0.071	0.067
	[0.001]	[0.001]	[0.001]	[0.001]	[0.001]	[0.001]
$T_0=20$	0.072	0.065	0.061	0.124	0.105	0.096
	[0.001]	[0.001]	[0.001]	[0.002]	[0.002]	[0.001]
$T_0=50$	0.081	0.071	0.067	0.136	0.118	0.109
	[0.001]	[0.001]	[0.001]	[0.002]	[0.002]	[0.002]
$T_0=100$	0.088	0.077	0.072	0.139	0.124	0.115
	[0.001]	[0.001]	[0.001]	[0.002]	[0.002]	[0.002]

Notes: this table presents MC simulations results on a permutation test where the SC estimator is asymptotically biased. Simulations are the same as in Tables 1 and 3 with the exception that in columns 1 to 3 we add 1 to λ_t^1 when $t > T_0$, while in columns 4 to 6 we add 2. Columns 1 to 6 present results for a stationary model, while columns 7 to 12 present results for a model with both non-stationary and stationary common factors. Standard errors in brackets.

Table 6: **Permutation Test with Heteroskeadsticity**

	Stationary model				Non-stationary model			
	without serial correlation		with 0.9 serial correlation		without serial correlation		with 0.9 serial correlation	
	t_i^{ratio}	$\tilde{t}_i$	t_i^{ratio}	$\tilde{t}_i$	t_i^{ratio}	$\tilde{t}_i$	t_i^{ratio}	$\tilde{t}_i$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$T_0 = 5$	0.140	-	0.238	-	0.116	-	0.187	-
	[0.002]	-	[0.003]	-	[0.002]	-	[0.003]	-
$T_0 = 20$	0.089	0.085	0.178	0.163	0.072	0.067	0.133	0.114
	[0.002]	[0.002]	[0.003]	[0.003]	[0.002]	[0.002]	[0.002]	[0.002]
$T_0 = 50$	0.071	0.071	0.131	0.121	0.063	0.058	0.111	0.097
	[0.002]	[0.002]	[0.002]	[0.002]	[0.002]	[0.002]	[0.002]	[0.002]
$T_0 = 100$	0.062	0.060	0.106	0.091	0.056	0.053	0.094	0.079
	[0.002]	[0.002]	[0.002]	[0.002]	[0.002]	[0.002]	[0.002]	[0.002]
$T_0 = 1000$	0.051	0.052	0.072	0.054	0.049	0.050	0.073	0.054
	[0.002]	[0.002]	[0.002]	[0.002]	[0.002]	[0.002]	[0.002]	[0.002]

Notes: this table presents rejection rates when the variance of the transitory shocks for the treated unit is 0.1 while the variance of the transitory shocks for the control unit is 1. Columns 1 and 2 consider the stationary model when the common factor is serially uncorrelated using, respectively, the test statistic suggested in Abadie et al. (2010) and the one suggested in equation 28. Columns 3 and 4 present results when the serial correlation of the common factor is 0.9. Columns 5 to 8 present results for the non-stationary model. It is not possible to calculate $\tilde{t}_i$ with $T_0 = 5$. Standard errors in brackets.

