Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Consistent Estimation of Optimal Synthetic Control Weights

Bernd Lucke[1]
Department of Economics
University of Hamburg

January 2022

Abstract

This paper proposes a new method to estimate synthetic control weights. We derive the true predictor weights from a standard factor model for potential outputs and show that these can be consistently estimated by OLS or maximum likelihood methods. We focus on post-treatment data and use pre-treatment data solely as predictors. Optimal synthetic control weights are defined as minimizing the *post*-treatment mean squared synthetic control error. These weights can be found easily by solving a simple quadratic minimization problem. We compare this to the complex standard optimistic bilevel minimization problem and show that the latter may suffer from lack of identification and inconsistencies in the usage of pre-treatment outcomes or other endogenous variables as predictors.

[1] Prof. Dr. Bernd Lucke, von Melle Park 5, 20146 Hamburg. Tel. +49-40-42838-3996. Email: bernd.lucke@uni-hamburg.de

I. Introduction

Synthetic control methods (SCM) have attracted much interest in recent years and have been used in countless empirical applications. The attractiveness of the method is rooted in its potential to estimate causal effects of policy interventions or other well-defined events. In the words of Athey and Imbens (2017), synthetic control methods are „arguably the most important innovation in the policy evaluation literature in the last 15 years."

Most papers rely on methodological foundations due to Abadie et al. (2010, henceforth ADH). Their approach envolves solving an optimistic bilevel minimization problem in order to determine the optimal weights for the construction of synthetic controls. This type of problem is mathematically complex and a number of unwelcome issues (which initially went unnoticed), have been detected and discussed in the more recent literature. While all of these issues are rooted in the mathematical structure of the bilevel minimization problem, they can broadly be classified as numerical, statistical or economic.

Numerically, a number of authors have reported that the commonly used Synth-algorithm proposed by ADH is unstable and may not converge to the global minimum of the objective function, cf. Becker and Klößner (2017), Becker et al. (2018) and Klößner et al. (2018). Malo et al. (2020) developed a competing algorithm which seems to solve most of these problems. Statistically, Ferman and Pinto (2016) noted that the ADH-estimator is asymptotically biased under plausible assumptions and suggest improved estimators. A broader range of alternative models and estimators is considered by e. g. Chernozhukov et al. (2020).

Economically, the ADH-method has been a source of concern because covariates which are believed to have predictive power for the outcomes of interest, have empirically been found to have very little (or even zero) impact on the construction of the synthetic controls (SC). For instance, Kaul et al. (2015) showed that covariates always receive zero weight when they compete with all pre-treatment outcomes - irrespective of the predictive power of the covariates. This is a trivial consequence of the mathematical structure of the bilevel problem. But even if – informationally inefficient - only a subset of the pre-treatment outcomes is used in the bilevel problem, several researchers have noted that the influence of covariates on the constructed synthetic control is usually surprisingly small.

How strongly predictors affect the choice of SC-weights depends on a different set of weights which we call the predictor weights. These predictor weights enter the bilevel problem in the form of a diagonal weighting matrix, commonly denoted V. As no a priori knowledge on V exists, V is usually data-determined. In fact, it is the determination of V which leads to the protracted mathematical structure of an (optimistic) bilevel minimization problem.

In this paper, we first discuss some problems related to the estimation of V in the standard approach. We do so in the common factors model widely used in the literature, e. g. in ADH (2010), Ferman and Pinto (2016) and Chernozhukov et al. (2020). We show that V may not be identified under standard assumptions. We then argue that the currently used methods to determine V suffer from misspecification because V is estimated from an objective function which minimizes the sum of squared pre-treatment synthetic control errors. But the aim of the SC-method is to choose weights such that the best fit between the *post*-treatment synthetic control and the potential outcome for the treated region under counterfactual non-treatment is found.

Under the assumption that treatment is, conditional on predictors, as good as randomly assigned, we derive the optimal predictor weights (analogous to *V*) from the factor model and then show that standard methods yield consistent estimators for these weights. Both the cross section and the time series dimension are required to approach infinity, but they can do so on arbitrary paths (simultaneous limit) and some of our results also hold for fixed *T*. Although the factor loadings of the idiosyncratic shocks cannot be consistently estimated, we show that the optimal SC-weights (i. e. the weights which minimize the post-treatment mean squared synthetic control error) can be found by solving a standard quadratic problem rather than a bilevel problem.

The rest of the paper is organized as follows. In section II we present the standard factor model for potential outcomes and discuss identification and restrictions on covariates. In section III we critically review the widely used ADH approach and argue that it is unlikely to yield reliable estimates of the optimal synthetic control weights. In section IV we propose a new method to to derive consistent estimates of these weights. We show that this can be achieved by standard methods. Section V concludes.

II. The Model

Suppose we observe a balanced panel of $J+1$ units over $T = T_0 + T_1$ periods of time. Unit 1 has been randomly assigned to a policy intervention (the treatment) from period T_0+1 onward. The policy has no impact on units $2,...,J+1$ and no impact on unit 1 prior to period T_0+1. We are interested in the effect the treatment has had on a specific cardinal variable *y*, which we call the outcome. The observed outcome for region *i* in period *t* is denoted y_{it} and is either the outcome under treatment or under non-treatment, whatever applies.

We assume that for each unit *i* and period $t > T_0$ the *potential* outcome in the case of non-treatment can be expressed as a linear function of *R*-dimensional, random vector z_i of observables and of an *F*-dimensional random vector λ_t of unobservable shocks. The observable, unit-specific variables z_i have deterministic, but time-variant coefficients $\theta_t \in \mathbb{R}^R$, $t = T_0+1,...,T$, and the unobservable time-variant shocks have deterministic, but unit-specific loading coefficients $\mu_i \in \mathbb{R}^F$, $i = 1,...,J+1$:

$$y_{it}^N = \mathbf{z}_i'\theta_t + \lambda_t'\mu_i \qquad (1)$$

Here the superscript *N* indicates that y_{it}^N is the potential outcome in the case of non-intervention. We assume that there are *C* common factors and $J+1$ idiosyncratic shocks, i. e. $F = C + J + 1$. We sometimes use the partition $\lambda_t = \left(\lambda_t^C' \quad \lambda_t^I'\right)'$ where λ_t^C contains the common factors and λ_t^I contains the idiosyncratic shocks. We partition $\mu_i = \left(\mu_i^C' \quad \mu_i^I'\right)'$ accordingly. A shock is idiosyncratic for unit *i* iff μ_i is nonzero only in row $C+i$ and all other μ_j 's are zero in this row.

