Hypothesis Testing with the Synthetic Control Method

Michihito Ando* Fredrik Sävje†

June 5, 2013

PRELIMINARY, DO NOT DISTRIBUTE

Abstract

The usage of Synthetic Controls (SC) to estimate the causal effect of events or interventions is a case study method gaining traction in the social sciences. Hypothesis testing in the SC method relies on non-standard permutations tests that thus far not been investigated extensively. In this paper we highlight a condition, not previously discussed in the literature, that is required for the validity of the current test procedure and show that it is unlikely to be fulfilled in a typical SC setting. The consequences of a violation however depend on the specific context. It is therefore a priori unclear how severely this affects the performance of the test. To investigate this, we conduct a Monte Carlo simulation study. This study shows that a violation could render the test both uninformative and misleading, but also that it remains informative in some settings. Additionally we investigate two alternative tests previously discussed in the literature, and introduce a new test. These tests are expected to be valid in a wider range of settings, which is confirmed by the simulation study. In particular the test we introduce fares considerably better than all other test when the parametric assumptions often made in a SC setting are true.

*Department of Economics, Uppsala University; UCFS; and Institute for Housing and Urban Research, Uppsala University.

†Department of Economics, Uppsala University; and UCLS. Correspondence: fredrik.savje@nek.uu.se

1 Introduction

The Synthetic Control (SC) method is a novel approach to case studies, where the causal effect of an intervention or event affecting a single unit is of interest. It was first used by Abadie and Gardeazabal (2003) to estimate the effect of terrorism on economic activity in the Basque Country, Spain. Important methodical discussions were later given by Abadie, Diamond and Hainmueller (2010), hereafter referred to as ADH. It has since then been used to estimate the effect of the reunification of West and East Germany (Abadie et al. 2012); the impact of natural disasters (Cavallo et al. 2013; Coffman and Noy 2012); the effect of economic liberalization (Billmeier and Nannicini 2013); the effect of terrorism on voting behavior (Montalvo 2011); and the effect of banning affirmative action on college enrollment and educational outcomes (Hinrichs 2012), to name a few.

Common to all of these applications is that they present an exceptionally challenging setting to investigate. There are only a small number of observations and only one, or a few, of the units are affected by the event. Furthermore, we have reason to believe that the affected units differ from the unaffected in substantial and unobserved ways. While many quantitative methods would be unusable in this situation, the SC method is tailor-made for it. By constructing an artificial, or synthetic, control unit from a subset of the unaffected units we could allow for confounders, at least to some extent. Further, since this control unit is constructed for each affected unit separately we could allow the number of affected units to be very small—including only one.

The focus of this paper is the hypothesis test proposed by ADH, which been used extensively in the recent literature. One of the caveats of the SC method is that the distribution of the resulting estimator in general cannot be derived, not even asymptotically. Recognizing this, ADH propose a test similar to a permutation test where the analysis is extended by estimating the effect also on unaffected units. The estimates for these units are, by assumption, not influenced by the event. If the true effect on the investigated unit is zero, then no unit is affected by the event and there would arguably be no systematical difference between the estimators. That is, under the null hypothesis of no effect we have reason to believe that the derived estimates for all units are drawn from the same underlying distribution. If the realized estimates indicate that this is not the case, we are motivated to reject the null and deem it likely that the event had an effect on the investigated unit.

Our first contribution is to highlight an important implicit condition required by this test, namely that the estimators must follow the same distribution and be uncorrelated under the null of no effect. The importance of this condition has to our knowledge not fully been appreciated previously in the literature, and subsequently remained largely

undiscussed in the applications that use the test. If the estimators are not independent and identically distributed (IID) under the null, then finding that they differ would not provide support for rejection—they would differ also under the null. Specifically, we will discuss three possible violations to the IID condition. They are all based on the notion that the distribution of the estimator depend on how well the SC unit reproduce the outcome for investigated unit. First, our ability to construct a suitable SC unit will depend on the size of the control group. The estimators in the permutation test are constructed with different control group sizes, thereby implying different distributions. Second, the existence of confounders will constrain our ability to construct a suitable SC unit for the investigated unit but not for the comparison units, which also could invalidate the test. Third, the usability of the SC method in any particular application is judged by how well the SC unit can reproduce the counterfactual outcome for the investigated unit (which to some degree is observable). We would therefore only use the SC method when a suitable SC unit can be constructed. This selectiveness is however not applied to the comparison units and thereby implies different distributions in *realized* applications.

While these issues are expected to be present in most applications to some extent, their magnitude and relative importance greatly depend on the specific setting. If the distributions differ substantially the test would be uninformative, and in worst case misleading, but only slight differences would arguably not render it unusable. To investigate the sensitivity of the test we conduct a Monte Carlo simulation study with a range of different specifications. The specifications are chosen so that the influence of each of the three issues can be isolated, and thereby assessing their relative importance. In addition, we investigate the test with a specification that uses real data on average wage levels in U.S. counties where we assign fictitious events to random counties, thereby providing indications of the performance in a more realistic setting.

A second contribution is the introduction of an alternative test. This test builds on the recognition that we can measure the fitness of the SC unit for each unit and thereby, to some degree, correct the estimators for eventual differences in distributions. Since this would render the estimates more alike, under the null, the discussed issues would be mitigated. Although the connection between the fitness of the SC unit and the behavior of the estimator has been recognized previously in the literature—for example ADH discuss two alternative tests also exploiting this fact—we derive a more detailed relationship based on the parametric assumptions most commonly made in the SC literature. As a consequence the test fares better compared to the previous tests when these assumptions are true, as indicated by the simulation study. It would however not necessarily outperform the existing tests when parametric assumptions do not hold. The

proposed test is therefore best seen as a complement to the previous tests.

The remainder of this paper is structured as follows. The next section introduces the SC method and discusses some previously overlooked issues which are important to the following discussion. Section 3 presents the current procedure of hypothesis testing and investigates the conditions needed for its validity. Section 4 introduces the alternative tests, including the test proposed by us, and Section 5 presents the results from the simulation study. Section 6 concludes.

2 The SC method

An instructive interpretation of the SC method is as a mixture of two other methods commonly used in case studies. From a quantitative perspective a prevalent approach is the *difference-in-difference* method,[1] where it is assumed that the average counterfactual change in the outcome for the units of interest is equal to the average change of the outcome of the control units—the assumption of "parallel trends." While this method often makes a convincing case, in some applications—including many case studies—the parallel trends assumption might be overly restrictive. For instance, in the setting of Abadie and Gardeazabal (2003) it seems unlikely that the average change of all other regions of Spain would be a good representation of the counterfactual change in the Basque Country.

From a qualitative perspective a common approach is the *comparative case study*, where one would compare the unit of interest with unaffected units much in the same way as with the difference-in-difference method. This method is however not constrained by the assumption that all included units share the same expected counterfactual change, instead it flexibly compare different aspects between different sets of units. The unfortunate consequence of this flexibility is that it, arguably, provides too much leeway to the researcher.

The SC method would here be seen as their synthesis. Like the qualitative comparative case study method, it is not required to assume that all control units (nor their average) is representative of the unit of interest, but instead that different units contain different aspects which, taken together, are representative of the unit. Furthermore, like the quantitative methods the SC method constrains the researcher by standardizing the construction of the comparison unit and thereby relies on the subjective judgment of the researcher to a lesser extent.[2]

In this section we will present the SC method more in detail. While the described

[1] See Card (1990) and Card and Krueger (1994) for notable examples.
[2] Refer to Abadie et al. (2012) for a more in-depth discussion of this synthesis.

method in itself does not differ from ADH, the structure of the description differs in some important aspects. This is motivated by that it highlights some features fundamental to the pursuing analysis.

In any particular SC application we are interested in the effect that some event, program, reform or *intervention* has on a single unit. To this end we have data on $J+1$ units, denoted by i, in T time periods, denoted by t. In particular, we have access to the outcome of interest, Y_{it}, for all units in all periods and a number of time-invariant covariates, collected in a $[r \times 1]$ column vector $\mathbf{Z}_i$.[3] We let $i = 1$ denote the unit of interest—the unit potentially affected by the intervention—and let $i \in \{2, \cdots, J+1\}$ denote the unaffected *control units*. The start of the intervention is in period $T_0 + 1$ and we allow the effect of the intervention to linger to all subsequent periods. All periods $t \leq T_0$ we thus refer to as the *pre-intervention periods*, and $t > T_0$ is referred to as the *intervention periods*.

We assume there to be two states of the world, one in which the intervention took place and one in which it did not. The first state is the actual state and thus the outcome in this state is observed (Y_{it}). The unobserved, and hypothetical, outcome in the state without the intervention is denoted by Y_{it}^N.[4] Building on this framework, the SC method rests on three assumptions:

i. Only one unit has the potential of being affected by the intervention, namely the unit of interest (hereafter the *treated unit*). This assumption thus mandates that $Y_{it} = Y_{it}^N$ for $i \neq 1$ in all t, essentially assuming the absence of spill-over effects.

ii. There are no effect of the intervention prior to $T_0 + 1$, or in other words that $Y_{it} = Y_{it}^N$ for all $t \leq T_0$. If we expect there to be effects preceding the intervention (e.g. anticipation effects) then $T_0 + 1$ is adjusted so that $Y_{it} = Y_{it}^N$ for $t \leq T_0$ becomes reasonable.

iii. The counterfactual outcome of the treated unit can be represented, or approximated, by some stable convex combination of the counterfactual outcomes of the control units.

The first two assumptions are familiar from, for example, the difference-in-difference method. The third assumption however differs from traditional methods in that the SC

[3] If the covariates vary over time the standard approach has been to average each covariate over all pre-intervention periods, but any linear combination could in principle be used.

[4] Previously in the SC literature a more traditional potential outcome framework has been used, where treatment is assigned individually. This would however imply that each unit has the potential of being treated, which in a case study may be nonsensical. For example the question "what would have happened if West and East Germany had reunited in France?" seems to be largely invalid.

method only requires that *some* combination of the controls is representative, rather than one particular pre-defined combination (as with the parallel trends assumption).

Building on these assumptions, the SC method can be characterized as follows. Since the (unobserved) counterfactual outcome of the treated unit is given by a combination of controls which are unaffected by the intervention we can reconstruct that outcome—if the combination is known. The reconstructed counterfactual outcome could then be compared to the actual outcome to assess the effect of the intervention. While the specific combination is not known a priori, since no unit is affected by the intervention in the pre-intervention periods (and thus the observed outcomes are equal to the counterfactual) the specific combination could be derived using the observed outcomes from these periods.

In line with the previous literature we will make some parametric assumptions to clarify the discussion and the analysis. Consider a situation where the outcome in the counterfactual state, Y_{it}^N, is given by a common factor model:

$$Y_{it}^N = \delta_t + \boldsymbol{\theta}_t \mathbf{Z}_i + \boldsymbol{\lambda}_t \boldsymbol{\mu}_i + \varepsilon_{it} \tag{1}$$

where δ_t is an unobserved period-specific intercept, $\mathbf{Z}_i$ is the vector of observed covariates mentioned above, $\boldsymbol{\theta}_t$ is an $[1 \times r]$ row vector of corresponding unknown and time-variant coefficients, $\boldsymbol{\lambda}_t$ is an $[1 \times F]$ row vector of unobserved common factors, $\boldsymbol{\mu}_i$ is a $[F \times 1]$ column vector of corresponding unobserved factor loadings (potentially correlated with "treatment") and ε_{it} is a period and unit specific transitory shock, assumed to have zero mean and be independent and identically distributed between units.