Table 7: **Conditional Permutation Test**

	Stationary model			Non-stationary model		
	$\sigma_\epsilon^2 = 0.1$	$\sigma_\epsilon^2 = 0.5$	$\sigma_\epsilon^2 = 1$	$\sigma_\epsilon^2 = 0.1$	$\sigma_\epsilon^2 = 0.5$	$\sigma_\epsilon^2 = 1$
	(1)	(2)	(3)	(4)	(5)	(6)
			Panel i: conditional on $R^2 > 0.8$			
$T_0 = 5$	0.057	0.066	0.067	0.055	0.065	0.068
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
	(0.871)	(0.756)	(0.751)	(0.913)	(0.768)	(0.731)
$T_0 = 20$	0.056	0.154	0.276	0.050	0.061	0.074
	[0.000]	[0.003]	[0.009]	[0.000]	[0.000]	[0.001]
	(0.731)	(0.035)	(0.006)	(0.988)	(0.622)	(0.366)
$T_0 = 50$	0.055	0.337	-	0.050	0.052	0.056
	[0.000]	[0.052]	-	[0.000]	[0.000]	[0.000]
	(0.751)	(0.000)	(0.000)	(1.000)	(0.858)	(0.590)
$T_0 = 100$	0.053	-	-	0.050	0.050	0.051
	[0.000]	-	-	[0.000]	[0.000]	[0.000]
	(0.794)	(0.000)	(0.000)	(1.000)	(0.976)	(0.835)
			Panel ii: conditional on $R^2 > 0.9$			
$T_0 = 5$	0.071	0.089	0.090	0.063	0.084	0.090
	[0.000]	[0.001]	[0.001]	[0.000]	[0.001]	[0.001]
	(0.704)	(0.560)	(0.557)	(0.796)	(0.591)	(0.551)
$T_0 = 20$	0.088	0.388	0.600	0.053	0.079	0.105
	[0.001]	[0.031]	[0.098]	[0.000]	[0.001]	[0.002]
	(0.196)	(0.001)	(0.000)	(0.876)	(0.250)	(0.092)
$T_0 = 50$	0.099	-	-	0.050	0.057	0.063
	[0.002]	-	-	[0.000]	[0.001]	[0.001]
	(0.050)	(0.000)	(0.000)	(0.989)	(0.481)	(0.217)
$T_0 = 100$	0.107	-	-	0.050	0.052	0.054
	[0.006]	-	-	[0.000]	[0.000]	[0.001]
	(0.007)	(0.000)	(0.000)	(1.000)	(0.757)	(0.459)

Notes: this table presents rejection rates conditional on the a good pre-treatment fit for the treated unit. Columns 1 to 3 present results for the stationary model (as in Table 1), while columns 4 to 6 present results for the non-stationary model (as in Table 3). Panel i defines good pre-treatment fit as a $R^2 > 0.8$ for the regression of the pre-treatment outcomes of the treated unit on the pre-treatment outcomes of the SC unit. Panel ii defines good pre-treatment fit as $R^2 > 0.9$. In parenthesis, we present the probability of having a good match. 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 ($R^2 > 09$)			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)
$T_0 = 5$	0.415	0.708	0.804	0.703	0.559	0.552	0.424	0.738	0.828
	[0.002]	[0.002]	[0.002]	[0.003]	[0.004]	[0.004]	[0.003]	[0.003]	[0.002]
$T_0 = 20$	0.196	0.493	0.653	0.195	0.001	0.000	0.149	-	-
	[0.001]	[0.001]	[0.001]	[0.003]	[0.000]	[0.000]	[0.001]	-	-
$T_0 = 50$	0.150	0.415	0.573	0.050	0.000	0.000	0.106	-	-
	[0.000]	[0.001]	[0.001]	[0.002]	[0.000]	[0.000]	[0.001]	-	-
$T_0 = 100$	0.130	0.384	0.540	0.007	0.000	0.000	0.087	-	-
	[0.000]	[0.001]	[0.001]	[0.001]	[0.000]	[0.000]	[0.002]	-	-

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 ($R^2 > 08$)			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)
$T_0 = 5$	0.090	0.200	0.263	0.911	0.787	0.768	0.088	0.197	0.263
	[0.001]	[0.001]	[0.002]	[0.002]	[0.003]	[0.003]	[0.001]	[0.002]	[0.002]
$T_0 = 20$	0.064	0.136	0.192	0.948	0.268	0.077	0.062	0.106	0.137
	[0.000]	[0.001]	[0.001]	[0.002]	[0.003]	[0.002]	[0.000]	[0.001]	[0.003]
$T_0 = 50$	0.052	0.112	0.158	0.988	0.039	0.000	0.052	0.067	-
	[0.000]	[0.001]	[0.001]	[0.001]	[0.001]	[0.000]	[0.000]	[0.002]	-
$T_0 = 100$	0.044	0.095	0.134	0.999	0.003	0.000	0.044	0.043	-
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.005]	-