The observable variables z_i are called the predictors. Note that the predictors may contain a constant term $z_{1i} = 1 \ \forall i = 1,..., J+1$, and that we may, therefore, assume $E(\lambda_t) = 0 \ \forall t$ without loss of generality. Further, z_i is assumed to involve only variables unaffected by the policy intervention, i. e. either variables which are strictly exogenous or variables which have been determined prior to treatment. For instance, in a macroeconomic application where outcomes are measures of GDP, z_i may involve endogenous variables like investment, human capital, infrastructure etc., provided these variables were determined not later than period T_0.

Note that the right-hand side of (1) involves three unobservables, θ_t, λ_t and μ_i. Many observationally equivalent choices for these unobservables exist. To see this, let G_1 be any nonzero $R \times F$ matrix and let G_2 be any nonsingular $F \times F$ matrix. Then for any given θ_t, λ_t and μ_i it is easy to find an observationally equivalent representation with, in general, different unobservables $\tilde{\theta}_t$, $\tilde{\lambda}_t$ and $\tilde{\mu}_i$:

$$
\begin{aligned}
y_{it}^N &= z_i'\theta_t + \lambda_t'\mu_i = z_i'\left(\theta_t - \underbrace{G_1}_{R\times F}\lambda_t + G_1\lambda_t\right) + \mu_i'\lambda_t \\
&= z_i'\left(\theta_t - G_1\lambda_t\right) + z_i'G_1\lambda_t + \mu_i'\lambda_t \\
&=: z_i'\tilde{\theta}_t + \left(z_i'G_1 + \mu_i'\right)\lambda_t \\
&=: z_i'\tilde{\theta}_t + \tilde{\tilde{\mu}}_i'\lambda_t \\
&= z_i'\tilde{\theta}_t + \left(\tilde{\tilde{\mu}}_i'\underbrace{G_2}_{F\times F}\right)\left(G_2^{-1}\lambda_t\right) \\
&=: z_i'\tilde{\theta}_t + \tilde{\mu}_i'\tilde{\lambda}_t
\end{aligned}
$$

Here, $\tilde{\mu}_i := G_2'\left(\mu_i + G_1'z_i\right)$ depends on the predictors. Hence, $\tilde{\mu}_i'\tilde{\lambda}_t$ will, in general, correlate with z_i. In order to uniquely identify θ_t and $\eta_{it} := \mu_i'\lambda_t$ in (1) we therefore impose the identifying assumption

A0: Orthogonality condition
For all $t > T_0$ and all $i = 1,..., J$ we have $E(z_i\eta_{it}) = 0_R$.

Note that for the purpose of this paper we are not interested in identifying a „true“ shock λ_t or a „true“ unit-specific shock η_{it}. This would require a structural analysis and specific, probably controversial identifying assumptions. But since we do not aim at an economic interpretation of the shocks, any identification will do. We just need to ensure that the shocks we work with are indeed uniquely identified.

The most convenient identification for θ_t is achieved by assumption A0, provided $E(z_iz_i')$ is non-singular $\forall i$. This can be seen by premultiplying (1) by z_i

$$z_iy_{it}^N = z_iz_i'\theta_t + z_i\eta_{it},$$

taking expectations and solving for θ_t:

$$\theta_t = E\left(z_i z_i'\right)^{-1} E\left(y_{it}^N z_i\right)$$

This implies that θ_t can be consistently estimated by least squares – a fact we will use below.

Note that predictors may contain endogenous variables determined in period T_0 or earlier. Hence, an identifying assumption analogous to A0 is not possible for periods $1,...,T_0$ since endogenous predictors would typically depend on some of the unobserved shocks.

This is important because in many models outcomes of period *t* can be written as functions of variables which were determined in the current or a previous period. However, (1) does not provide for any right-hand side observable variable dated T_0+1 or later. It is useful to think of (1) as representing a dynamic model which has been solved backwards in time until all observable variables werde determined in period T_0 or earlier. Therefore, θ_t must be time-dependent since it depends on the time difference $t-T_0$, while z_i describes initial conditions prior to treatment.

The initial conditions joint with unobservable shocks in subsequent periods eventually give rise to y_{it}^N. Needless to say, the effects of such unobservable shocks accumulate in a complicated way over time. Their cumulative effect is expressed in the λ_t-vector, which is, therefore, very likely autocorrelated and whose variance probably increases over time.

We have defined (1) only for $t > T_0$. Extending (1) to hold for $1 \le t \le T_0$ would require that all predictors are known already in the initial period 1. This would greatly limit the predictive power of z_i for the treatment period if the time span T_0 prior to treatment is substantial.

On the other hand, if (1) is defined only for $t > T_0$, predictors may involve functions of pre-treatment outcomes y_{is}, $s \le T_0$. In this case the number of pre-treatment outcomes used for the construction of synthetic controls is held fixed and is not increased when asymptotic arguments are invoked.

The potential outcomes in the case of treatment are denoted y_{it}^{Tr} and modeled as the potential outcome in the case of non-treatment plus a treatment effect α_{it} which is unit and time specific:

$$y_{it}^{Tr} = \alpha_{it} + y_{it}^N \qquad (2)$$

Since y_{it}^N is, by definition, independent of treatment, it follows directly that α_{it} is uncorrelated with the random variables z_i and λ_t.

Denoting the treatment status of unit *i* in period *t* by d_{it}, where $d_{it}=1$ in the case of treatment and $d_{it}=0$ otherwise, observed outcomes are

$$y_{it} = d_{it} y_{it}^{Tr} + (1 - d_{it}) y_{it}^{N} \tag{3}$$

The primary object of causal analysis is knowledge of α_{1t} for $t > T_0$ or of its average over the treatment period $\bar{\alpha}_1 := T_1^{-1} \sum_{t=T_0+1}^{T} \alpha_{1t}$. More generally, knowledge of any treatment effect α_{it} may be desired. Since either y_{it}^{Tr} or y_{it}^{N} is unobserved for unit *i* in period *t*, the key question is how observations on y_{it}, d_{it} and z_i can be used to estimate the unobserved components of (2) as well as possible.