The outcomes in the state of the world where the intervention took place are directly observed and thus we have no reason to make any parametric assumption on that process. The *treatment effect* is defined as the difference between the outcomes in the two states of the world: $\beta_{it} \equiv Y_{it} - Y_{it}^N$. Notice that, by assumption, the treatment effect is zero ($\beta_{it} = 0$) for the controls ($i \neq 1$) and in the pre-intervention periods ($t \leq T_0$). The aim of the SC method is to estimate the β_{it} which are not zero by assumption. That is the treatment effect of the treated unit in the intervention periods. As Y_{it} is observed the SC method reduces to the imputation of the unobserved Y_{it}^N.

With this aim, we introduce a unit-specific $[(J+1) \times 1]$ column vector describing a convex combination of the units, denoted by $\mathbf{W}_i = (w_{i1}, \cdots, w_{i(J+1)})'$. This convex combination is what we refer to as the *synthetic control unit*. The elements in this vector describe the weight, w_{ij}, that unit j has in the SC unit for i. Since the unit itself is not allowed to be in its own SC unit we restrict so that $w_{ii} = 0$. Further, in order to confine the SC unit to a *convex* combination we impose that all weights, for a given $\mathbf{W}_i$, are

non-negative and that their sum is one. For the moment we restrict the attention to the SC unit of the treated unit, later on we will however extend the analysis to all units.

Now, assume that there exists a specific SC unit, denoted by $\mathbf{W}_1^*$, that fulfills:

$$\mathbf{Z}_1 = \sum_{j=1}^{J+1} w_{1j}^* \mathbf{Z}_j \quad \text{and} \quad \boldsymbol{\mu}_1 = \sum_{j=1}^{J+1} w_{1j}^* \boldsymbol{\mu}_j. \tag{2}$$

In that case, the third assumption required by the SC method—that the counterfactual outcome can be reconstructed—is fulfilled. To see this, consider the difference between the counterfactual outcome of the treated unit and the convex combination of the controls:

$$\begin{aligned} Y_{1t}^N - \sum_{j=1}^{J+1} w_{1j}^* Y_{jt}^N &= (\delta_t + \boldsymbol{\theta}_t \mathbf{Z}_1 + \boldsymbol{\lambda}_t \boldsymbol{\mu}_1 + \varepsilon_{1t}) - \sum_{j=1}^{J+1} w_{1j}^* \left(\delta_t + \boldsymbol{\theta}_t \mathbf{Z}_j + \boldsymbol{\lambda}_t \boldsymbol{\mu}_j + \varepsilon_{jt}\right), \\ &= \varepsilon_{1t} - \sum_{j=1}^{J+1} w_{1j}^* \varepsilon_{jt}. \end{aligned} \tag{3}$$

Net of the transitory shocks, the outcome is reconstructed exactly. Together with the assumption of no spill-over effects this will enables us to estimate the effect of the intervention. The estimator (which we denote by $\hat{\beta}_{1t}^*$) is given by the difference between the observed outcome of the treated unit and the combination of the controls described by $\mathbf{W}_1^*$. Notice that the observed outcomes can be expressed as $Y_{it} = \beta_{it} + Y_{it}^N$, by a reshuffling of the definition of the treatment effect. The estimator thereby collapses to:

$$\begin{aligned} \hat{\beta}_{1t}^* = Y_{1t} - \sum_{j=1}^{J+1} w_{1j}^* Y_{jt} &= (\beta_{1t} + Y_{1t}^N) - \sum_{j=1}^{J+1} w_{1j}^* (\beta_{jt} + Y_{jt}^N), \\ &= \beta_{1t} + Y_{1t}^N - \sum_{j=1}^{J+1} w_{1j}^* Y_{jt}^N, \\ &= \beta_{1t} + \varepsilon_{1t} - \sum_{j=1}^{J+1} w_{1j}^* \varepsilon_{jt}, \end{aligned} \tag{4}$$

where the equality in the second row follows from that $\beta_{jt} = 0$ for all $j \neq 1$ (due to the absence of spill-overs), and the last equality follows from (3). The resulting expression contains the treatment effect and a linear combination of transitory shocks. Since the shocks have zero mean and are uncorrelated with $\mathbf{W}_1^*$ we have that $E(\hat{\beta}_{1t}^*) = \beta_{1t}$. In other words, (4) is an unbiased estimator of β_{1t}. In particular, notice that this is true for *any* joint distribution of the parameters—given that we have $\mathbf{W}_1^*$ we do not need to

constrain $\boldsymbol{\lambda}_t$ and $\boldsymbol{\mu}_i$ in any way.

This situation is however largely hypothetical—in any particular application we would not be given $\mathbf{W}_1^*$ and could therefore not derive $\hat{\beta}_{1t}^*$. While there often are good prospects of deriving the SC unit, as will become apparent, this fact still confronts us with two problems. First, there is no guarantee that there exists a set of weights so that (2) is fulfilled, i.e. $\mathbf{W}_1^*$ may not exist. If we stack both $\mathbf{Z}_i$ and $\boldsymbol{\mu}_i$ into a $[(r+F)\times 1]$ column vector, $\boldsymbol{\psi}_i = (\mathbf{Z}_i', \boldsymbol{\mu}_i')'$, then $\mathbf{W}_1^*$ exists only when $\boldsymbol{\psi}_1$ is in the convex hull formed by $\{\boldsymbol{\psi}_i : i \neq 1\}$.[5] This is not necessarily true in any particular application. Second, since we never directly observe $\boldsymbol{\mu}_i$ we cannot, with certainty, confirm that (2) is fulfilled.

With these two issues in mind, consider the consequences of using some set of weights, $\mathbf{W}_1$, which not necessarily is equal to $\mathbf{W}_1^*$. In that case (2) would not necessarily hold and the SC unit may not reconstruct the counterfactual outcome. Of the terms in equation (3) only δ_t will be reconstructed with any convex combination (since it is constant between units). For a given set of weights, let the differences in the two remaining terms—the covariates $\mathbf{Z}_i$ and factor loadings $\boldsymbol{\mu}_i$—between the unit and its SC be denoted by $\mathbf{B}_i(\mathbf{W}_i) = \left(\boldsymbol{\psi}_i - \sum_{i=1}^{J+1} w_{ij}\boldsymbol{\psi}_j\right)$, where $\boldsymbol{\psi}_i$ is defined as in the previous paragraph. Using some weights $\mathbf{W}_1 \neq \mathbf{W}_1^*$ would thus add $\boldsymbol{\phi}_t\mathbf{B}_1(\mathbf{W}_1)$ to the expression in (3), where $\boldsymbol{\phi}_t = (\boldsymbol{\theta}_t, \boldsymbol{\lambda}_t)$ collects the coefficients of the covariates and the common factors in an $[1\times(r+F)]$ row vector. The estimator resulting from these weights would therefore be:

$$\begin{aligned}\hat{\beta}_{1t}(\mathbf{W}_1) = Y_{1t} - \sum_{j=1}^{J+1} w_{1j}Y_{jt} &= \beta_{1t} + Y_{1t}^N - \sum_{j=1}^{J+1} w_{1j}Y_{jt}^N, \\ &= \beta_{1t} + \varepsilon_{1t} - \sum_{j=1}^{J+1} w_{1j}\varepsilon_{jt} + \boldsymbol{\phi}_t\mathbf{B}_1(\mathbf{W}_1). \qquad (5)\end{aligned}$$

Although $\mathbf{W}_1^*$ would pin the last term of (5) to zero and thereby result in that $\hat{\beta}_{1t}(\mathbf{W}_1^*)$ collapses into (4), any other set of weights would not. In these situations the estimator would not necessarily be unbiased and could to a high degree depend on the underlying distributions. While this fact causes most weights to be unusable for estimating the effect of the intervention, it also enables us to derive a feasible set of weights.

To this end, notice that (5) could be derived for any period, including the pre-intervention periods. For clarity we hereafter refer to $P_{it}(\mathbf{W}_i) \equiv \hat{\beta}_{it}(\mathbf{W}_i)$ in $t \leq T_0$ as the *pre-intervention fits*, while $\hat{\beta}_{it}(\mathbf{W}_i)$ is used exclusively in the intervention periods. Since there are no effect of the intervention prior to $T_0 + 1$ the first term (β_{1t}) of the

[5] This geometric interpretation will turn out to be a particularly fruitful in the analysis of the condition required by the hypothesis test, as presented in the following section.

pre-intervention fits is, by assumption, zero. That is, $P_{1t}(\mathbf{W}_1)$ is only affected by the transitory shocks and $\boldsymbol{\phi}_t\mathbf{B}_1(\mathbf{W}_1)$. Now, if we were to use $\mathbf{W}_1 = \mathbf{W}_1^*$ then this would have observable effects on the pre-intervention fits. Specifically, we would have $\mathbf{B}_1(\mathbf{W}_1^*) = \mathbf{0}$ and only the shocks remain in the expression. Since the shocks have zero mean we would, for a sufficiently large number of pre-intervention periods, expect the average of all $P_{1t}(\mathbf{W}_1)$ to be zero. Thus finding a $\mathbf{W}_1$ that results in an average pre-intervention fit close to zero would lead us to believe that this combination is, at least, close to $\mathbf{W}_1^*$.

That the average $P_{1t}(\mathbf{W}_1)$ is zero is however not a sufficient condition for that the used weights are equal, or close, to $\mathbf{W}_1^*$ (even disregarding the distorting effect of the transitory shocks). There is another circumstance, not previously discussed in the literature, under which the average pre-intervention fit is zero. If the distribution of $\boldsymbol{\phi}_t$ is stable over the pre-intervention periods, then using weights so that $\boldsymbol{\phi}_t$ on average is orthogonal to $\mathbf{B}_1(\mathbf{W}_1)$ would also result in that the average $P_{1t}(\mathbf{W}_1)$ is close to zero. The estimator implied by these weights would however only be unbiased if $\boldsymbol{\phi}_t$ in the intervention periods remain distributed as in the pre-intervention periods. While this additional assumption could be reasonable in some situations, we would prefer finding weights that satisfies (2) if possible—with $\mathbf{W}_1^*$ we can allow any distribution of $\boldsymbol{\phi}_t$.

We could however, in theory, differentiate between the two situations. The variance of $P_{1t}(\mathbf{W}_1)$ would decrease when $\mathbf{B}_i(\mathbf{W}_1)$ approaches zero, but this would not necessarily be the case when $\boldsymbol{\phi}_t$ becomes more orthogonal to $\mathbf{B}_1(\mathbf{W}_1)$.[6] If we, besides the mean, also focus on the variance of the pre-intervention fits then we could avoid finding a SC unit that is based on orthogonalization. Finding some $\mathbf{W}_1$ that results in pre-intervention fits that is centered around zero with only small deviations would make us confident that the resulting SC unit fulfill (2), at least approximately so.

We will not further discuss the exact procedures for deriving weights. Our analysis is applicable to any procedure that aims to find a SC unit that fulfills (2). In short, the approach proposed by ADH, which is almost exclusively used in the literature, tries to find weights as to minimize the mean squared pre-intervention fit (thereby considering both the mean and the variance), but it also exploits the fact that one part of (2) is observed—the covariates. In practice the weights are chosen as to minimize a weighted average of the squared pre-intervention fits and squared difference in the covariates. We direct the reader to ADH for further details.