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 Minimum Distance Problem

Using the notation of Abadie et al. (2010), the SC weights will solve the following optimization problem:

$$\|X_1 - X_0 W\|_V$$

where $\sum_{j=2}^{J} w_1^j = 1$ and $w_1^j > 0$ for all $j = 2, ..., J$, and

$$X_1 - X_0 W = \begin{bmatrix} Z_1 - \sum_{j \neq 1} w_1^j Z_j \\ \sum_{s=1}^{T_0} k_s^1 Y_{1s} - \sum_{j \neq 1} w_1^j \sum_{s=1}^{T_0} k_s^1 Y_{1s} \\ \vdots \\ \sum_{s=1}^{T_0} k_s^K Y_{1s} - \sum_{j \neq 1} w_1^j \sum_{s=1}^{T_0} k_s^K Y_{1s} \end{bmatrix}$$

We prove the properties of the M-estimator for the weights for the special case in which we use all the pre-treatment periods as predictors. In this case, V becomes the identity matrix, and the optimization problem for this particular case is:

$$\arg\min_{w \in W} \frac{\sum_{t_0=1}^{T} \left[\left(y_{1t} - \sum_{j \neq 1} w_1^j y_{jt} \right)' \left(y_{1t} - \sum_{j \neq 1} w_1^j y_{jt} \right) \right]}{T_0}$$

subject to $\sum_{j=2}^{J} w_1^j = 1$ and $w_1^j > 0$ for all $j = 2, ..., J$.. Define the vector Jx1 $\widehat{\mathbf{w}} \equiv \{\hat{w}_1^j\}_{j \neq 1}$ as the solution of this minimization problem.[24] Using the population model for y_{it}, we can write this optimization problem as:

$$\arg\min_{w \in W} \frac{\sum_{t_0=1}^{T} \left[\left(\epsilon_{1t} - \sum_{j \neq 1} w_1^j \epsilon_{jt} \right) + \lambda_t \left(\mu_{1t} - \sum_{j \neq 1} w_1^j \mu_{jt} \right) \right]^2}{T_0}$$

In order to show the uniform convergence of the objective function, we need to impose assumptions about the stochastic processes of $\{\epsilon_{jt}\}_{t=1}^{T_0}$ and $\{\lambda_t\}_{t=1}^{T_0}$.

Assumption 1: $(\epsilon_{jt}, \lambda_t)'$ is weakly stationary and second moment ergodic.

Lemma 1 *Define* $g(y_{1t}, y_{jt}, w) \equiv \left[\left(\epsilon_{1t} - \sum_{j \neq 1} w_1^j \epsilon_{jt} \right) + \lambda_t \left(\mu_{1t} - \sum_{j \neq 1} w_1^j \mu_{jt} \right) \right]^2$. *Under assumption 1,*

$$\sup_{w \in W} \left\| \frac{1}{T_0} \sum_{t_0=1}^{T} g(y_{1t}, y_{jt}, w) - \mathbb{E}\left[g(y_{1t}, y_{jt}, w) \right] \right\| \to_p 0 \tag{31}$$

[24]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.

Proof. Note that $g\left(y_{1t}, y_{jt}, w\right)$ is continuous a each set of $\left\{\widehat{w}_1^j\right\}_{j=2}^J$. In addition,

$$\begin{aligned}\left\|g\left(y_{1t}, y_{jt}, w\right)\right\| &\leq \left\|y_{1t} - \sum_{j=2}^J w_1^j y_{jt}\right\|' \left\|y_{1t} - \sum_{j=2}^J w_1^j y_{jt}\right\| \\ &\leq C\end{aligned}$$

By lemma 2.4 of Newey and McFadden (1994), we have uniform convergence:

$$\sup_{w \in W}\left\|\frac{1}{T_0}\sum_{t_0=1}^{T} g\left(y_{1t}, y_{jt}, w\right) - \mathbb{E}\left[g\left(y_{1t}, y_{jt}, w\right)\right]\right\| \rightarrow_p 0$$