For the following, let us introduce the following notation: Collect all covariates of the control regions in the $R \times J$ matrix $Z_0 := (z_2 \quad ... \quad z_{J+1})$ and collect all factor loadings of the control regions in the $F \times J$ matrix $M_0 := (\mu_2 \quad ... \quad \mu_{J+1})$. Denote by $\Theta^{post} := (\theta_{T_0+1} \quad \cdots \quad \theta_T)'$ the $T_1 \times R$ matrix of time-dependent coefficients for the covariates and by $\Lambda^{post} := (\lambda_{T_0+1} \quad \cdots \quad \lambda_T)'$ the $T_1 \times F$ random matrix of shocks. Moreover, let $y_i^{post} := (y_{iT_0+1} \quad ... \quad y_{iT})' \ \forall i = 1, ..., J+1$ and collect all observations for the controls in the post-treatment period in the $T_1 \times J$ matrix $Y_0^{post} := (y_2^{post} \quad ... \quad y_{J+1}^{post})$. Finally, define y_i^{pre} and Y_0^{pre} analogously for the pre-treatment periods $1, ..., T_0$.

Using (3) we can then rewrite (1) with observable variables on the left hand side. For the *J* control units, we have

$$Y_0^{post} = \Theta^{post} Z_0 + \Lambda^{post} M_0 \tag{4}$$

while for the treated unit 1 (1), (2) and (3) yield

$$y_1^{post} = \alpha_1 + \Theta^{post} z_1 + \Lambda^{post} \mu_1 \tag{5}$$

where $\alpha_1 := (\alpha_{1T_0+1} \quad \cdots \quad \alpha_{1T})'$ is the parameter of interest.

III. The Standard SC-Approach

The standard synthetic control approach has been popularized by Abadie and Gardeazabal (2003) and Abadie, Diamond and Hainmueller (ADH) (2010). The key idea is that $y_{1t}^N,\ t > T_0$, the potential output of unit 1 in the counterfactual case of non-treatment, can be approximated by a weighted average of the observed contemporaneous outcomes of the control units. Formally, if $y_1^{N,post} := \left(y_{1T_0+1}^N \quad \cdots \quad y_{1T}^N \right)'$, the ADH approach aims at finding a suitable nonnegative vector of weights $w^* \in \Delta_J := \left\{ w \in \mathbb{R}^J \,\middle|\, \iota_J' w = 1 \ \wedge \ w_i \geq 0 \ \ \forall i = 1,...,J \right\}$ such that

$$y_1^{N,post} \approx Y_0^{post} w^* \tag{6}$$

Here, ι_J is a vector $J \times 1$ vector of ones.

To find the desired weights w^*, ADH's approach relies on the predictors z_1, Z_0 and on all pre-treatment outcomes y_1^{pre}, Y_0^{pre}, where the predictors may also include functions of some or all of the pre-treatment outcomes. Since not all predictors may be equally informative for potential outputs, let $v \in \Delta_R$ be a vector of non-negative predictor weights and let $V := \text{diag}(v)$ be the corresponding diagonal $R \times R$ matrix.

ADH propose to solve the following optimistic bilevel minimization problem:

$$\min_{v \in \Delta_R, w \in \Delta_J} L_{out}(v,w) := \frac{1}{T_0} \left(y_1^{pre} - Y_0^{pre} w \right)' \left(y_1^{pre} - Y_0^{pre} w \right) \tag{7}$$

s. t.

$$w \in \Psi(v) := \underset{w \in \Delta_J}{\text{argmin}} \ L_{in}(v,w) := (z_1 - Z_0 w)' V (z_1 - Z_0 w)$$

$$V = \text{diag}(v)$$

This formulation is due to Malo et al. (2020). For a given V we call

$$\min_{w \in \Delta_J} L_{in}(v,w) = (z_1 - Z_0 w)' V (z_1 - Z_0 w) \tag{8}$$

the inner minimization problem and

$$\min_{v \in \Delta_R, w \in \Psi(v)} L_{out}(v,w) = \frac{1}{T_0} \left(y_1^{pre} - Y_0^{pre} w \right)' \left(y_1^{pre} - Y_0^{pre} w \right) \tag{9}$$

the outer maximization problem. Note that the outer problem requires w to be from $\Psi(v)$.

Define $\Phi_1^Z := \{w \in \Delta_J | z_1 = Z_0 w\}$, $\Phi_1^{Y^{pre}} := \{w \in \Delta_J | y_1^{pre} = Y_0^{pre} w\}$ and, for later purposes, $\Phi_1^M := \{w \in \Delta_J | \mu_1 = M_0 w\}$ [2]. ADH assume the existence of weights $w^0 \in (\Phi_1^Z \cap \Phi_1^{Y^{pre}})$, i. e. there exist weights w^0 such that linear combinations of the columns of Z_0 can exactly reproduce z_1 and the same linear combinations of columns of Y_0^{pre} can exactly reproduce y_1^{pre} . ADH show that under this (and some additional) assumptions $Y_0^{post} w^0$ is an asymptotically unbiased estimator of $y_1^{N,post}$ when the number of pre-treatment observations T_0 approaches infinity.[3]

Unfortunately, the ADH approach is problematic in multiple regards:

1. **If λ_t contains at least one idiosyncratic shock which affects the treated unit, then, asymptotically, $\Phi_1^{Y^{pre}} = \varnothing$ with probability 1.** This was noticed by Ferman and Pinto (2016). Since this idiosyncratic shock would be independent of the shocks hitting the control units, no linear combination of control units can exactly reproduce the pre-treatment time series of the treated unit for $T_0 >> J$. Hence, in this case, there is no reason to believe that the ADH-estimator is asymptocially unbiased.

2. **The weights solving (7) are, in general, not unique.** Define $x_1 := (y_1^{pre\,\prime} \quad z_1{}')'$, $X_0 := (Y_0^{pre\,\prime} \quad Z_0{}')'$. Then the solution set $\Phi_1^X := \Phi_1^Z \cap \Phi_1^{Y^{pre}}$ can be written as $\Phi_1^X = \{w \in \Delta_J | w = w^0 + (I - X^+ X)\zeta,\ \zeta \in \mathbb{R}^J\}$, where X^+ is the Moore-Penrose inverse of X. It is easy to show that Φ_1^X is a convex subset of Δ_J . There are, in general, many admissible choices of ζ , in which case Φ_1^X has infinitely many elements. Only in very special cases would the solution w^0 be unique.