[6] To see this, consider $\boldsymbol{\phi}_t\mathbf{B}_1(\mathbf{W}_1) = \phi_t^B\|\mathbf{B}_1(\mathbf{W}_1)\|$ where ϕ_t^B is the scalar projection of $\boldsymbol{\phi}_t$ onto $\mathbf{B}_1(\mathbf{W}_1)$ and $\|\mathbf{B}_1(\mathbf{W}_1)\|$ is the Euclidean norm of $\mathbf{B}_1(\mathbf{W}_1)$. Then the expected value and variance of this term, conditional on the weights, would be given by $E(\boldsymbol{\phi}_t\mathbf{B}_1(\mathbf{W}_1)|\mathbf{W}_1) = E(\phi_t^B|\mathbf{W}_1)\|\mathbf{B}_1(\mathbf{W}_1)\|$ and $V(\boldsymbol{\phi}_t\mathbf{B}_1(\mathbf{W}_1)|\mathbf{W}_1) = V(\phi_t^B|\mathbf{W}_1)\|\mathbf{B}_1(\mathbf{W}_1)\|^2$, since $\mathbf{B}_1(\mathbf{W}_1)$ is constant over time for a given $\mathbf{W}_1$. Thus if we find some $\mathbf{W}_1$ so that $\boldsymbol{\phi}_t$ on average is orthogonal to $\mathbf{B}_1(\mathbf{W}_1)$ (i.e. $E(\phi_t^B|\mathbf{W}_1) = 0$) then the mean would be zero even if (2) do not hold. However it is only when $\|\mathbf{B}_1(\mathbf{W}_1)\| = 0$ that $V(\boldsymbol{\phi}_t\mathbf{B}_1(\mathbf{W}_1)|\mathbf{W}_1) = 0$.

Once appropriate weights are derived, the estimates of the treatment effect is given by $\hat{\beta}_{1t}(\mathbf{W}_1)$. If $\mathbf{W}_1$ in fact does fulfill (2), or approximately does so, then $\boldsymbol{\phi}_t\mathbf{B}_1(\mathbf{W}_1)$ would be zero, or close to zero, and the estimator would be unbiased. This is however not automatic, if even the best SC unit does not reproduce counterfactual outcome of the treated unit well then the estimator could still be biased. It is therefore important to examine the pre-intervention fit of the investigated unit. When $P_{1t}(\mathbf{W}_1)$ is not centered around zero, or varies considerably, then we would be motivated to re-think whether to study the particular intervention with the SC method. Or in the words of ADH: "[F]or each particular application, the analyst can decide if the characteristics of the treated unit are sufficiently matched by the synthetic control. In some instances, the [pre-intervention] fit may be poor and then we would not recommend using a synthetic control."

3 Hypothesis testing

While the point estimate might be of greatest interest, in most settings we would also like to judge the statistical significance of the estimates, i.e. test against the hypothesis of no treatment effect. The distribution of the estimator, under the null of no effect, is however largely unknown and not easily derived. Since we cannot confirm that the weights do fulfill (2), even if the pre-intervention fits are close to zero, we can never discard the eventuality that the estimator is directly and to a large extent affected by the covariates and the common factors in the intervention periods. Furthermore, even if we were to find weights so that (2) hold exactly, the transitory shocks would still directly enter the estimator. Notice that this is true for any J, T_0 and T. While higher J and T_0 tend to mitigate the first issue by enabling us to derive a more suitable SC unit, the transitory shocks would never converge to some known distribution. In all, if we are not ready to make strong assumptions on the underlying distributions, there is little hope to derive the distribution of the estimator under the null, even asymptotically.

To salvage the situation ADH propose a permutation-like test where the analysis is extended to the control units. For each control we estimate the "treatment effect" as if that unit potentially was affected by the intervention. In order to avoid confusion we refer to the control units as *placebos* whenever they are the investigated unit. With only some minor modifications, the estimation procedure outlined in the previous section is applicable also to the placebos. Since the treated unit may be affected by the intervention including it in the control group of the placebos would imply that the SC unit is not made up of only unaffected units, and thereby not necessarily reconstructing the counterfactual outcome. We therefore impose, not only that the investigated unit cannot act as a

control for itself, but also that the treated unit never can act as a control by restricting so $w_{i1} = 0$.[7] With these restrictions we can express the estimated effect for any i as:[8]

$$\hat{\beta}_{it} \quad = \quad \beta_{it} + \varepsilon_{it} - \sum_{j=1}^{J+1} w_{ij}\varepsilon_{jt} + \boldsymbol{\phi}_t \mathbf{B}_i. \tag{6}$$

By assumption the first term of this expression—the treatment effect—is zero for the placebos. The basis of the test proposed by ADH is the recognition that under null of no treatment effect, β_{it} is zero for all units. The estimator would in that case, for all units, only consist of the last three terms of (6). Since this is always the case for the placebos their estimators could be seen as a representation of the estimator of the treated unit when the null is true. The formal requirement is that the estimators are independent and identically distributed (IID) under the null of no treatment effect. If that requirement is met we can compare the estimate of the treated unit, $\hat{\beta}_{1t}$, with the set of estimates of the placebos, $\{\hat{\beta}_{it} : i \neq 1\}$, to assess the likelihood that they are drawn from the same underlying distribution. If the estimate for the treated differ considerably from the placebos we would deem it unlikely that they are from the same distribution, and therefore reject the null.

Building on this idea, mainly two types of tests have been used in the literature. The most prevalent has been to compare the time-series of $\hat{\beta}_{it}$ for the treated unit and the placebos graphically. Figure 1 presents an illustration of two such graphical comparisons. In the left panel the estimated effect for the investigated unit (the bold black line) is clearly positioned above all the placebo estimates (the gray lines) in the intervention periods. We would in this case deem it unlikely that we would derive this set of estimates if there was no treatment effect—we can reject the null hypothesis. In the right panel however, the estimated effects for the unit of interest, while different from zero, are not exceptional compared to the placebos. We can therefore not rule out the possibly that the deviation from zero is just by chance—we cannot reject the null.

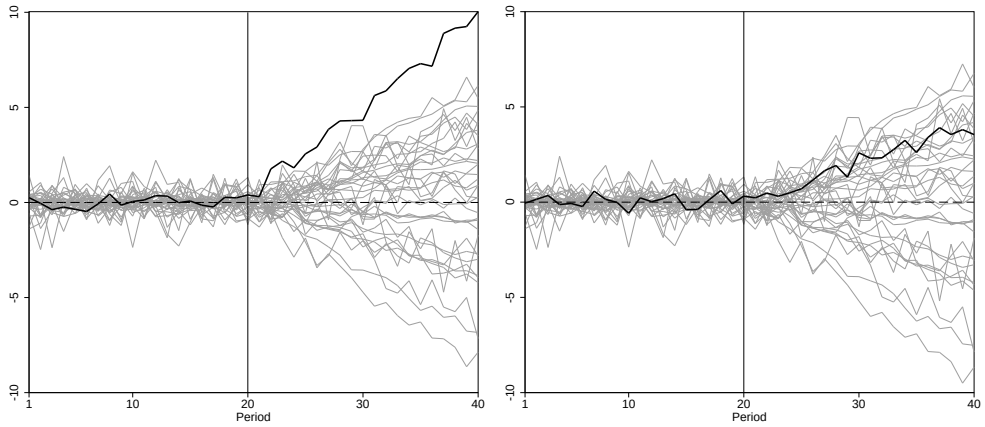
The other type is a slightly more formalized test, resulting in quantitative measures of significance such as p-values. The mean of the placebo estimates are expected to be zero in all settings, since they are unaffected by the intervention.[9] A fruitful comparison

[7] If the only propose is to produce a valid test this restriction is superfluous, as we only need the estimators to be IID under the null. Under the null the treated unit is unaffected by the intervention and could therefore act as a control. Subsequently, many of the previous applications has not imposed this restriction. However for the placebo estimates to be representative also when the null is false, which for example is implicit in the graphical comparison presented below, this restriction is needed. For completeness we will here impose it, but as seen in the simulation study it has, at most, a marginal effect in that setting.

[8] Henceforth we let $\hat{\beta}_{it}$ denote the $\hat{\beta}_{it}(\mathbf{W}_i)$ that utilizes the optimal $\mathbf{W}_i$, and similarly with $P_{it}(\mathbf{W}_i)$ and $\mathbf{B}_i(\mathbf{W}_i)$.

[9] While we cannot be certain that (2) is fulfilled for each placebo, in fact we have strong reasons

Figure 1: Illustration of graphical hypothesis test



Notes: The figure shows an illustration of the graphical hypothesis test. Each line represent the time-series of the difference between the actual outcome and the outcome of the SC unit, i.e. the estimated effect. The black line is the time-series for the treated unit while the gray lines are the series for the placebos. The vertical line indicates the start date of the intervention. The left panel shows a situation where we are motivated to reject the test, while the right panel shows a situation where we are not.

is therefore how much the estimates deviate from zero. When using this test we first calculate the rank of the absolute deviation from zero of the estimate of the treated unit among the estimates for all units. Given the IID assumption and under the null of no treatment effect, the distribution of the rank of the treated unit is uniform. We can thereby derive the probability of observing the actual set of estimates under the null, resulting in a p-value. In practice, let $R_t = \sum_{i=1}^{J+1} \mathbb{1}\left(|\hat{\beta}_{1t}| \leq |\hat{\beta}_{it}|\right)$ denote the rank of the absolute value of the estimated effect for the treated unit in t among all units, where $\mathbb{1}(\cdot)$ is the indicator function. Then, under the null of no effect, the probability that the rank would be R_t or lower is $R_t/(J+1)$, which provides us with our p-value. If $R_t/(J+1)$ is lower than the specified significance level we reject the null.[10]

3.1 The IID condition

In this section we will discuss the required IID condition, specifically why we have reason to believe that it do not hold in most situations. While there are settings where

to expect the opposite, the distributions of the covariates and factors loadings will not systematically differ between the placebos. Thus the expected estimated effect over all placebos is zero.

[10]This test investigates the effect in a single intervention period. We could however easily extended it to a test of the period-average treatment effect. We would in that case simply exchange $\hat{\beta}_{it}$ above for the average effect calculated post-estimation, namely $\hat{\beta}_i^A = \frac{1}{T-T_0}\sum_{t=T_0+1}^{T}\hat{\beta}_{it}$.

the condition would hold, the following discussion indicates that they are few and far between. The lack of acknowledgment previously in the literature is therefore notable.[11] The consequences of a violation are however not obvious—it could lead to both over- and under-rejection. Furthermore, it does not necessarily render the test uninformative.

Our investigation is divided into two parts. First we will discuss the IID condition under the assumption that we have found weights so that the counterfactual outcome can be reconstructed, up to the transitory shocks, for the treated unit. In other words, we assume that the derived weights fulfill (2). We show that the IID condition does not hold in this situation except under some very exceptional circumstances, which nearly never would be true in a real application. Then we investigate a setting where (2) only holds in an approximate sense and show that the IID condition is unlikely to be fulfilled also in this setting.

In the first situation, where (2) hold for the treated unit, the IID condition would require that (2) hold for all placebos as well. If that is the case then all estimators are only linear combinations of the transitory shocks and we would, for all intends and purposes, be able to consider them IID.[12] This is however a largely hypothetical situation—it would require that a large number of units share the exact covariates and factors loadings.

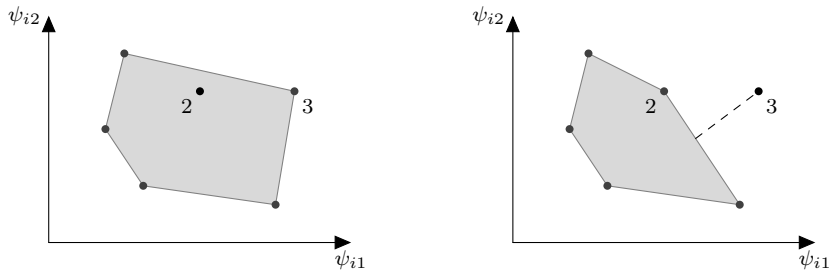
To see this we first briefly remind us about the geometric interpretation that was mentioned above. As then discussed there exists a set of weight so that (2) hold if, and only if, $\boldsymbol{\psi}_i$ is in the convex hull formed by $\{\boldsymbol{\psi}_j : j \neq i, 1\}$. Notice that this has a direct correspondence to the definition of $\mathbf{B}_i$—it is the difference between $\boldsymbol{\psi}_i$ and the point in the convex hull that is described by the weights $(\sum_{j=1}^{J+1} w_{ij}\boldsymbol{\psi}_j)$. The Euclidean norm of $\mathbf{B}_i$ would thus give us the distances between these two points. When (2) is true these two points are the same and subsequently we have that $\|\mathbf{B}_i\| = 0$. In fact, an algorithm that is devised as to derive weights to fulfill (2) could be interpreted as to try to minimize the norm of $\mathbf{B}_i$. We will denote the minimal norm for unit i as $\|\mathbf{B}_i^*\| \equiv \min_{\mathbf{W}_i} \|\mathbf{B}_i\|$. A necessary condition for that the derived weights make (2) hold for all units is that $\|\mathbf{B}_i^*\| = 0$ is true for all i.