■

Now, we need to show that there is one only set of $\mathbf{w}_0 \equiv \left\{w_1^j\right\}_{j=2}^J$ that maximizes $\mathbb{E}\left[g\left(y_{1t}, y_{jt}, w\right)\right]$.

$$\arg\min_{w \in W} \mathbb{E}\left[\left[\left(\epsilon_{1t} - \sum_{j \neq 1} w_1^j \epsilon_{jt}\right) + \lambda_t\left(\mu_{1t} - \sum_{j \neq 1} w_1^j \mu_{jt}\right)\right]^2\right]$$

In order to simplify the problem, we impose assumptions about the second moments of $\{\epsilon_{jt}\}_{t=1}^{T_0}$ and $\{\lambda_t\}_{t=1}^{T_0}$.

Assumption 2: ϵ_{jt} is uncorrelated with λ_t for $t = 1, ..., T_0$. In addition, $Var\left[\epsilon_{jt}\right] = \sigma^2$ and $\mathbb{E}\left[\lambda_t'\lambda_t\right] = \Omega$.

Under assumption 2, the population objective function simplifies to:

$$\mathbb{E}\left[g\left(y_{1t}, y_{jt}, w\right)\right] = \sigma^2\left(1 + \sum_{j \neq 1}\left(w_1^j\right)^2\right) + \left(\mu_{1t} - \sum_{j \neq 1} w_1^j \mu_{jt}\right)' \Omega \left(\mu_{1t} - \sum_{j \neq 1} w_1^j \mu_{jt}\right)$$

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, $\{w_1^j\}_{j \neq 1}$, 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 population optimization problem as:

$$\arg\min_{w \in W} \mathbf{w}'\mathbf{w} + \left(\mu_1 - \mu_0\mathbf{w}\right)' \Omega \left(\mu_1 - \mu_0\mathbf{w}\right)$$

subject to $W = \{\mathbf{w} : \mathbf{w}'\iota = 1, \mathbf{w} \geq 0\}$, with ι being a vector $(J \times 1)$ of 1's. This is a minimization of a quadratic function in a convex space, and has a unique solution $\mathbf{w}_0$.

Using the results above, we could use the theory about M-estimator to show consistent of $\widehat{\mathbf{w}} \equiv \left\{\widehat{w}_1^j\right\}_{j=2}^J$.

Theorem 2 *Under assumptions 1 and 2,* $\widehat{\mathbf{w}} \rightarrow_p \mathbf{w}_0$

Proof. *Using the results of previous lemma and the fact that* $\mathbf{w}_0$ *is the unique maximum of* $Q_0(w) \equiv \mathbb{E}\left[g\left(y_{1t}, y_{jt}, \mathbf{w}_0\right)\right]$ *and* W *is compact, we can use Theorem* 2.1 *of Newey and McFadden (1994) to show that* $\widehat{\mathbf{w}} \rightarrow_p \mathbf{w}_0$. ■

A.2 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:

$$\begin{aligned} E[\hat{\alpha}_{1t}-\alpha_{1t}] &= \gamma + E\left[\sum_{j\neq 1}(\bar{w}_1^j-\hat{w}_1^j)\epsilon_{jt}\right] + E\left[\lambda_t\sum_{j\neq 1}(\bar{w}_1^j-\hat{w}_1^j)\mu_j\right] \\ &= \gamma + \sum_{j\neq 1}E\left[(\bar{w}_1^j-\hat{w}_1^j)\epsilon_{jt}\right] + \sum_{j\neq 1}E\left[\lambda_t(\bar{w}_1^j-\hat{w}_1^j)\right]\mu_j \end{aligned}$$

Given that $\hat{w}_1^j$ is a consistent estimator for $\bar{w}_1^j$, if we have that ϵ_{it} has finite variance, then:

$$\left|E\left[(\bar{w}_1^j-\hat{w}_1^j)\epsilon_{jt}\right]\right| \leq E\left[\left|(\bar{w}_1^j-\hat{w}_1^j)\epsilon_{jt}\right|\right] \leq \sqrt{E\left[(\bar{w}_1^j-\hat{w}_1^j)^2\right]E\left[(\epsilon_{jt})^2\right]} \to 0$$

Similarly, if λ_t^f has finite variance for all $f=1,...,F$, then $E\left[\lambda_t(\bar{w}_1^j-\hat{w}_1^j)\right]\mu_j \to 0$.