3. ***V* is not identified at the optimum:** Φ_1^Z is non-empty by assumption, hence we have $z_1 = Z_0 w$ and $L_{in}(v, w) = 0$ for *any* matrix $V = \mathrm{diag}(v)$ and any $w \in \Phi_1^X$. Since, as acknowledged by ADH, in practice one may not achieve more than approximate equality $z_1 \approx Z_0 w$, *V* will be nearly unidentified close to the optimum and, therefore, estimates of *V* may be numerically unstable. This is particularly troubling since Φ_1^X is, in general, a non-degenerate convex set, so that, loosely speaking, $w \in \Phi_1^X$ is not identified either. Therefore, estimates of both *V* and *w* may be numerically unstable and the instability of the former may reinforce the instability of the latter and vice versa.

4. **Suboptimal weights *w* are chosen if all pre-treatment outcomes are included in the matrix of predictors:** This result is due to Kaul et al. (2015). Suppose that $z_1 := (y_1^{pre\,\prime} \quad \tilde{z}_1{}')'$

[2] This definition is meaningful only if (1) holds also for the preintervention periods – which ADH assume. As we have pointed out, this requires that predictors are determined in period 1 or earlier. We do not know how ADH justify their usage of preintervention outcomes as predictors.

[3] Some papers (e. g. Malo et al. (2020)) state that ADH prove the „consistency" of the SC-estimator. This is not true. Their proof shows asymptotic unbiasedness. ADH do not claim that the variance of $y_1^{N,post} - Y_0^{post} w^*$ converges to zero.

, $Z_0 := \left(Y_0^{pre}{}' \quad \tilde{Z}_0{}'\right)'$ where $\tilde{z}_1, \tilde{Z}_0$ collect all predictors different from y_1^{pre}, Y_0^{pre}, respectively. Then, a solution to the bilevel problem (7) is given by any $w \in \Phi_1^{Y^{pre}}$ along with $v = T_0^{-1}\left(\iota_{T_0}{}' \quad 0_R{}'\right)'$, i. e. with a *V*-matrix which has equal nonzero elements for the T_0 pre-treatment outcomes on the main diagonal and is zero everywhere else. In words: The choice of weights is solely driven by the pre-treatment outcomes and all other predictors have no impact at all.

This is suboptimal for causal inference as can be seen by the following example. Suppose there are two control units, i. e. $J = 2$, and two periods, T_0 and *T*. Treatment occurs in *T*. Moreover, assume that potential outcomes in the non-treatment case are generated by

$$y_{it}^N = \tilde{z}_i + \varepsilon_{it} \quad \forall i \forall t \tag{10}$$

where ε_{it} is idiosyncratic (i.i.d.) white noise with variance σ_ε^2. Let $\tilde{z}_1 = 2, \tilde{z}_2 = 0, \tilde{z}_3 = 3$ and assume that pre-treatment shocks are $\varepsilon_{1T_0} = -2, \varepsilon_{2T_0} = -1, \varepsilon_{3T_0} = 3$. Hence, the pre-treatment potential outcomes are $y_{1T_0}^N = y_1^{pre} = 0, \; y_{2T_0}^N = y_2^{pre} = -1, \; y_{3T_0}^N = y_3^{pre} = 6$.

Minimizing $\left(y_1^{pre} - Y_0^{pre} w\right)'\left(y_1^{pre} - Y_0^{pre} w\right)$ s. t. $w \in \Delta_2$ is equivalent to solving $y_1^{pre} = w_1 y_2^{pre} + \left(1 - w_1\right) y_3^{pre}$. Hence $w_1 = \left(y_1^{pre} - y_3^{pre}\right) \Big/ \left(y_2^{pre} - y_3^{pre}\right) = 6/7$, i. e. the solution is $w = \left(w_1 \quad w_2\right)' = \left(6/7 \quad 1/7\right)' \in \Delta_2$. It is easy to see that this along with $v = \left(1 \quad 0\right)'$ solves the bilevel problem (7).

However, the weights minimizing the mean squared error in the treatment period would be the solution of

$$\min_{0 \le \tilde{w}_1 \le 1} E\left(\left(y_{1T}^N - \tilde{w}_1 y_{2T}^N - \left(1 - \tilde{w}_1\right) y_{3T}^N\right)'\left(y_{1T}^N - \tilde{w}_1 y_{2T}^N - \left(1 - \tilde{w}_1\right) y_{3T}^N\right)\right) \tag{11}$$

Since

$$y_{1T}^N - \tilde{w}_1 y_{2T}^N - \left(1 - \tilde{w}_1\right) y_{3T}^N = \tilde{z}_1 + \varepsilon_{1T} - \tilde{w}_1\left(\tilde{z}_2 + \varepsilon_{2T}\right) - \left(1 - \tilde{w}_1\right)\left(\tilde{z}_3 + \varepsilon_{3T}\right)$$

the minimand in (11) is

$$\begin{aligned}
&E\left(\left(\tilde{z}_1 + \varepsilon_{1T} - \tilde{w}_1\left(\tilde{z}_2 + \varepsilon_{2T}\right) - \left(1 - \tilde{w}_1\right)\left(\tilde{z}_3 + \varepsilon_{3T}\right)\right)'\left(\tilde{z}_1 + \varepsilon_{1T} - \tilde{w}_1\left(\tilde{z}_2 + \varepsilon_{2T}\right) - \left(1 - \tilde{w}_1\right)\left(\tilde{z}_3 + \varepsilon_{3T}\right)\right)\right) \\
&= \tilde{z}_1{}'\tilde{z}_1 + \sigma_\varepsilon^2 - 2\tilde{w}_1 \tilde{z}_1{}'\tilde{z}_2 - 2\left(1 - \tilde{w}_1\right)\tilde{z}_1{}'\tilde{z}_3 + 2\tilde{w}_1\left(1 - \tilde{w}_1\right)\tilde{z}_2{}'\tilde{z}_3 + \tilde{w}_1^2\left(\tilde{z}_2{}'\tilde{z}_2 + \sigma_\varepsilon^2\right) + \left(1 - \tilde{w}_1\right)^2\left(\tilde{z}_3{}'\tilde{z}_3 + \sigma_\varepsilon^2\right) \\
&= \tilde{z}_1{}'\tilde{z}_1 + 2\sigma_\varepsilon^2 - 2\tilde{z}_1{}'\tilde{z}_3 + \tilde{z}_3{}'\tilde{z}_3 + 2\tilde{w}_1\left(\tilde{z}_1{}'\tilde{z}_3 - \tilde{z}_1{}'\tilde{z}_2 + \tilde{z}_2{}'\tilde{z}_3 - \tilde{z}_3{}'\tilde{z}_3 - \sigma_\varepsilon^2\right) + \tilde{w}_1^2\left(\tilde{z}_2{}'\tilde{z}_2 - 2\tilde{z}_2{}'\tilde{z}_3 + \tilde{z}_3{}'\tilde{z}_3 + 2\sigma_\varepsilon^2\right)
\end{aligned}$$