[11] An often stated condition for the validity of this test is that treatment should be as-if randomly assigned among the units. Leaving aside that this seldom is the case, this condition would only ensure that the are no confounders. The other issues discussed in this section would still be relevant.

[12] Note however that in the strictest sense, the IID condition is not even fulfilled when (2) holds true for all units. The transitory shock of unit i would enter the estimator of i but also the estimators for all units j with $w_{ji} > 0$, thereby violating the independence condition. Furthermore, conditional on the weights the variance of $\sum_{j=1}^{J+1} w_{ij}\varepsilon_{jt}$ will depend on the concentration of the weights—few units with high weights results in higher variance than many units with low weights. Since the concentration is likely to differ between units the condition of identically distributed estimators would be violated. While this indicates that the test never can be perfect, we would expect these violations to have relatively small implications in an actual application.

Because the placebos also act as controls, a situation where $\|\mathbf{B}_i^*\| = 0$ is true for all i will be a very rare encounter. Assume that for some unit $i \neq 1$ we have that $\|\mathbf{B}_i^*\| = 0$ hold. This would then, as mentioned, imply that $\boldsymbol{\psi}_i$ is inside the convex hull formed by $\{\boldsymbol{\psi}_j : j \neq i, 1\}$. Now consider another unit $k \neq 1$ which is a vertex of this convex hull. This unit must either be outside its convex hull, implying $\|\mathbf{B}_i^*\| > 0$, or there must exist a duplicate unit, l, for which we have $\boldsymbol{\psi}_l = \boldsymbol{\psi}_k$. Assume that k does not have a duplicate, then since k is a vertex in the hull of i no other unit in $\{\boldsymbol{\psi}_j : j \neq i, 1\}$ would extend it to the point $\boldsymbol{\psi}_k$ (by the definition of a vertex). And since $\boldsymbol{\psi}_i$ is in its convex hull, neither it can extend the hull to $\boldsymbol{\psi}_k$. That is, $\boldsymbol{\psi}_k$ must be outside the convex hull formed by $\{\boldsymbol{\psi}_j : j \neq k, 1\}$. On the other hand, if there exists a duplicate (either i or some other unit) it would take the place of k and thereby extend the convex hull to $\boldsymbol{\psi}_k$.

Figure 2: Illustration of the existence of imperfect SC units



Notes: The figure shows an illustration of why, in general, not all units can have a SC unit that reproduce the counterfactual outcome, net of transitory shocks. In this illustration six control units are plotted and there are two common factors. In the left panel unit 2 is the investigated unit which is positioned inside the convex hull formed by the remaining controls. In the right panel unit 3 is the investigated unit and since it is a vertex of the convex hull for unit 2 it will be positioned outside its convex hull. The SC unit for unit 3 could therefore never reproduce the counterfactual outcome exactly.

The reasoning above is illustrated in Figure 2, where we have plotted $\boldsymbol{\psi}_i$ for all placebos in a setting where $J = 6$ and $\boldsymbol{\psi}_i$ consists of two elements. In the left panel unit 2 is the unit currently investigated. The convex hull formed by the remaining units is marked by the shaded surface. Since it is positioned in the hull we have $\|\mathbf{B}_2^*\| = 0$. In the right panel we change focus to unit 3. This unit is a vertex in the convex hull of unit 2, and will therefore be positioned outside its convex hull. Subsequently we have $\|\mathbf{B}_3^*\| > 0$ (here illustrated with a dashed line). The only other possibility would be if there exist another unit, l, for which we have $\boldsymbol{\psi}_l = \boldsymbol{\psi}_3$, in which case the convex hull would be as in the left panel.

Notice that this applies to all units that are vertices in the convex hull for any other unit with $\|\mathbf{B}_i^*\| = 0$ (which coincide with the vertices of the convex hull of the treated unit). In order for $\|\mathbf{B}_i^*\| = 0$ to be true for all i we would therefore require that each of

these units have a duplicate unit—a highly unlikely situation.

Motivated by this result, in the remaining discussion we will focus on the case where at least some $\|\mathbf{B}_i^*\|$ deviate from zero, that is we have $E(\|\mathbf{B}_i^*\|) > 0$. In general we would require that the expected distance to the convex hull for the placebos is equal to the expected distance for the treated unit, that is $E(\|\mathbf{B}_1^*\|) = E(\|\mathbf{B}_i^*\|)$ for any $i \neq 1$. If not, $\boldsymbol{\phi}_t \mathbf{B}_i$ in (6) is expected to differ systematically between the treated unit and the placebos, implying estimators that follow different distributions. There are three points that could be raised as to why we have reason to believe this is not the case.[13]

First, notice that the size of the convex hull will affect $\|\mathbf{B}_i^*\|$. The larger the convex hull the higher probability that $\boldsymbol{\psi}_i$ is in it, everything else equal. The size of the convex hull weakly increases in the number of points used in its construction, in this case how many controls are used when deriving the weights. In the permutation test the treated unit has access to J controls—all unaffected units. The placebos on the other hand have only access to $J-1$ controls since the treated unit is prohibited to act as a control. Only considering this factor would lead us to believe that $E(\|\mathbf{B}_i^*\|)$ is lower for the treated unit than for the placebos.

Second, the distribution of the convex hull relative to the distribution of $\boldsymbol{\psi}_i$ for the investigated unit will also affect the expected distance. If the two distributions are similar there is a high probability that the realized $\boldsymbol{\psi}_i$ is in, or close to, the realized convex hull. Notice that the convex hull is always made up of units unaffected by the intervention. Its distribution will thus depend on the distribution of $\boldsymbol{\psi}_{i\neq 1}$. If there are confounders (i.e. when $E(\boldsymbol{\psi}_1)$ differ from $E(\boldsymbol{\psi}_{i\neq 1})$) we would expect the distribution of the convex hull of the treated unit to differ substantially from the distribution of $\boldsymbol{\psi}_1$. This would however not be the case with the placebos. Their convex hull will always be constructed by other placebos, which share the same distribution of $\boldsymbol{\psi}_i$. In short, the convex hull is always expected to cover a large part of the sample space of the random variable that $\boldsymbol{\psi}_{i\neq 1}$ is drawn from. This is not necessarily true for the random variable that $\boldsymbol{\psi}_1$ is drawn from when there are confounders. Or yet in other words, if there are confounders the treated unit will differ systematically from the controls and thereby constraining our ability to construct a suitable SC unit. In isolation, this factor would imply a higher $E(\|\mathbf{B}_i^*\|)$ for the treated unit.

Third, and a slightly more subtle point, is due to the practice of selectiveness in application. As mentioned in Section 2, in order to minimize the risk of bias the researcher

[13] We leave the investigation of whether the estimators are independent as this arguably is secondary to whether they are identically distributed. There is however reason to believe that they are negatively correlated. To see this, notice that if $\|\mathbf{B}_i^*\|$ is high for some $i \neq 1$ this implies that $\boldsymbol{\psi}_i$ is positioned far from its convex hull. Since i will be included in the set that form the convex hull for all other units it will tend to extend their hull, implying a lower $\|\mathbf{B}_j^*\|$ for $j \neq i$.

is recommended to investigate the pre-intervention fits of the chosen estimator. If the fits deviate from zero to a large extent, we would expect the estimator to behave badly and are therefore motivated to use another method in the particular application. While this ensures that the SC method is only used when it is close to unbiased, it does not ensure that the hypothesis test is valid. If there are confounders, so that the issues discussed in the previous paragraph have a large influence, then the selectiveness could improve the testing (in realized applications) since settings where confounders are particularly problematic are discarded.[14] There are however no way of knowing the optimal level of selectiveness. While the point estimate monotonically improves the more restrictive we are, from the perspective of the hypothesis test we can be too restrictive. The more restrictive we are the lower $E(\|\mathbf{B}_1^*\|)$ will be in the remaining applications, and at some point it will be lower than $E(\|\mathbf{B}_i^*\|)$ for the placebos. For instance, in a setting without confounders any level of selectiveness will imply that the expected distances differ.

In any particular application the total effect of these factors is not clear, and they are not necessarily present in all situation. The first and third point tend to make the SC more suitable for the treated unit, which imply less deviation from zero under the null and thereby under-rejection. The second point tends to make the SC unit less suitable for the treated, and thereby over-rejection. The total effect depends on the specific setting, which cannot be observed. In some applications they might, more or less, cancel each other and thereby produce an informative test. It should however be noted this would happen by chance. There is no automatic force that would make the total effect negligible. We can therefore conclude that in most applications there is a risk that the hypothesis test is affected by these issues.

4 Alternative tests

ADH discusses two additional tests that potentially could mitigate some the problems arising when the IID condition is violated. Both these tests exploit the fact that the pre-intervention fits are informative not only for the treated unit but also for the placebos. This information could aid us in creating a pool of comparison units that is more similar to the treated unit, under the null, and thereby producing a valid test. In this section we will briefly discuss the two additional tests by ADH and then introduce our test.

Starting with the first alternative test by ADH, remember that the issues discussed in the previous section are all rooted in that $\mathbf{B}_i$ is expected to differ systematically between the treated unit and the controls. We can however, in principle, observe this

[14]Note however that the selectiveness do not imply that we limit the attention to situations where there is no confounders. We only discard cases where SC method is unable to correct for them.

difference for the same reasons that enabled us to derive the weights. Since there is no effect of the intervention in the pre-intervention periods, the only possible source of systematic difference in the pre-intervention fits must come from $\boldsymbol{\phi}_t \mathbf{B}_i$. Restricting the pool of placebos to only those that share similar pre-intervention fits with the treated unit would thereby result in a more valid comparison group. ADH propose that we exclude all placebos with a mean square pre-intervention fit that is greater than some factor ζ of the mean square fit of the treated unit. That is, only units i that fulfill:

$$\sum_{t=1}^{T_0} P_{it}^2 \leq \zeta \sum_{t=1}^{T_0} P_{1t}^2, \tag{7}$$

are to be compared to the treated unit. If we set ζ so that there are only negligible difference between $\mathbf{B}_i$ for the remaining placebos and the treated unit, their distributions would be (roughly) the same and thereby render the test valid. There are however an unattractive consequence with this test—it excludes comparison units. Ideally we would like to set ζ so that there is no difference between the remaining units—in the extreme, only to include placebos that have identical pre-intervention fits as the treated unit. There are however often no units, or only very few, that falls under this category. So even if the remaining units are very similar to the treated unit, they are so few that the test would be uninformative for this reason (e.g. in the rank test the p-value is in chunks of $1/(J+1)$). There is, in other words, a trade-off between the number of comparison units and their similarity with the treated unit. Since the solution to this trade-off is not obvious, the researcher is provided with some leeway in with this test. This leeway is not trivial since, in many situations, the test would reject the null for some ζ but not for others. ADH propose that several tests with different ζ are used which partly circumvent this problem (they use $\zeta = 20$, $\zeta = 5$ and $\zeta = 2$). It is however not obvious which conclusions should be drawn when the different levels of ζ present diverging results.