A.3 Demeaned Estimator

In this section, we formalize the alternative SC estimator that we propose in section ?? of the paper. In this new method, before finding the weights, we calculate the pre-treatment average of all units and demean the data. The "within-model" for treatment and control units are, respectively:

$$y_{it}^C - \overline{y}_i = \left(\lambda_t - \overline{\lambda}\right)'\mu_i + (\epsilon_{it}-\overline{\epsilon}_i)$$

$$y_{it}^T - \overline{y}_i = \alpha_{it} + \left(y_{it}^C - \overline{y}_i\right)$$

where $\overline{y}_i = \frac{1}{T_0}\sum_{t=1}^{T_0} y_{it}$, $\overline{\lambda} = \frac{1}{T_0}\sum_{t=1}^{T_0}\lambda_t$ and $\overline{\epsilon}_i = \frac{1}{T_0}\sum_{t=1}^{T_0}\epsilon_{it}$.

Note that we can write this model as,

$$\widetilde{y}_{it}^C = \widetilde{\lambda}_t'\widetilde{\mu}_i + \widetilde{\varepsilon}_{it}$$

$$\widetilde{y}_{it}^T = \alpha_{it} + \widetilde{y}_{it}^C$$

where $\widetilde{\lambda}_t$ does not include any time-invariant common factor, and $\widetilde{\mu}_i$ does not involve factor loadings associated with a constant common factor. This model is the same as before, but using the demeaned variables. In this case,

$$\widehat{\alpha}_{1t} \to \alpha_{1t} + \left(\widetilde{\epsilon}_{1t}\sum_{j\neq 1}\overline{w}_j^1\widetilde{\epsilon}_{jt}\right) + \widetilde{\lambda}_t'\left(\widetilde{\mu}_1\sum_{j\neq 1}\overline{w}_j^1\widetilde{\mu}_j\right)$$

Under the assumptions of the Difference-in-Difference Model,

$$\mathbb{E}\left[\widetilde{\lambda}_t\right] = 0$$

and

$$\mathbb{E}\left[\widetilde{\epsilon}_{1t} - \sum_{j\neq 1}\overline{w}_j^1\widetilde{\epsilon}_{jt}\right] = 0$$

In this case, the SC estimator is asymptotically unbiased.

A.4 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:

$$\mu_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \mu_2 = \begin{pmatrix} 0.5 \\ 1 \end{pmatrix}, \mu_3 = \begin{pmatrix} 1.5 \\ 1 \end{pmatrix}, \mu_4 = \begin{pmatrix} 0.5 \\ 0 \end{pmatrix}, \mu_5 = \begin{pmatrix} 1.5 \\ 1 \end{pmatrix} \tag{32}$$

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 1. 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 $\sigma_\epsilon^2 = E[(\lambda_t^1)^2] = E[(\lambda_t^2)^2] = 1$, then the factor loadings of the SC unit would be given by $(1.038, 0.8458)$. Therefore, there is small loss in the construction of the first factor loading and a gain in the construction of the second factor loading. Therefore, if selection into treatment is correlated with the common shock λ_t^1, then the SC estimator would be more asymptotically biased than the DID estimator.

A.5 Relation with Powell (2016) and Wong (2015)

In this section of the Appendix, we show how the proofs in Wong (2015) and Powell (2016) differ from our approach.

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

$$\widehat{\mathbf{w}} - \mathbf{w} = \left((Y'Y)^{-1} - (Y'Y)^{-1}\mathbf{j}(\mathbf{j}'(Y'Y)^{-1}\mathbf{j})^{-1}\mathbf{j}'(Y'Y)^{-1}\right) Y'(\zeta - Y'\mathbf{w}) \tag{33}$$

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.