and, therefore, the solution to (11) is

$$\tilde{w}_1 = -\frac{\tilde{z}_1{}'\tilde{z}_3 - \tilde{z}_1{}'\tilde{z}_2 + \tilde{z}_2{}'\tilde{z}_3 - \tilde{z}_3{}'\tilde{z}_3 - \sigma_\varepsilon^2}{\tilde{z}_2{}'\tilde{z}_2 - 2\tilde{z}_2{}'\tilde{z}_3 + \tilde{z}_3{}'\tilde{z}_3 + 2\sigma_\varepsilon^2}$$

provided the denominator is nonzero and $0 \le \tilde{w}_1 \le 1$. In our example this is easily verified to be the case:

$$\tilde{w}_1 = -\frac{\tilde{z}_1'\tilde{z}_3 - \tilde{z}_1'\tilde{z}_2 + \tilde{z}_2'\tilde{z}_3 - \tilde{z}_3'\tilde{z}_3 - \sigma_\varepsilon^2}{\tilde{z}_2'\tilde{z}_2 - 2\tilde{z}_2'\tilde{z}_3 + \tilde{z}_3'\tilde{z}_3 + 2\sigma_\varepsilon^2} = -\frac{6-9-\sigma_\varepsilon^2}{9+2\sigma_\varepsilon^2} = \frac{3+\sigma_\varepsilon^2}{9+2\sigma_\varepsilon^2} < 1$$

If, e. g., $\sigma_\varepsilon^2 = 3$, then the MSE-optimal solution is $\tilde{w} = (2/5 \quad 3/5)' \neq (6/7 \quad 1/7)'$. Note that the implied bias of the ADH-solution $w = (6/7 \quad 1/7)'$ is

$$E\left(Y_0^{post} w - y_1^{post}\right) = \frac{6}{7}\tilde{z}_2 + \frac{1}{7}\tilde{z}_3 - \tilde{z}_1 = -\frac{11}{7},$$

while the bias of the MSE-optimal solution $\tilde{w} = (2/5 \quad 3/5)'$ is much smaller in absolute value:

$$E\left(Y_0^{post} \tilde{w} - y_1^{post}\right) = \frac{2}{5}\tilde{z}_2 + \frac{3}{5}\tilde{z}_3 - \tilde{z}_1 = -\frac{1}{5}$$

In this example, the main point to see is that the MSE-optimal weights depend on $\tilde{z}_1$ and $\tilde{Z}_0$. But if all pre-treatment outcomes are used as predictors, then the ADH-approach determines weights which are completely independent of $\tilde{z}_1$ and $\tilde{Z}_0$: They depend solely on the pre-treatment outcomes. The reason is that the pre-treatment outcomes are – trivially - a sufficient statistic for the pre-treatment outcomes in the outer minimization problem (9). But they are not a sufficient statistic for the *post*-treatment outcomes we are interested in. For these, the predictors $\tilde{z}_1$, $\tilde{Z}_0$ contain important information which helps to reduce the confounding effects of idiosyncratic shocks in the pre-treatment period – and which should not be discarded.

In other words: The bilevel minimization problem (7) determines the „wrong“ *V*-matrix.

5. **No endogenous variables allowed as predictors:** One might suspect that the problem of insufficient predictor weights for $\tilde{z}_1, \tilde{Z}_0$ is not confined to the case where *all* pre-treatment outcomes are used as predictors. If T_0 is large and all but a few pre-treatment outcomes are in the predictor matrix it is likely, that the *V*-matrix will still assign most of the weight to the pre-treatment outcomes in the predictor matrix and only very little to all other predictors. The distortion is probably the less severe, the fewer pre-treatment outcomes (relative to all pre-treatment outcomes in the outer minimization problem) are used as predictors. But there is little merit in studying this in more detail, because under plausible settings there is a more fundamental problem which disallows the use of endogenous variables as predictors in ADH's approach altogether.

   To see this, suppose that potential outputs are autocorrelated. This is quite a natural assumption since pre-treatment outcomes would have no value as predictors if potential outcomes were uncorrelated over time. With autocorrelation, however, pre-treatment outcomes (and other variables dated T_0 and earlier) can serve as initial conditions in a dynamically evolving system. Since errors add up over time, estimation of treatment effects

should condition on initial conditions in T_0 or at least not much earlier. But since, according to (1), the same predictors must be used for all outcomes irrespective of their timing, the dynamic equations for post-treatment outcomes must be solved backward in time, while the same equation for pre-treatment outcomes prior to T_0 must be solved forward.

This in itself is not a problem because the resulting equations are merely descriptive – they do not need to admit a causal interpretation. However, solving forward necessarily results in a representation where pre-treatment outcomes are functions of future shocks. While this can still be seen as merely descriptive, such a representation is not in line with the assumed model in equation (1). Equation (1) holds uniformly for all periods and requires that period-*t* outcomes depend on shocks which are known in period *t*. Hence, z_i can only accomodate predictors which allow a representation as in (1).

To illustrate the issue, consider the following example: Suppose there are four periods, T_0-1, T_0, T_0+1 and T_0+2. Treatment occurs in T_0+1. Moreover, assume that potential outcomes in the non-treatment case are generated by

$$y_{it}^N = \rho_y y_{it-1}^N + \varepsilon_{it} \quad \forall i \forall t, \; 0 < |\rho| < 1 \tag{12}$$

where ε_{it} is idiosyncratic (i.i.d.) white noise with variance σ_ε^2.

For the last period T_0+2, equation (12) can be written as

$$y_{iT_0+2}^N = \rho_y y_{iT_0+1}^N + \varepsilon_{iT_0+2} = \rho_y \left(\rho_y y_{iT_0}^N + \varepsilon_{iT_0+1} \right) + \varepsilon_{iT_0+2} = \rho_y^2 z_i + \lambda_{T_0+2}, \tag{13}$$

where $z_i := y_{iT_0}^N$ and $\lambda_{T_0+2} := \varepsilon_{iT_0+2} + \rho_y \varepsilon_{iT_0+1}$ is a first order moving average. Clearly, (13) is a special case of (1), as is the analogous equation for the preceding treatment period

$$y_{iT_0+1}^N = \rho_y y_{iT_0}^N + \varepsilon_{iT_0+1} =: \rho_y z_i + \lambda_{T_0+1} \tag{14}$$

Here, λ_{T_0+1} is white noise. Moreover, $\theta_{T_0+1} := \rho_y$ and $\theta_{T_0+2} := \rho_y^2$.