The second alternative test discussed by ADH tries, instead of excluding placebos, to account for their differences in a more direct manner. Specifically, if the pre-intervention fits deviate from zero to a large extent for a certain unit, we would expect the estimates of its effects to do the same. Fundamental to this test is the idea that the *magnitude* of the deviation in the pre-intervention periods is informative of the *magnitude* of the deviation in the estimator. By scaling the estimates with the pre-intervention fits we would thereby be able to, in some sense, control for the problematic difference. In practice we would calculate the ratio between the mean squared estimated effect and

the mean squared pre-intervention fit for all units. That is, we calculate:

$$Q_i = \frac{\frac{1}{T-T_0}\sum_{t=T_0+1}^{T}\hat{\beta}_{it}^2}{\frac{1}{T_0}\sum_{t=1}^{T_0}P_{it}^2}, \tag{8}$$

for all i. This statistic could then be either graphed or be used in a rank test as above. In the rank test, if the rank of Q_1 among all Q_i is less than the cut-off implied by the stipulated significance level we would reject the null. This test has the attractive feature that, while accounting for eventual difference in the distributions, all units remain in the comparison pool and thereby maintain the highest possible level of detail. Furthermore, the procedure is completely standardized—there are no free parameters to be chosen—so no leeway is given to the researcher who wants to apply it. The main disadvantage is however that it is rather crude. When deriving the statistic we would effectively scale the treatment effect with the pre-intervention fits, thereby reducing the power of the test. Furthermore, it implicitly assumes that the deviation in pre-intervention fit is proportional to the deviation in estimator, which is not necessarily the case.[15]

The disadvantages inherent in these two tests motivate us to seek alternatives. In the remainder of this section we will introduce and discuss a new test procedure that does not suffer from these issues. It also builds on the insight that the pre-intervention fits provide useful information. Compared to the previous tests we will however derive a more detailed relation, based on the parametric assumptions.

Consider the estimator from above, but with some minor alternations concerning the definition of the weights. The alternative weights are denoted with $\bar{\mathbf{W}}_i = (\bar{w}_{i1}, \cdots, \bar{w}_{i(J+1)})'$. First, we now drop the restriction that the treated unit cannot act as a control for the placebos, that is we no longer impose that $\bar{w}_{i1} = 0$. Second, to economize on notation we will change the sign of the weights (given that the composition of the SC unit is unchanged we have $\bar{w}_{ij} = -w_{ij}$) and change the restriction of the units' own weights from $w_{ii} = 0$ to $\bar{w}_{ii} = 1$. With these changes the vector of weights describes the difference between the investigated unit and the convex combination of control, rather than only the convex combination. The "estimator," $\bar{\beta}_{it}$, resulting from these weights—the equivalent to (6)—is then given by:

$$\bar{\beta}_{it} = \sum_{j=1}^{J+1}\bar{w}_{ij}\left(\beta_{jt} + \varepsilon_{jt}\right) + \boldsymbol{\phi}_t\bar{\mathbf{B}}_i, \tag{9}$$

[15] Another unappealing property, while seldom an issue, is that test in theory could reject the null hypothesis of no period-average effect even if the point estimate of the period-average effect is exactly zero. This would arise when the estimates of the treated, although have zero mean, varies much more than the placebos.

where $\bar{\mathbf{B}}_i = \sum_{j=1}^{J+1} \bar{w}_{ij} \boldsymbol{\psi}_j$ is the difference between the covariates and factor loadings when using $\bar{\mathbf{W}}_i$. Notice that the first issue discussed in Section 3.1—different sized control groups—no longer is relevant. When constructing $\bar{\beta}_{it}$, all units have a control group of size J. We will now show that the last term of this expression, $\boldsymbol{\phi}_t \bar{\mathbf{B}}_i$, can be expressed as a linear combination of the observed pre-intervention fits.[16]

Notice that the pre-intervention fits given by $\bar{\mathbf{W}}_i$ are:

$$P_{it} = \sum_{j=1}^{J+1} \bar{w}_{ij} \varepsilon_{jt} + \boldsymbol{\phi}_t \bar{\mathbf{B}}_i.$$

Now stack all pre-intervention fits for unit i into a $[T_0 \times 1]$ column vector as such $\mathbf{P}_i = (P_{i1}, \cdots, P_{iT_0})'$. Then, since $\bar{\mathbf{B}}_i$ is constant over all periods, we have:

$$\begin{aligned} \mathbf{P}_i &= \sum_{j=1}^{J+1} \bar{w}_{ij} \tilde{\boldsymbol{\varepsilon}}_j + \tilde{\boldsymbol{\phi}} \bar{\mathbf{B}}_i, \\ \Rightarrow \tilde{\boldsymbol{\phi}} \bar{\mathbf{B}}_i &= \mathbf{P}_i - \sum_{j=1}^{J+1} \bar{w}_{ij} \tilde{\boldsymbol{\varepsilon}}_j, \end{aligned} \tag{10}$$

where $\tilde{\boldsymbol{\varepsilon}}_i = (\varepsilon_{i1}, \cdots, \varepsilon_{iT_0})'$ is a $[T_0 \times 1]$ column vector and $\tilde{\boldsymbol{\phi}} = (\boldsymbol{\phi}_1', \cdots, \boldsymbol{\phi}_{T_0}')'$ is a $[T_0 \times (r+F)]$ matrix.

Assume that $\tilde{\boldsymbol{\phi}}' \tilde{\boldsymbol{\phi}}$ has full rank so that $\left(\tilde{\boldsymbol{\phi}}' \tilde{\boldsymbol{\phi}}\right)^{-1}$ exists. This will in general be true if the number of elements in $\boldsymbol{\phi}_t$ is less or equal to the number of pre-intervention periods (i.e. $r + F \leq T_0$). We can then, for any $[1 \times (r+F)]$ row vector $\mathbf{v}$, express $\mathbf{v}\bar{\mathbf{B}}_i$ as:

$$\begin{aligned} \mathbf{v}\bar{\mathbf{B}}_i &= \mathbf{v} \left(\tilde{\boldsymbol{\phi}}' \tilde{\boldsymbol{\phi}}\right)^{-1} \tilde{\boldsymbol{\phi}}' \tilde{\boldsymbol{\phi}} \bar{\mathbf{B}}_i, \\ &= \boldsymbol{\Phi}(\mathbf{v}) \mathbf{P}_i - \sum_{j=1}^{J+1} \bar{w}_{ij} \boldsymbol{\Phi}(\mathbf{v}) \tilde{\boldsymbol{\varepsilon}}_j, \end{aligned}$$

where the last equality follows from (10) and $\boldsymbol{\Phi}(\mathbf{v}) = \mathbf{v} \left(\tilde{\boldsymbol{\phi}}' \tilde{\boldsymbol{\phi}}\right)^{-1} \tilde{\boldsymbol{\phi}}'$ is a $[1 \times (r+F)]$ row vector.

Note that $\mathbf{v}$ can be any $[1 \times (r+F)]$ vector—including the unknown realizations of $\boldsymbol{\phi}_t$ in the intervention periods. Therefore by setting $\mathbf{v} = \boldsymbol{\phi}_t$ we can express $\boldsymbol{\phi}_t \bar{\mathbf{B}}_i$ in any

[16] The following procedure is in some aspects similar to the procedure used in the proof presented in Appendix B in ADH.

period as:

$$\boldsymbol{\phi}_t\bar{\mathbf{B}}_i \quad = \quad \boldsymbol{\Phi}_t\mathbf{P}_i + \sum_{j=1}^{J+1} \bar{w}_{ij}\boldsymbol{\Phi}_t\tilde{\boldsymbol{\varepsilon}}_j, \tag{11}$$

where we define $\boldsymbol{\Phi}_t = \boldsymbol{\Phi}(\boldsymbol{\phi}_t)$. In other words, we can reconstruct the unknown term $\boldsymbol{\phi}_t\bar{\mathbf{B}}_i$ with a linear combination of all pre-intervention fits and a weighted linear combination of all shocks in the pre-intervention periods. To see the intuition behind this result, notice that if the number of elements in $\boldsymbol{\phi}_t$ is less or equal to the number of pre-intervention periods then there is at least one linear combination of the pre-intervention $\boldsymbol{\phi}_t$ that reconstructs any $\boldsymbol{\phi}_t$ in the intervention period—that is $\{\boldsymbol{\phi}_t : t \leq T_0\}$ span the complete space of possible realizations of $\boldsymbol{\phi}_t$. This linear combination is exactly what is given by $\boldsymbol{\Phi}_t$. For illustration assume that there are no transitory shocks, in which case $P_{it} = \boldsymbol{\phi}_t\bar{\mathbf{B}}_i$, then $\boldsymbol{\Phi}_t\mathbf{P}_i$ would effectively be the linear combination of $\{\boldsymbol{\phi}_t\bar{\mathbf{B}}_i : t \leq T_0\}$ which collapses into $\boldsymbol{\phi}_t\bar{\mathbf{B}}_i$ (since $\bar{\mathbf{B}}_i$ is constant over periods). Adding the shocks would add distortions so that we would not be able to reconstruct $\boldsymbol{\phi}_t\bar{\mathbf{B}}_i$ perfectly, but since the transitory shocks, by assumption, are uncorrelated with $\boldsymbol{\Phi}_t$ it would not be a systematic distortion. In other words, $\boldsymbol{\Phi}_t$ captures, in some sense, the relation between the realization of $\boldsymbol{\phi}_t$ in the pre-intervention periods and $\boldsymbol{\phi}_t$ in an intervention period.

If we know $\boldsymbol{\Phi}_t$ we could derive $\boldsymbol{\phi}_t\bar{\mathbf{B}}_i$, net of transitory shocks, and subtract that from $\bar{\beta}_{it}$ leaving only treatment effects and transitory shocks. However, in an actual setting $\boldsymbol{\Phi}_t$ is not known and this route would not be feasible. Instead we substitute $\boldsymbol{\phi}_t\bar{\mathbf{B}}_i$ in (9) for the expression derived in (11) and arrive at:

$$\begin{aligned}
\bar{\beta}_{it} &= \sum_{j=1}^{J+1} \bar{w}_{ij}\left(\beta_{jt} + \varepsilon_{jt}\right) + \left(\boldsymbol{\Phi}_t\mathbf{P}_i - \sum_{j=1}^{J+1} \bar{w}_{ij}\boldsymbol{\Phi}_t\tilde{\boldsymbol{\varepsilon}}_j\right), \\
&= \sum_{j=1}^{J+1} \bar{w}_{ij}\left(\beta_{jt} + \varepsilon_{jt} - \boldsymbol{\Phi}_t\tilde{\boldsymbol{\varepsilon}}_j\right) + \boldsymbol{\Phi}_t\mathbf{P}_i, \\
&= \sum_{j=1}^{J+1} \bar{w}_{ij}\alpha_{jt} + \boldsymbol{\Phi}_t\mathbf{P}_i,
\end{aligned}$$

where we in the last equality have defined $\alpha_{it} = (\beta_{it} + \varepsilon_{it} - \boldsymbol{\Phi}_t\tilde{\boldsymbol{\varepsilon}}_i)$. Two things are particularly noteworthy in this expression. First, since $\boldsymbol{\Phi}_t$ is given by $\boldsymbol{\phi}_t$ (and therefore uncorrelated with the shocks) we have $E(\alpha_{it}) = \beta_{it}$. If we could derive α_{it}, it would be an unbiased estimator of the treatment effect (also when $\|\mathbf{B}_i\| > 0$). Second, all α_{jt} and $\boldsymbol{\Phi}_t$ are constant in $\bar{\beta}_{it}$ over i (for a given t)—what differ is the weights and pre-intervention fits. Put differently, $\bar{\beta}_{it}$ is a linear combination of observed variables

and unknown coefficients common to all units.