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 SC 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 $\frac{1}{T_0}Y'Y \rightarrow E[\mathbf{y}_t\mathbf{y}_t']$ and $\frac{1}{T_0}Y'(\zeta - Y\mathbf{w}) \rightarrow E[\mathbf{y}_t(y_{1t} - \mathbf{y}_t'\mathbf{w})]$. Wong (2015) argues that $E[\mathbf{y}_t(y_{1t} - \mathbf{y}_t'\mathbf{w})] = 0$. However, we have that:

$$\begin{aligned} E[\mathbf{y}_t(y_{1t} - \mathbf{y}_t'\mathbf{w})] &= E[\mathbf{y}_t y_{1t}] - E[\mathbf{y}_t\mathbf{y}_t'\mathbf{w}] = E[(\mathbf{y}_t^* + \epsilon_t)(y_{1t}^* + \epsilon_{1t})] - E[(\mathbf{y}_t^* + \epsilon_t)(\mathbf{y}_t^* + \epsilon_t)'\mathbf{w}] \\ &= \mathbf{y}_t^* y_{1t}^* - \mathbf{y}_t^*\mathbf{y}_t^{*\prime}\mathbf{w} - E[\epsilon_t\epsilon_t']\mathbf{w} = -E[\epsilon_t\epsilon_t']\mathbf{w} \end{aligned} \tag{34}$$

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.

In another article, Powell (2016) proposes a generalization of the SC method where the treatment can be multivalued and more than one unit may be treated. He jointly estimates the treatment effect and the SC weights, and argues that the estimator for the treatment effect is consistent. In Theorem 3.1 of his paper, he argues that the following objective function has a unique minimum at $b = \alpha_0$ (although there might be multiple choices of weights):

$$\Gamma(b, \{w_i^j\}) = E\left[||Y_{it} - D_{it}'b - \sum_{j \neq i}\left(w_i^j(Y_{jt} - D_{jt}'b)\right)|| \right] \tag{35}$$

where D_{it} is a $(K \times 1)$ vector of treatment variables and α_0 is the $(K \times 1)$ vector of treatment effects.

We show that this generally will not be the case. For simplicity, we assume that μ_i is fixed and that $\mu_i - \sum_{j \neq i} {w_i^j}^* \mu_j = 0$ for some $\{{w_i^j}^*\}_{j \neq i}$. Therefore:

$$\begin{aligned} \Gamma(b, \{w_i^j\}) &= E\left[\left(\epsilon_i - \sum_{j \neq i} w_i^j \epsilon_j\right)^2\right] \\ &+ \left(\mu_i - \sum_{j \neq i} w_i^j \mu_j\right)' E[\lambda_t' \lambda_t] \left(\mu_i - \sum_{j \neq i} w_i^j \mu_j\right) \\ &+ (\alpha_0 - b)' \left(D_{it} - \sum_{j \neq i} w_i^j D_{jt}\right) \left(D_{it} - \sum_{j \neq i} w_i^j D_{jt}\right)' (\alpha_0 - b) \\ &+ \left(\mu_i - \sum_{j \neq i} w_i^j \mu_j\right)' cov\left(\lambda_t', \left(D_{it} - \sum_{j \neq i} w_i^j D_{jt}\right)'\right) (\alpha_0 - b) \end{aligned} \tag{36}$$

Note that we can set the second, third, and the forth terms of this objective function equal to zero by setting $w_i^j = {w_i^j}^*$ and $b = \alpha_0$. However, there is a first order gain in moving in the direction of weights that minimize the first term. Therefore, there is a set of parameters $\tilde{w}_i^j \neq {w_i^j}^*$ and $b = \alpha_0$ that attains a lower value than ${w_i^j}^*$ and $b = \alpha_0$ (unless ${w_i^j}^*$ minimizes the first term). Since $b = \alpha_0$ minimizes the objective function conditional on setting $w_i^j = {w_i^j}^*$, then it cannot be that the optimal weights will be given by ${w_i^j}^*$. Let $\tilde{\tilde{w}}_i^j$ be the weights that minimize the objective function. Therefore, $\mu_i - \sum_{j \neq i} \tilde{\tilde{w}}_i^j \mu_j \neq 0$.