For period T_0-1, however, it is impossible to write (12) in a form compatible with (1):

$$y_{iT_0-1}^N = \rho_y^{-1} \left(y_{iT_0}^N - \varepsilon_{iT_0} \right) =: \theta_{T_0-1} z_i - \rho_y^{-1} \varepsilon_{iT_0} \tag{15}$$

Since ε_{iT_0} is white noise, there is no way how this could be written as a shock which is determined in period T_0-1.

Consequently, the model used by ADH, Ferman and Pinto and others does not allow lagged endogenous variables in the predictor matrix (unless they predate period 1 rather than period T_0+1). Asymptotic results which let $T_0 \to \infty$ implicitly assume that only strictly exogenous predictors are allowed. This is not a desirable property in models with autocorrelated outcomes.

IV. Consistent estimates for synthetic controls weights

As argued above, a major weakness of the standard ADH-type SC-estimation strategy is the weighting matrix *V*. It is central to the bilevel minimization problem (7), but it is far from clear whether a data-driven determination of *V* from within the bilevel problem results in a reasonable estimate. And estimate of what? The first question to be addressed is: Is there something like a true, optimal *V*-matrix and, if so, how is *V* related to the parameters of the potential outcomes model (1)-(3)?

The overarching aim of synthetic control analysis is a good estimate of the counterfactual $y_1^{N,post}$. For this, let us focus on the synthetic control error $\varepsilon_1^{post}(w) := y_1^{N,post} - Y_0^{post} w,\ w \in \Delta_J$. A standard optimality criterion would be the least-squares criterion, i. e. we may want to minimize the mean squared error, defined as the conditional expectation

$$MSE_1^{post} := \frac{1}{T_1} E\left(\varepsilon_1^{post}(w)'\varepsilon_1^{post}(w)\middle| I_0\right)$$

where $I_0 := \left\{y_1^{pre}, Y_0^{pre}, Y_0^{post}, z_1, Z_0\right\}$ is the relevant information set.

We have

$$\begin{aligned}
\varepsilon_1^{post}(w)'\varepsilon_1^{post}(w) &= \left(y_1^{N,post} - Y_0^{post} w\right)'\left(y_1^{N,post} - Y_0^{post} w\right) \\
&= \left(\Theta^{post}\left(z_1 - Z_0 w\right) + \Lambda^{post}\left(\mu_1 - M_0 w\right)\right)'\left(\Theta^{post}\left(z_1 - Z_0 w\right) + \Lambda^{post}\left(\mu_1 - M_0 w\right)\right) \\
&= \left(z_1 - Z_0 w\right)'\Theta^{post\,\prime}\Theta^{post}\left(z_1 - Z_0 w\right) + \left(\mu_1 - M_0 w\right)'\Lambda^{post\,\prime}\Lambda^{post}\left(\mu_1 - M_0 w\right) \\
&\quad + 2\left(z_1 - Z_0 w\right)'\Theta^{post\,\prime}\Lambda^{post}\left(\mu_1 - M_0 w\right)
\end{aligned}$$

i. e. the appropriate problem to solve is

$$\begin{aligned}
\min_{w \in \Delta_J}\ MSE_1^{post} &= \left(z_1 - Z_0 w\right)' E\left(T_1^{-1}\Theta^{post\,\prime}\Theta^{post}\middle| I_0\right)\left(z_1 - Z_0 w\right) \\
&\quad + \left(\mu_1 - M_0 w\right)' E\left(T_1^{-1}\Lambda^{post\,\prime}\Lambda^{post}\middle| I_0\right)\left(\mu_1 - M_0 w\right) \\
&\quad + 2\left(z_1 - Z_0 w\right)' E\left(T_1^{-1}\Theta^{post\,\prime}\Lambda^{post}\middle| I_0\right)\left(\mu_1 - M_0 w\right)
\end{aligned} \tag{16}$$

While (16) is quite different from the bilevel problem (7), it is apparent that under ADH's assumption $\Phi_1^Z \cap \Phi_1^{Y^{pre}} \neq \varnothing$, a solution to (7) is also a solution to (16). To see this, note that $y_1^{pre} - Y_0^{pre} w = \Theta^{pre}\left(z_1 - Z_0 w\right) + \Lambda^{pre}\left(\mu_1 - M_0 w\right)$ and, therefore, any $w^0 \in \Phi_1^Z \cap \Phi_1^{Y^{pre}}$ satisfies both $z_1 = Z_0 w^0$ and $\mu_1 = M_0 w^0$ with probability one[4]. Hence, w^0 solves (16). However, *V* would not be identified. (Note that ADH's solution set $\Phi_1^Z \cap \Phi_1^{Y^{pre}}$ can equivalently be written in the more lucid form $\Phi_1^Z \cap \Phi_1^M$ - which is the form we will use in the sequel.)

[4] $\mu_1 = M_0 w^0$ holds only with probability one because, in principle, we might have that Λ^{pre} is orthogonal to $\mu_1 - M_0 w^0$ with $\mu_1 \neq M_0 w^0$. However, this event occurs only with probability zero.

Since w^0 solves (16) under ADH's assumption $\Phi_1^Z \cap \Phi_1^M \neq \varnothing$, it is no surprise that, as ADH prove, $Y_0^{post} w^0$ is a consistent estimator for $y_1^{N,post}$. However, this result hinges crucially on the assumption $\Phi_1^Z \cap \Phi_1^M \neq \varnothing$. In fact, under this assumption and provided we know the loading coefficients μ_1, M_0, we might ignore the bilevel problem (7) completely and simply solve the linear equations

$$\begin{pmatrix} Z_0 \\ M_0 \end{pmatrix} w = \begin{pmatrix} z_1 \\ \mu_1 \end{pmatrix}, \quad w \in \Delta_J \tag{17}$$

A solution to (17) would exist by assumption and it would also solve (7) and (16). It would, however, have no relation whatsoever to the structural information like Θ^{post}, Θ^{pre} or the moments of the shocks $\lambda_1^{post}, \Lambda_0^{post}, \lambda_1^{pre}, \Lambda_0^{pre}$. (This does not invalidate the solution. It merely emphasizes that in the case of $\Phi_1^Z \cap \Phi_1^M \neq \varnothing$ the post-treatment controls Y_0^{post} would be a sufficient statistic for this structural interpretation, since $Y_0^{post} w^0$ would be equal to $y_1^{N,post}$.)