To economize on notation, let $\mathbf{A}_t = (\alpha_{1t}, \cdots, \alpha_{(J+1)t}, \mathbf{\Phi}_t)'$ be a $[(J+1+T_0) \times 1]$ column vector that collects the unknown coefficients and $\mathbf{S}_i = (\bar{w}_{i1}, \cdots, \bar{w}_{i(J+1)}, \mathbf{P}_i')$ be an $[1 \times (J+1+T_0)]$ row vector that collects all observed variables. We could then write the above expression as $\bar{\beta}_{it} = \mathbf{S}_i \mathbf{A}_t$. Since $\mathbf{A}_t$ is constant over i, the set containing all $\bar{\beta}_{it}$ in a given period would produce a system of linear equations. Stacking the $\bar{\beta}_{it}$ in t into a $[(J+1) \times 1]$ column vector $\bar{\boldsymbol{\beta}}_t = (\bar{\beta}_{1t}, \cdots, \bar{\beta}_{(J+1)t})'$ and stacking all $\mathbf{S}_i$ into a $[(J+1) \times (J+1+T_0)]$ matrix, $\mathbf{S} = (\mathbf{S}_1', \cdots, \mathbf{S}_{(J+1)}')'$, we get the system $\bar{\boldsymbol{\beta}}_t = \mathbf{S}\mathbf{A}_t$.

The main caveat here is that this system is underdetermined—there are $(J+1+T_0)$ unknowns but only $(J+1)$ equations. Furthermore, since the SC unit is constrained to a convex combination of the controls the weights are perfectly collinear. Both these facts imply that neither $\mathbf{S}^{-1}$ nor $(\mathbf{S}'\mathbf{S})^{-1}$ will ever exist, thus we cannot solve for $\mathbf{A}_t$ nor estimate it using for example ordinary least square. This shatters the hopes to use α_{it} as an estimator for β_{it}—we can never derive it. However, for the purpose of the hypothesis test we would only require that all α_{it} follow the same distribution under the null, and preferably that they do not under the alternative. We could therefore accept a procedure where we derive α_{it} with a bias as long as it is, on average, equal between the units. We will use two methods for deriving α_{it}.

With the first method we will force a unique solution using the (Moore–Penrose) pseudoinverse of $\mathbf{S}$, denoted by $\mathbf{S}^+$. One property of the pseudoinverse is that if $\bar{\boldsymbol{\beta}}_t = \mathbf{S}\mathbf{A}_t$ has any solutions then $\mathbf{S}\mathbf{S}^+\bar{\boldsymbol{\beta}}_t = \bar{\boldsymbol{\beta}}_t$. In that case $\hat{\mathbf{A}}_t = \mathbf{S}^+\bar{\boldsymbol{\beta}}_t$ would be one of these solution (specifically the minimum norm solution). To see this notice that $\mathbf{S}\hat{\mathbf{A}}_t = \mathbf{S}\mathbf{S}^+\bar{\boldsymbol{\beta}}_t = \bar{\boldsymbol{\beta}}_t$ due to the property of the pseudoinverse. While this solution would derive a biased $\mathbf{A}_t$, which cannot be used as a point estimator, we do not expect this bias to differ between units under the null. It could therefore be used in the permutation test.

The second method rests on the fact that we actually can increase the number of estimates artificially. While a larger number of controls always are preferred, we have no reason to prefer any estimator over another conditional on their pre-intervention characteristics. By reducing the number controls used in the analysis we could increase the number of estimates by alternating the content of the control group. Each of these estimates would be less informative than with the full set of controls, but still informative to some extent. This idea was partly exploited by Abadie et al. (2012) in a robustness test, where they drop each of the control units with positive weights, one at a time, and derive the treatment effect (after recalculating the weights). They could thereby investigate whether the estimated effect is driven by the transitory shocks in a single control. We could however just as well interpret them as separate point estimates.

Building on this notion, with this method we resample the control group several times for each investigated units and estimates the treatment effect using the reduced control group. We subject each unit to M resampling rounds, denoted by m, where we, from the full set of controls, randomly draw $J' < J$ units (with replacement) to act as the control group. After deriving the optimal weights only using the J' drawn controls (i.e. restricting the weights of the units not drawn to zero) we derive $\bar{\beta}_{it}$ as above. If the drawn set of controls does not reproduce the exact convex hull as in some other round, we expect the weights in m, which is denoted with $\mathbf{W}_{im}$, to be unique. Therefore all variables that depend on the units' identity (i.e. $\bar{\beta}_{imt}$, $\mathbf{B}_{im}$ and $\mathbf{P}_{im}$) will differ between rounds. Importantly, notice that all other variables will remain constant over rounds—$\mathbf{A}_t$ is unchanged. Stacking the variables as above, so that $\bar{\boldsymbol{\beta}}_t = (\bar{\beta}_{11t}, \cdots, \bar{\beta}_{(J+1)Mt})'$ now is a $[M(J+1)\times 1]$ column vector and $\mathbf{S} = (\mathbf{S}'_{11}, \cdots, \mathbf{S}'_{(J+1)M})'$ is a $[M(J+1)\times(J+1+T_0)]$ matrix, we still have $\bar{\boldsymbol{\beta}}_t = \mathbf{S}\mathbf{A}_t$.

If we set M so that $M(J+1) > (J+1+T_0)$ we have moved from an underdetermined system to an overdetermined. However, $(\mathbf{S}'\mathbf{S})^{-1}$ still do not exist since the weights are restricted to convex combinations also in the resampled case, implying perfect collinearity. To solve this we will, instead of the ordinary least squares technique, use ridge regressions—where we besides minimizing the norm of the residual ($\|\bar{\boldsymbol{\beta}}_t - \mathbf{S}\hat{\mathbf{A}}_t\|$) also minimize the norm of the coefficients ($\|\hat{\mathbf{A}}_t\|$) thereby providing a unique solution. In practice we derive the coefficients with $\hat{\mathbf{A}}_t = (\mathbf{S}'\mathbf{S} + \mathbf{I})^{-1}\mathbf{S}'\bar{\boldsymbol{\beta}}_t$.[17]

Regardless of how we derive $\hat{\mathbf{A}}_t$ the remaining procedure is the same. The first $J+1$ elements of $\hat{\mathbf{A}}_t$, denoted with $\hat{\alpha}_{it}$, will be biased versions of α_{it}. If the bias is not systematically different between the treated unit and the placebos we would expect that $\hat{\alpha}_{it}$ is IID under the null of no effect. Employing $\hat{\alpha}_{it}$ in a permutation-based test would in that case produce a valid test of the hypothesis. If α_{it} is not too distorted by the procedure we expect that $\hat{\alpha}_{it}$ is not IID under the alternative hypothesis thereby providing an informative test.

While this test seems promising it should be stressed that it relies heavily on the parametric assumptions—deviations from them could potentially lead to serious complications. While this do not necessarily invalidates its use, it should be approached with caution. We therefore present this test as a complement to the existing tests, which, arguably, are more robust against misspecification. Some indications of the performance of the tests in a setting similar to actual applications are presented in the following section.

[17] Notice that there is a close connection between these two methods, as will be apparent in the simulation study, where the pseudoinverse can be derived as a limiting case of the ridge regression estimator: $\mathbf{S}^{+} = \lim_{\kappa\to 0}(\mathbf{S}'\mathbf{S} + \kappa^2\mathbf{I})^{-1}\mathbf{S}'$.

5 Simulation study

To investigate the issues discussed above we conduct a simulation study in two parts. The first part is a controlled simulation study where the data is randomly generated with an artificial data generating process. By carefully choosing the generating process we can investigate the issues affecting the validity of the hypothesis test separately. In particular we will use five different specifications, starting with a relatively simple setting and then stepwise adding complexity.

The second part consists of simulations based on actual data where we randomly pick units to which we assign fictitious interventions. Since this setting is closer to the typical situation where the SC method is used, at least with respect to the data generating process, it will be informative of the performance of tests in an actual application.

For the five specifications in the first part we generate data for 40 units ($J = 39$) in 35 periods ($T = 35$) and set the first 20 periods to be the pre-intervention periods ($T_0 = 20$). This setting is chosen as to reflect a typical SC application. Common to all these specifications is that the data is generated with the common factor model from the previous analysis:

$$Y_{it}^N = \delta_t + \boldsymbol{\theta}_t \mathbf{Z}_i + \boldsymbol{\lambda}_t \boldsymbol{\mu}_i + \varepsilon_{it},$$

where $\mathbf{Z}_i$ contain three elements and $\boldsymbol{\mu}_i$ contain six. We set δ_t to be a random walk (starting at zero with a standard normal shock term) and let $\boldsymbol{\theta}_t$, $\mathbf{Z}_i$ and ε_{it} be independent standard normal. The specifications differ in how we model the common factors and the factor loadings. Specifically, labeling the specifications with the letters A to E, we set:

A. The common factors and the factor loadings to be independent standard normal, that is $\boldsymbol{\lambda}_t \sim \mathcal{N}_6(\mathbf{0}, \mathbf{I})$ and $\boldsymbol{\mu}_i \sim \mathcal{N}_6(\mathbf{0}, \mathbf{I})$.

B. Like A, but where we add confounders by setting $\boldsymbol{\mu}_1 \sim \mathcal{N}_6(\mathbf{1}, \mathbf{I})$.

C. Like B, but where we increase the variance of the common factors so that $\boldsymbol{\lambda}_t \sim \mathcal{N}_6(\mathbf{0}, 2\mathbf{I})$.

D. Like B, but where we shift the mean of the common factors so that $\boldsymbol{\lambda}_t \sim \mathcal{N}_6(\mathbf{1}, \mathbf{I})$.

E. The combination of C and D, where we both increase the variance and shift the mean of the common factors so that $\boldsymbol{\lambda}_t \sim \mathcal{N}_6(\mathbf{1}, 2\mathbf{I})$.

The different specifications will have predictable consequences on the test discussed in Section 3. In a setting where we do not restrict the applications by the pre-intervention

fits, the only issue present in the first specification (A) is that the size of the control group differs between the units. If this is an important factor we expect that we here under-reject the test. In the following four specifications we have introduced confounders. Subsequently, both the first and the second issue discussed in Section 3.1 are relevant. Since they affect the test in different directions—one leading to under-rejection and the other to over-rejection—the total effect in this case is not clear.

Between the specifications with confounders (B to E) we expect the changes to the distributions of the common factors to influence the test. There are two possible effects following such a change. Whenever a SC unit does not reconstruct the counterfactual outcome exactly, net of transitory shocks, the common factors will determine the consequences of that deviation. Specifically, the common factors will decide to which level any eventual difference in $\mathbf{B}_i$ is amplified. If the common factors are close to zero, then even large deviations would not affect the estimates to any greater degree. The test would in that case remain informative. As we move from B to E the common factors deviate more and more from zero. For a given level of fitness of the SC unit we would therefore expect the performance of the test to deteriorate as we move to E.

The second effect counteracts the first. Common factors deviating from zero will also influence the pre-intervention fits. It will therefore be easier to spot unsuitable SC units—leading to more suitable SC units, on average. In the pre-intervention periods the transitory shocks will reduce our ability to observe the effect of the common factors thereby adding distortions to the weights. If the common factors deviate more from zero this reduces the relative influence of the shocks and thereby improves the signal-to-noise ratio. This effect will thus lead to better weights and reduce the impact of the confounders. Since these two effects counteract each other, it is not clear if the effect of the confounders is mitigated or magnified as we move from specification B to E.