Now we consider whether $\tilde{\tilde{w}}_i^j$ and $b = \alpha_0$ can be the solution to the problem. Note that the third term can be set to zero by choosing $b = \alpha_0$. However, if treatment assignment is correlated with λ_t, then we could make the forth term lower than zero. Since the first order effect of moving away from $b = \alpha_0$ on the third term is equal to zero, while we can have a first order gain in the forth term, then α_0 would not be the solution to this minimization problem. Note that $b = \alpha_0$ minimizes this problem if treatment assignment is uncorrelated with the common factors. Again, this is consistent with the results we find that the SC is asymptotically unbiased in this case.

A.6 Permutation test

We now prove that that the test statistic $\tilde{t}_i$ has, asymptotically, the same expected value and variance for all permutations. We have that:

$$\tilde{t}_i = \frac{\frac{1}{T-T_0} \sum\limits_{t=T_0+1}^{T} \left[y_{it} - \sum\limits_{j \neq i} \widehat{w}_i^j y_{jt}\right]^2 - \widehat{E[S_i]}}{\sqrt{\widehat{Var[S_i]}}}$$

where $S_i = \frac{1}{T-T_0} \sum\limits_{t=T_0}^{T} \left[\epsilon_{it} - \sum\limits_{j \neq i} \overline{w}_i^j \epsilon_{jt} + \lambda_t \left(\mu_i - \sum\limits_{j \neq i} \overline{w}_i^j \mu_j\right)\right]^2$. We use $T - T_0$ blocks of a combination of pre-treatment variables defined by $\widehat{P_{ik}} = \frac{1}{T-T_0} \sum\limits_{s=k}^{k+T-T_0-1} \left(y_{is} - \sum\limits_{j \neq i} \widehat{w}_i^j y_{js}\right)^2$ for $k = 1, ..., 2T_0 - T$. In this case, the expectation of S_i is estimated by:

$$\widehat{E[S_i]} = \frac{1}{2T_0 - T} \sum_{k=1}^{2T_0 - T} \left[\widehat{P_{ik}}\right]$$

and the estimator of the variance is:

$$\widehat{Var\left[S_i\right]} = \frac{1}{2T_0 - T} \sum_{k=1}^{2T_0-T} \left[\widehat{P_{ik}} - \widehat{E\left[S_i\right]}\right]^2$$

We need to impose the following assumptions. Consider the sequence $\{P_{ik}\}_{k=1}^{2T_0-T}$. We assume that:

1. P_{ik} is a covariance-stationary sequence.

2. P_{ik} is α-mixing with size $-\frac{r}{r-1}$, $r > 4$.

3. $E\left[|P_{ik}|^{r+\delta}\right] < \Delta < 0$ for some $\delta > 0$ at all s.

4. $\frac{1}{T_0}\sum_{s=1}^{T_0} P_{ik}^2 \to_p E\left[P_{ik}^2\right]$

Lemma 3 *Under assumptions 1-5, and assuming that $\widehat{w}_i^j \to_p \overline{w}_i^j$, then we have that the expected value of the asymptotic distribution of $\tilde{t}_i$ is equal to zero and the asymptotic variance is equal to 1.*

Proof. *Using the result that $\widehat{w}_i^j \to_p \overline{w}_i^j$,*

$$\frac{1}{2T_0 - T} \sum_{k=1}^{2T_0-T} \left[\widehat{P_{ik}}\right] = \frac{1}{2T_0 - T} \sum_{k=1}^{2T_0-T} \left[P_{ik}\right] + o_p\left(1\right)$$

Under assumptions 1-5, and using Corollary 3.48 in White(1999),

$$\frac{1}{2T_0 - T} \sum_{k=1}^{2T_0-T} \left[\widehat{P_{ik}}\right] \to_p E\left[P_{ik}\right] = E\left[\frac{1}{T - T_0}\sum_{t=T_0}^{T}\left(y_{is} - \sum_{j\neq i}\overline{w}_i^j y_{js}\right)^2\right]$$