Unfortunately, the ADH assumption $\Phi_1^Z \cap \Phi_1^M \neq \varnothing$ is far from innocuous. There is no theoretical reason why it should hold and it apparently almost never holds in practice. But if a trivial solution $w^0 \in \Phi_1^Z \cap \Phi_1^M$ does not exist, we do not know if the solution to (7) gives rise to a consistent estimator of $y_1^{N,post}$. Moreover, comparing the bilevel problem (7) to the problem associated with minimizing the conditional mean squared error of ε_1^{post}, i. e. problem (16), suggests that the solutions to these problems may be very different for the following reasons:

1. The ADH-approach requires that a quadratic form of $z_1 - Z_0 w$ be minimized with respect to *V*, where *V* is constrained to be a diagonal matrix. The minimand in (16) also contains a quadratic form of $z_1 - Z_0 w$, but with respect to the weighting matrix $E\left(T_1^{-1}\Theta^{post\,\prime}\Theta^{post} \middle| I_0\right)$, which is almost certainly a different and in particular not a diagonal matrix.

2. The ADH-approach requires a quadratic form of $y_1^{pre} - Y_0^{pre} w$ to be minimized. Since $y_1^{pre} - Y_0^{pre} w = \Theta^{pre}\left(z_1 - Z_0 w\right) + \Lambda^{pre}\left(\mu_1 - M_0 w\right)$, this involves minimizing a quadratic form of $\mu_1 - M_0 w$ with respect to a weighting matrix $T_1^{-1}\Lambda^{pre\,\prime}\Lambda^{pre}$, whereas the minimand in (16) defines the analogous quadratic form with respect to $E\left(T_1^{-1}\Lambda^{post\,\prime}\Lambda^{post} \middle| I_0\right)$. These weighting matrices are, in general, different.

3. Similarly, the ADH-approach involves a mixed term $\left(z_1 - Z_0 w\right)'\left(T_0^{-1}\Theta^{pre\,\prime}\Lambda^{pre}\right)\left(\mu_1 - M_0 w\right)$. The analogous term in the minimand of (16) is $\left(z_1 - Z_0 w\right)' E\left(T_1^{-1}\Theta^{post\,\prime}\Lambda^{post} \middle| I_0\right)\left(\mu_1 - M_0 w\right)$. Again, the two weighting matrices are different. For instance, actual shocks Λ^{pre} from the pre-treatment period have an impact on the weights in the ADH-approach, while minimizing the MSE in (16) requires weights which rely on conditional expectations of shocks in the treatment period. Moreover, there is usually no reason to believe that the Θ-coefficients are

the same in pre- and post-period or that Θ^{pre} correlates in the same way with actual pre-treatment shocks as we may Θ^{post} expect to do with shocks of the treatment period[5].

Since it seems reasonable to accept the minimization of the MSE of the synthetic control error ε_1^{post} as optimality criterion, the above discussion suggests that ADH's SC-approach suffers from a generally misspecified choice of weighting matrices. Only if an exact solution $y_1^{pre} = Y_0^{pre} w^0$ exists, would this choice not matter and the ADH-solution would asymptotically coincide with the solution of (16). But, as Ferman and Pinto (2016) have correctly pointed out, the probability for the existence of such a solution is asymptotically zero.

It would, therefore, be desirable to solve (16) directly. This approach has so far been discarded in the discipline because the counterfactual potential outcome $y_1^{N,post}$ is unobserved. But if consistent estimates of Θ^{post}, Λ^{post}, μ_1 and M_0 can be obtained, (16) could be approximated arbitrarily well with increasing *J* or *T* (or both).

In the following we will explore if such estimates are possible. The result will be negative, because it will turn out that there is no way to estimate the idiosyncratic loading coefficients μ_1^I consistently – not even if both *J* and *T* become very large. But, as we will show, Θ^{post}, Λ^{post}, M_0^C and μ_1^C can be estimated consistently and this is sufficient to solve (16).

To show this we proceed in three steps. First, suppose that $J >> R$ and estimate equation (1) for each period $t > T_0$ as a cross section regression over the *J* control units for which the potential outcome in the case of non-treatment is observed:

$$\begin{pmatrix} y_{2t}^N \\ \vdots \\ y_{J+1,t}^N \end{pmatrix} = \underbrace{\begin{pmatrix} z_2' \\ \vdots \\ z_{J+1}' \end{pmatrix}}_{J\times R} \underbrace{\theta_t}_{R\times 1} + \eta_t = Z_0' \theta_t + \eta_t \qquad (18)$$

where $\eta_{jt} := \lambda_t' \mu_j \quad \forall j = 2,...,J+1$ and $\eta_t := \left(\eta_{2t} \cdots \eta_{J+1t}\right)'$. The error terms η_{jt} are linear combinations of the λ_t's, some of which are common factors, and, hence, the covariance matrix of η_t will not be diagonal, i. e. we will have non-zero covariances $E\left(\eta_{kt}\eta_{jt}\right) \neq 0,\ k \neq j$. Moreover, while the identifying assumption A0 ensures $E\left(z_i \eta_{it}\right) = 0$, the regressor matrix is not strictly exogenous since $E\left(z_k \eta_{jt}\right),\ k \neq j$, may well be nonzero. (This would, e. g., typically be the case if the common shocks λ_t are autocorrelated and if z_k involves endogenous variables dated T_0 or earlier.)

[5] Note that it may be a reasonable assumption to postulate that observed and unobserved variation in potential outcomes is orthogonal to each other, i. e. any linear combination of predictors is uncorrelated with any linear combination of shocks, $\left(z_1 - Z_0 w\right)'\left(T_0^{-1}\Theta^{pre\,\prime}\Lambda^{pre}\right)\left(\mu_1 - M_0 w\right)$ in ADH's approach vanishes asymptotically. But it would not imply that the analogous term involving the conditional expectation $E\left(T_1^{-1}\Theta^{post\,\prime}\Lambda^{post}\middle| I_0\right)$ converges to zero, so that even asymptotically ADH's approach differs, in general, from an approach which minimizes the MSE of the synthetic control error ε_1^{post}.