In order to investigate the third issue—selectiveness in application—we must select a subset of the generated data sets to analyze. As discussed, the primary purpose of this selectiveness is to ensure unbiased point estimates. There is however no standardized condition to select applications. While not completely mirroring a true application, we will in this study impose the selectiveness by excluding half of the data sets with the highest mean squared pre-intervention fit of the treated unit. This procedure will suffice to investigate how restricting the applications is affecting the test. However from the perspective of achieving unbiased estimates this may be both too restrictive and not restrictive enough, depending on the specification. For instance, whenever the common factors are symmetrically centered on zero (as in specification A to C) the estimator is unbiased for any set of weights. Selectiveness would in this case only reduce the variance of the estimator. In an actual application we are however unable to observe which type

of setting we are in and would therefore always be selective to some degree. For this reason selectiveness is relevant to all specifications, even if they are unbiased from the start. Restricting the applications in this way would improve the SC unit for the treated unit, but not the placebos, resulting in decreased rejection rates in all specifications.

In the second part, referred to as specification F, we use quarterly wage data for U.S. counties. In particular we will investigate how a fictitious intervention affects the wage level for a randomly chosen county. We compile the data set used in this simulation from three sources. The main source is the *Quarterly Census of Employment and Wages*, available from the Bureau of Labor Statistics, which provides the average quarterly wage level for each U.S. county between the first quarter of 1990 to the third quarter of 2012. It also provides industry sector employment shares which will be used as covariates. The second source is the *Local Area Unemployment Statistics*, also from the Bureau of Labor Statistics, which provides monthly unemployment rates for each county during the relevant period. Last we use data on the population level of each county in 1990, as provided by the U.S. Census Bureau.

We restrict our focus to the states that contain at least 21 counties with non-missing variables during the relevant periods, resulting in 2 940 counties in 39 states. For each simulation round we randomly draw one of these counties to act as the treated unit, and the remaining counties in the same state act as controls. In most states there are more than 40 counties. For these states we restrict the control group to the 39 counties with a population in 1990 that is closest to the treated county. As in the first part, we use 20 pre-intervention periods and 15 intervention periods. Of the 91 available quarters we randomly assign the intervention start date to any quarter between the first quarter in 1995 and the first quarter of 2009, thereby ensuring an adequate number of quarters before and after the intervention. The data set is then constructed by setting Y_{it}^N to the wage level, in the form of average weekly wage, in county i in quarter t. When analyzing the power of tests we will artificially add an effect of the intervention to the outcome, as detailed below. Three covariates are also included in the analysis, in the form of the pre-intervention average unemployment rate for each county and the pre-intervention average sector employment shares of government institutions and of service providing industries.

With this specification treatment is randomly assigned to the units—by construction there are no confounders. In this aspect specification F is most similar to A. While this may not perfectly reflect a true application, where there often would be confounders, the purpose of this part is to investigate the effect of a more realistic data generating process.

For each generated data set we will investigate the performance of all five test dis-

cussed in this paper. Three of them are the tests proposed by ADH: the test using all placebos and thereby disregard the pre-intervention characteristics (labeled "ADH 1"), the test restricting the pool of placebos by the pre-intervention fits ("ADH 2"), and the test using the ratio between the estimates and the pre-intervention fits as test statistic ("ADH 3"). The other two are the newly introduced tests, both without and with resampling (labeled "AS 1" and "AS 2," respectively). For all tests we use their rank-based version, and test against the null hypothesis that the intervention period-average treatment effect is zero. Refer to Appendix A for a detailed discussion of the exact procedure for each test.

Table 1: Size — Rejection rates at 5% with no treatment effect.

Panel 1: Unrestricted						
	A	B	C	D	E	F
ADH 1	0.055	0.145	0.145	0.420	0.385	0.043
ADH 2	0.043	0.142	0.133	0.420	0.383	0.063
ADH 3	0.052	0.050	0.037	0.013	0.028	0.048
AS 1	0.048	0.037	0.033	0.033	0.037	0.037
AS 2	0.048	0.037	0.030	0.035	0.037	0.037
Panel 2: Restricted by pre-intervention fit						
	A	B	C	D	E	F
ADH 1	0.015	0.090	0.070	0.105	0.145	0.000
ADH 2	0.005	0.090	0.055	0.110	0.135	0.015
ADH 3	0.105	0.080	0.065	0.025	0.050	0.050
AS 1	0.050	0.025	0.040	0.040	0.070	0.035
AS 2	0.050	0.025	0.035	0.045	0.070	0.035

Notes: Each cell presents the rejection rate at the 5% significance level of the null hypothesis of no (period-averaged) effect of the intervention when the null hypothesis is true ($\beta_{1t} = 0$). Columns indicate specifications and rows the tests. The first five columns represent the specifications (A to E) that use data from the controlled data generating process, while the last column represents the specification (F) that uses real data. The details of the specifications and the tests are described in the text. The number of simulations for each specification is 400. In the first panel all generated simulations are analyzed, while in the second panel the 200 simulations with the lowest mean squared pre-intervention fit of the treated unit are used.

Table 1 presents the results of the simulation study when the null hypothesis of no treatment effect is true—the size of the tests. Particularly, each cell presents the rejection rate at the 5% significance level for each test. The columns denote the six

specifications, A to F, and the rows denote the five tests. The first panel includes all generated data sets while we in the second panel have restricted the analysis to the data sets where the treated unit has a pre-intervention fit close to zero (thereby imposing selectiveness in application). Averaged over a large number of applications, as in this analysis, we expect the rejection rate to be approximately 5% for a valid test.

Starting with the first test, ADH 1, which is the most commonly used test in the past literature, we can largely confirm the prediction made in the discussion above. In specification A in the first panel, only the first issue discussed in Section 3.1—the different sized control groups—is relevant. Since this implies a more suitable SC unit for the treated unit we expect that we would under-reject the test in this case. As it turn out this is not the case, the rejection rate is very close to 5% and even slightly higher. This indicates that this issue does not have a large influence in the current setting. This may however not hold in other settings. Here the relative difference in the control group size is quite small. In a setting with fewer controls the relative difference is larger and thus we expect this issue to be of greater importance. For instance, in a setting with 10 controls the relative size between the treated and placebos is 9/10, which is a considerably larger difference than in the current setting where it is 38/39.

Continuing to the specifications containing confounders, as presented in the next four columns, we see that the rejection rates increase, in line with the predictions. In the first two specifications (B and C), where the common factors still are standard normal, the rejection rates almost triples. Notably, when increasing the variance of the common factors, as in specification C, the rejection rate remain unchanged. It seems that the two effects discussed above—increased ability to find a suitable SC unit and amplified differences when not—counteract each other, leading to approximately the same rejection rates. On the other hand, with a mean shift in the common factors the two effects do not counteract, as seen in the next two specifications (D and E). Here the rejection rates increases sharply and we reject the test more than eight times as often as what we would expect from a valid test. A unidirectional mean shift, as in this case, turns out be particularly troublesome. This is however an exceptionally demanding situation, since we largely have prevented orthogonalization by introducing the confounders also as unidirectional mean shifts in the factor loadings. Since orthogonalization often would be a second-best alternative, this setting could be seen as an extreme case. Increasing the variance of the mean shifted common factors, as in E, however increases our ability to derive a suitable SC unit leading to rejection rates closer to the desired target. In the last specification F, where we use real data, the rejection rate is close to 5%. This is explained by the lack of confounders.

Shifting focus to the setting where we impose selectiveness, as presented in the

second panel in Table 1, we see as anticipated that the rejection rate decreases for all specifications. This manipulation leads to under-rejection in specification A and F, where the rejection rates were at the desired level in the unrestricted setting. The tests that over-rejected the test in the unrestricted case move closer to a 5% rejection rate but are still at a level higher than desired. This verifies that, while the selectiveness can counteract problems arising with confounders, the optimal level of selectiveness is situation-dependent. Furthermore notice that the validity of the test is disconnected for the bias of the estimates, for example in specification A, B and C the estimator is unbiased in all settings.

Turning to the alternative tests, we can first note that the second test, ADH 2, performs virtually identically to ADH 1. In a setting where there is no treatment effect, we tend to reject the test when the SC units for the placebos are more suitable than the SC unit for the treated. If we with the first test reject the null, then this implies that treated unit is among the units with the most deviating pre-intervention fits. There would therefore not be many placebos with worse fits, leading to few discarded units with the second test and still a rejected null. If we were to be more restrictive, requiring even higher similarity and also excluding units with better fits, the tests would differ to a larger extent. Notably, in this simulation study we have used the most conservative level of restrictiveness previously used in the literature (requiring at most twice as high mean squared fit), any higher restrictiveness would in most applications exclude nearly all placebos.

The third test ADH 3 seems, on the other hand, to perform better than the previous two. The rejection rates are close to 5% in all specification, and never deviate more than five percentage points from that target. The remaining deviation confirms that the proportionality assumption implicit in this test only is a crude approximation, but that it in most cases can adjust eventual differences to reasonable levels.

The last two tests, AS 1 and AS 2, perform on par with ADH 3. The rejection rates are always close to 5%, but with a slight tendency of under-rejection. It seems that the procedure of deriving $\hat{\alpha}_{it}$ does not introduce bias differently between the units, even when confounders are present. Notably this is true also in specification F, in which the parametric assumption that these two tests are built on is not expected to hold exactly.

We now turn to the tests ability to reject a false null hypothesis—the power of the tests. To investigate this we will artificially add a treatment effect to the generated data sets. In order to make the added effect comparable between specifications we will add a standardized effect. Specifically we calculate the pre-intervention sample standard deviation of the outcome among all units and then add a quarter of a standard deviation

Table 2: Power — rejection rates at 5 % with treatment effect.

	A	B	C	D	E	F
Panel 1: Unrestricted						
ADH 1	0.300	0.333	0.328	0.613	0.613	0.085
ADH 2	0.235	0.302	0.300	0.630	0.625	0.117
ADH 3	0.092	0.070	0.058	0.063	0.087	0.142
AS 1	0.433	0.328	0.470	0.357	0.443	0.163
AS 2	0.433	0.328	0.470	0.360	0.443	0.160
Panel 2: Restricted by pre-intervention fit						
	A	B	C	D	E	F
ADH 1	0.305	0.310	0.275	0.335	0.395	0.010
ADH 2	0.240	0.250	0.235	0.370	0.435	0.060
ADH 3	0.180	0.125	0.110	0.105	0.155	0.210
AS 1	0.515	0.400	0.580	0.440	0.540	0.185
AS 2	0.515	0.400	0.580	0.445	0.540	0.185

Notes: Each cell presents the rejection rate at the 5% significance level of the null hypothesis of no (period-averaged) effect of the intervention when the null hypothesis is false. Columns indicate specifications and rows the tests. The first five columns represent the specifications (A to E) that use data from the controlled data generating process, while the last column represents the specification (F) that uses real data. The details of the specifications and the tests are described in the text. To generate a treatment effect we add one quarter of a standard deviation of the outcome in the pre-intervention periods to the treated unit in the intervention periods, as detailed in the text. The number of simulations for each specification is 400. In the first panel all generated simulations are analyzed, while in the second panel the 200 simulations with the lowest mean squared pre-intervention fit of the treated unit are used.

as a treatment effect to the treated unit in the intervention periods.[18] Using these data sets we redo the analysis as above and again test against the null hypothesis of no treatment effect.

Table 2 presents the rejection rates from the simulations where we have added the treatment effect, thereby indicating the power of the tests. The more informative a test is the higher we expect its rejection rate to be. As above the columns denote the specification and the rows the tests. The first panel consists of all generated data sets while the second panel only includes half of the sets with the lowest mean squared pre-intervention fit of the treated unit.