Under assumption 2 in the main text,

$$E\left[\frac{1}{T - T_0}\sum_{t=T_0}^{T}\left(y_{is} - \sum_{j\neq i}\overline{w}_i^j y_{js}\right)^2\right] = E\left[\left(y_{is} - \sum_{j\neq i}\overline{w}_i^j y_{js}\right)^2\right]$$

Using the model for y_{is} and under the condition that $\sum_{j=2}^{J+1}\overline{w}_i^j = 1$,

$$E\left[\left(y_{is} - \sum_{j\neq i}\overline{w}_i^j y_{js}\right)^2\right] = E\left[\left(\epsilon_{is} - \sum_{j\neq i}\overline{w}_i^j \epsilon_{js} + \lambda_s\left(\mu_{is} - \sum_{j\neq i}\overline{w}_i^j \mu_j\right)\right)^2\right]$$

At the end,

$$\widehat{E\left[S\right]} \to_p E\left[S\right]$$

Using a proof analogous to the lemma above, we can show that $\widehat{Var\left[S\right]} \to_p Var\left[S\right]$.

Therefore:

$$\widetilde{t}_i \to_d \frac{\frac{1}{T-T_0}\sum_{t=T_0+1}^{T}\left(\varepsilon_{it} - \sum_{j\neq i}\overline{w}_i^j\varepsilon_{jt} + \lambda_t\left(\mu_i - \sum_{j\neq i}\overline{w}_i^j\mu_j\right)\right)^2 - E\left[\frac{1}{T-T_0}\sum_{t=T_0+1}^{T}\left(\varepsilon_{it} - \sum_{j\neq i}\overline{w}_i^j\varepsilon_{jt} + \lambda_t\left(\mu_i - \sum_{j\neq i}\overline{w}_i^j\mu_j\right)\right)^2\right]}{\sqrt{Var\left[\frac{1}{T-T_0}\sum_{t=T_0+1}^{T}\left(\varepsilon_{it} - \sum_{j\neq i}\overline{w}_i^j\varepsilon_{jt} + \lambda_t\left(\mu_i - \sum_{j\neq i}\overline{w}_i^j\mu_j\right)\right)^2\right]}}$$

■

Notation

Variable	*Dimension*	*Description*
y_{it}	(1×1)	*Outcome for unit i at time t*
y_{it}^{C}	(1×1)	*Potential outcome for unit i at time t if not treated*
y_{it}^{T}	(1×1)	*Potential outcome for unit i at time t if treated*
$\boldsymbol{Y}_1^P$	$(T_0 \times 1)$	*Vector of pre-treatment outcome for the treated*
$\boldsymbol{Y}_0^P$	$(T_0 \times J)$	*Matrix of pre-treatment outcome for the controls*
$\boldsymbol{y}_t$	$(J \times 1)$	*Vector of outcomes for the controls at time t*
Z_i	$(R \times 1)$	*Vector of covariates*
X_1	$(K \times 1)$	*Vector of economic predictors for the treated*
X_0	$(K \times J)$	*Matrix of economic predictors for the controls*
λ_t	$(1 \times F)$	*Vector of common factors*
Ω	$(F \times F)$	$E[\lambda_t' \lambda_t]$
μ_i	$(F \times 1)$	*Vector of factor loadings*
μ_0	$(F \times J)$	*Matrix of factor loadings for the controls*
α_{it}	(1×1)	*Treatment effect for unit i at time t*
$\boldsymbol{w}$ *or* $\{w_1^j\}_{j \neq 1}$	$(J \times 1)$	*Vector of weights*
$\widehat{\boldsymbol{w}}$ *or* $\{\hat{w}_1^j\}_{j \neq 1}$	$(J \times 1)$	*M-estimator of weights*
$\bar{\boldsymbol{w}}$ *or* $\{\bar{w}_1^j\}_{j \neq 1}$	$(J \times 1)$	*Probability limit of the M-estimator of weights*
ϵ_{it}	(1×1)	*Idiosyncratic error for unit i at time t*
ϵ_t	$(J \times 1)$	*Idiosyncratic error for the control units at time t*