Also in this case, the first two tests performs similarly, with rejection rates around 30% in the first three specifications, slightly higher than 60% in the two following and only around 10% in F. These rejection rates should however be seen in light of the size of the two tests. Given the tendency of over-rejection when the null is true, especially the severe over-rejection in specification D and E, the high rejection rates exhibited here are neither surprising nor remarkable. When we restrict the applications, as presented in the second panel in Table 2, these two tests exhibit lower rejections rates which entirely is explained by the change in size.

The third test, ADH 3, which had rejection rates close to the desired level when the null was true, presents a low power with rejection rates ranging from 10 to 20%. One possible explanation would be that this test scales the estimates with the pre-intervention fit when deriving their ratio. While this is intended since it "controls" for inherent differences in the estimates, it also scales the treatment effect. Since this scaling reduces the difference between the estimates also when the null is false, the low power is anticipated.

Turning to the last two tests, AS 1 and AS 2, we see that the power in the first five specifications is consistently higher than for any other tests (when excluding cells where the size invalidates their use). This is explained by that in all of these specifications the parametric assumption this test is based on is true. Exploiting the more detailed relationship between the pre-intervention fits and the estimates, we can control for more of the unwanted differences. Since the treatment effect is unexplained by the pre-intervention fits, unlike the other factors, it is not inadvertently controlled for, unlike in ADH 3. This enables us to control for the unwanted factors while leaving the treatment effect largely unaffected, thereby explaining the high power.

As discussed these two tests rely to a higher extent on the parametric assumptions than the other tests. It is therefore of particular interest how they perform in the spec-

[18] More in detail, we derive $SD = \frac{1}{T_0(J+1)}\sqrt{\sum_{i=1}^{J+1}\sum_{t=1}^{T_0}(Y_{it}^N - \bar{Y}^N)^2}$ where $\bar{Y}^N$ is the average outcome in the pre-intervention periods, and then set $Y_{it} = Y_{it}^N + 0.25SD$ for $i = 1$ in $t > T_0$.

ification where we use real data. Turning to column F we see that the rejection rates are less than half of what they were in the other specifications. The high performance evidently depends on to which degree the assumptions hold. However, while the performance is considerably lower than in the other specifications it is still better or on par with the other tests. In the restricted version (as in panel 2) of specification F, the first two tests are almost completely uninformative, rejecting the test even less frequently than what is expected when the null is true. The third test, ADH 3, rejects the test in approximately one fifth of the simulations, which is roughly the same as with AS 1 and AS 2. So even if the very high rejection rates require that the parametric assumptions hold true, the test performs reasonably well even when they do not.

Worth noting is that the performance of AS 1 and AS 2 are virtually the same. It seems that extending the number of estimates as in AS 2 makes little difference in our ability to derive $\hat{\alpha}_{it}$. Since AS 2 is considerably more computationally demanding than AS 1—it effectively re-runs the analysis 25 times—it seems that AS 1 is to be preferred.

6 Concluding remarks

In this paper we have studied the SC method. While it is a welcomed addition to the toolbox of the quantitative researcher—mainly because that toolbox has been short on case study methods—its performance is to some degree unknown. One aspect which has been particularly overlooked is the procedures for hypothesis testing. We hope that this study partly will bridge this gap. We have showed that in many settings the fundamental condition for a valid test, that the statistic used in the permutation test are independent and identically distributed under the null for the included units, does not hold. As a consequence the current tests could be both uninformative and misleading, depending on the exact setting and procedure followed.

These findings provide motivation to more extensively investigate the validity of test procedures in each application where they are used. From the perspective of the most commonly used test, where the estimated effects are used as test statistic, one should ensure that the method is only applied when a suitable SC unit is found for the treated unit. This is both to avoid the severe over-rejection that this test otherwise could be plagued with, and to make certain that the point estimates are unbiased. This selectiveness does however not in itself results in a valid hypothesis test, as we have seen. It is therefore essential to take further actions to ensure that the used statistic follow the same distribution for all units under the null.

Of the methods that tries to achieve this, the arguably most intuitive procedure is to restrict the comparison to placebos with similar pre-intervention characteristics and

keep the estimates as test statistic. Using this method could however be problematic. To reach a pool of placebos similar enough to the treated unit we must often exclude so many that the test is rendered uninformative. Indeed, when we employ this test in the simulation study, there is essentially no difference in rejection rates compared to the original version. This indicate that we are motivated to restrict the pool of placebos even further (e.g. by also excluding placebos with better SC units than the treated). In theory this would result in a valid test, but in practice it leaves no placebos to compare with the estimate with.

Fortunately there are viable alternatives. The two types of tests that "control" for the differences in distribution by exploiting the pre-intervention characteristics—either by deriving the ratio between the fits and estimates, or using the detailed relationship implied by the parametric assumptions—show promising results in the simulation study. Specifically, both exhibit a size close to the expected level and reasonably high power using real data. The main difference is that the newly introduced test exhibits considerably higher power when the parametric assumption it builds on holds true. These tests are however less transparent than a direct comparison of the estimates and therefore are less apt for a graphical analysis.

Evidently, hypothesis testing with the SC method is not straightforward, but nevertheless still possible. Given the results in this study a constructive approach appears to be to use range of test procedures which together would provide a comprehensive picture of the statistical significance of the estimates. In particular, the graphical analysis using raw estimates provides an intuitive overview of the results, while subjected to a higher risk of being misleading. The tests that directly tries to account for eventual differences are on the other hand more robust to estimates that are not IID, thereby valid in a wider range of settings. The complementing features of these two types would thus, in many settings, make a convincing case if combined.

References

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller (2010) "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program," *Journal of the American Statistical Association*, Vol. 105, No. 490, pp. 493–505.

Abadie, Alberto, Alexis J. Diamond, and Jens Hainmueller (2012) "Comparative Politics and the Synthetic Control Method," Research Paper 2011-25, MIT Political Science Department.

Abadie, Alberto and Javier Gardeazabal (2003) "The Economic Costs of Conflict: A Case Study of the Basque Country," *The American Economic Review*, Vol. 93, No. 1, pp. 113–132.

Billmeier, Andreas and Tommaso Nannicini (2013) "Assessing Economic Liberalization Episodes: A Synthetic Control Approach," *The Review of Economics and Statistics*, Vol. X, p. Forthcoming.

Card, David (1990) "The Impact of the Mariel Boatlift on the Miami Labor Market," *Industrial and Labor Relations Review*, Vol. 43, No. 2, pp. 245–257.

Card, David and Alan B. Krueger (1994) "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania," *The American Economic Review*, Vol. 84, No. 4, pp. 772–793.

Cavallo, Eduardo, Sebastian Galiani, Ilan Noy, and Juan Pantano (2013) "Catastrophic Natural Disasters and Economic Growth," *The Review of Economics and Statistics*, Vol. X, p. Forthcoming.

Coffman, Makena and Ilan Noy (2012) "Hurricane Iniki: measuring the long-term economic impact of a natural disaster using synthetic control," *Environment and Development Economics*, Vol. 17, pp. 187–205.

Hinrichs, Peter (2012) "The Effects of Affirmative Action Bans on College Enrollment, Educational Attainment, and the Demographic Composition of Universities," *The Review of Economics and Statistics*, Vol. 94, pp. 712–722.

Montalvo, José G. (2011) "Voting after the Bombings: A Natural Experiment on the Effect of Terrorist Attacks on Democratic Elections," *The Review of Economics and Statistics*, Vol. 93, pp. 1146–1154.

A Details of the tests in the simulation study

For each constructed data set in the simulation study we first estimate the treatment effect for all units as outlined in ADH. Specifically, following the notation in ADH, we set $\mathbf{X}_i$ to include all three covariates and the outcome in the 7th, the 14th and the 20th period. We then derive the weights by minimizing $\sqrt{(\mathbf{X}_i - \mathbf{X}_0\mathbf{W}_i)'\mathbf{V}(\mathbf{X}_i - \mathbf{X}_0\mathbf{W}_i)}$, where $\mathbf{V}$ is chosen so as to minimize the mean squared pre-intervention fits (in all pre-intervention periods) that is implied by the $\mathbf{W}_i$ yielded by $\mathbf{V}$. The resulting estimates are then used in the three tests as detailed below.

In the first test, ADH 1, we calculate the intervention period average of the estimated effects, $\hat{\beta}_i^A = \frac{1}{T-T_0}\sum_{t=T_0+1}^{T}\hat{\beta}_{it}$. Which are used to derive the rank of the treated unit: $R_1 = \sum_{i=1}^{J+1}\mathbb{1}\left(|\hat{\beta}_1^A| \leq |\hat{\beta}_i^A|\right)$. The null is rejected whenever $R_1/(J+1) \leq 0.05$.

For the second test, ADH 2, we restrict the pool of placebos as detailed in Section 4. That is, we only include placebos that have a mean squared pre-intervention fit that is at most two times the mean squared pre-intervention fit of the treated unit (in (7) we set $\zeta = 2$). This results in $J' \leq J$ number of remaining placebos. Among these placebos, we calculate the rank of the treated unit, R_2, as above and reject the null if $R_2/(J'+1) \leq 0.05$.

Turning to the third test, ADH 3, we first calculate Q_i for all units as described by (8) in Section 4. We then rank the statistic of the treated unit among all units, $R_3 = \sum_{i=1}^{J+1}\mathbb{1}\left(Q_1 \leq Q_i\right)$. Again the null is rejected if $R_3/(J+1) \leq 0.05$.

The forth test, AS 1, we first derive the weights without the restriction that the treated unit cannot act as a control, but in all other aspects follow the procedure as above (e.g. $\mathbf{X}_i$ remain unchanged). Using these weights we derive the "estimates" ($\bar{\beta}_{it}$) and the pre-intervention fits ($\mathbf{P}_i$). For the first intervention period, we stack all $\bar{\beta}_{it}$ into $\bar{\boldsymbol{\beta}}_t$ and the weights and pre-intervention fits are stacked into $\mathbf{S}$, as detailed in Section 4. We then derive $\hat{\mathbf{A}}_t = \mathbf{S}^{+}\bar{\boldsymbol{\beta}}_t$. The first $J+1$ elements in $\hat{\mathbf{A}}_t$ will be the sought after $\hat{\alpha}_{it}$. We repeat this procedure for all $t > T_0$, resulting in a unique $\hat{\alpha}_{it}$ for each i in each $t > T_0$. To derive the statistic of the intervention period average effect we calculate the mean of all $\hat{\alpha}_{it}$ for each i: $\hat{\alpha}_i^A = \frac{1}{T-T_0}\sum_{t=T_0+1}^{T}\hat{\alpha}_{it}$. Then, similarly to above we calculate the rank of the treated unit among all units, $R_4 = \sum_{i=1}^{J+1}\mathbb{1}\left(|\hat{\alpha}_1^A| \leq |\hat{\alpha}_i^A|\right)$, and reject the null whenever $R_4/(J+1) \leq 0.05$.

Last, in the fifth test, AS 2, we resample the control group 25 times for each unit ($M = 25$). In each resample we reduce the control group to half of its original size by randomly drawing $\lceil J/2 \rceil$ units. In each resample round we derive the weights as in the previous test, but only allow the drawn units to have positive weights. This results in $25(J+1)$ estimates, $\bar{\beta}_{imt}$, which are all stacked into a vector $\bar{\boldsymbol{\beta}}_t$, and similarly the weights

and pre-intervention fits that are stacked into $\mathbf{S}$. We then derive $\hat{\mathbf{A}}_t = (\mathbf{S}'\mathbf{S} + \mathbf{I})^{-1}\mathbf{S}'\bar{\boldsymbol{\beta}}_t$ for each $t > T_0$. The resulting $\hat{\alpha}_{it}$ are then averaged over the intervention periods for each i. The average of the treated is then ranked among all units, as above, and we reject the null when $R_5/(J+1) \leq 0.05$.