So we cannot generally expect the OLS estimate $\hat{\theta}_t$ to be unbiased in finite samples. But by virtue of assumption A0 the estimate $\hat{\theta}_t$ is *J*-consistent for $\theta_t \ \forall t > T_0$. (Note that under the strong assumption of exogeneity of the z_i 's, $\hat{\theta}_t$ would even be unbiased.) Obviously, the associated estimate $\hat{\eta}_t$ is consistent for η_t.

Now define the $T_1 \times J$ matrix $H := \left(\eta_{T_0+1} \ \cdots \ \eta_T\right)'$[6]. We have $H = \Lambda M_0$, a decomposition which is unique if we impose the conventional restrictions that $M_0 M_0'$ be diagonal and $T_1^{-1}\Lambda'\Lambda = I_F$, i. e. all shocks are orthogonal to each other. We can partition $M_0' = \left(M_0^{C\prime} \ \ M_0^{I\prime}\right)$, where $M_0^{I\prime}$ is $J \times J+1$ and its first column is a column of zeros reflecting the fact that the idiosyncratic shock of the treated unit 1 does not affect any of the control units.

Moreover, we can partition $\Lambda = \left(\Lambda^C \ \ \Lambda^I\right)$ with the common factors $\Lambda^{C\prime} := \left(\lambda_{T_0+1}^C \ \cdots \ \lambda_T^C\right)$ and the idiosyncratic shocks $\Lambda^{I\prime} := \left(\lambda_{T_0+1}^I \ \cdots \ \lambda_T^I\right)$ being of dimensions $C \times T_1$ and $J+1 \times T_1$, respectively. Note that the first column of Λ^I is the idiosyncratic shock of the treated unit 1. We obtain

$$H = \Lambda M_0 = \left(\Lambda^C \ \ \Lambda^I\right)\left(M_0^{C\prime} \ \ M_0^{I\prime}\right)' = \Lambda^C M_0^C + \Lambda^I M_0^I + =: \Lambda^C M_0^C + \Omega \qquad (19)$$

where Ω is a $T_1 \times J$ matrix of the idiosyncratic shocks with typical element ω_{ti}, $t = T_0+1,...,T$, $i = 2,...,J+1$. Let $\omega_t := \left(\omega_{t2} \ \cdots \ \omega_{t\,J+1}\right)'$ be the period-*t* column of Ω' . Clearly, $\omega_t = M_0^{I\prime}\lambda_t^I$ and $E\left(\omega_t\omega_t'\right) = E\left(M_0^{I\prime}\lambda_t^I\lambda_t^{I\prime}M_0^I\right) = M_0^{I\prime}E\left(\lambda_t^I\lambda_t^{I\prime}\right)M_0^I = M_0^{I\prime}M_0^I$ since the J idiosyncratic shocks in each λ_t^I vector are, by definition, orthogonal to each other and have unit variance.

$\tilde{M}_0^I$ eine Note that the covariance matrix $M_0^{I\prime}M_0^I$ is diagonal and, therefore, the least squares criterion requires to choose Λ^C and M_0^C such that

$$tr\left(\Omega'\Omega\right) = tr\left(\left(H - \Lambda^C M_0^C\right)'\left(H - \Lambda^C M_0^C\right)\right)$$

is minimal. This is achieved by the standard principal components estimator, i. e. when $M_0^C = T_1^{-1}\Lambda^{C\prime}H$ and Λ^C equals $\sqrt{T_1}$ times the $T_1 \times F$ matrix of those orthonormal eigenvectors of HH' which correspond to the *C* greatest eigenvalues of HH'.

[6] We often suppress the superscript *post* in the following derivations, since all variables are from the post-treatment period.

The common factors Λ^C and their factor loadings M_0^C can be consistently estimated by the principal components estimator even if T_1 is fixed and only J approaches infinity. Bai (2003) shows that necessary and sufficient conditions for this case are the assumptions

A1 $$\frac{1}{J}\lim_{J\to\infty}\sum_{i=2}^{J+1}\omega_{ti}\omega_{si}=0,\ \forall t\neq s$$

and

A2 $$\frac{1}{J}\lim_{J\to\infty}\sum_{i=2}^{J+1}\omega_{ti}^2=\sigma^2,\ \forall t$$

which we may refer to as asymptotic orthogonality and asymptotic homoskedasticity, respectively.

Whether it is reasonable to assume A1 and A2 depends on the data. A2 is probably an assumption which can often be justified relatively easily, but absence of serial correlation is not necessarily guaranteed by taking the average of a large cross section (A1), since positive autocorrelation is plausible for many e. g. shocks with economic relevance such as technology shocks, monetary policy shocks or fiscal policy shocks. However, if the outcome variables are, say, measures of economic activity for entities which interact with each other in a common market, it is hard to think of such shocks as idiosyncratic shocks.

For any shock which hits an important activity of one entity will also affect the activities of all other entities with which it trades commodities, services or financial assets or with which it shares some degree of factor mobility. Equilibrium requirements and accounting identities ensure that any relevant shock will be transmitted to some degree to many, if not all other entities. Hence, systems in which individual entities coordinate their decisions in open markets are by construction, interdependent. Therefore, all shocks which disturb the interdependent system must be considered common shocks, i. e. common factors.

This does not mean that idiosyncratic shocks are non-existent. But a shock which affects one entity and leaves no trace for any other entity must be a shock which is not internalized in the interdependent system, but rather is extraneous to it. The prime candidate for such shocks are measurement errors. If the outcome variables are recorded with error, such errors may well be independent between cross section units. And even if measurement errors for some units are positively autocorrelated or if there are similar autocorrelation structures for a group of, say, neighboring entities, other entities may well have negative autocorrelation structures or the mere increase in the number of entities without any serial correlation will ensure that asymptotically the phenomenon of serial correlation vanishes.

Moreover, in health sciences or in educational research there may be many settings in which it can be reasonable to assume cross sections shocks without pronounced or systematic serial correlation. We will therefore in the following assume that assumptions A1 and A2 hold.

Suppose now that we estimate (18) by OLS for all $t>T_0$. We can collect the estimated coefficients in the $T_1\times R$ matrix $\hat{\Theta}^{post}:=\hat{\Theta}:=\left(\hat{\theta}_{T_0+1}\ \ ...\ \ \hat{\theta}_T\right)'$ and the residuals in the $T_1\times J$ matrix $\hat{H}:=\left(\hat{\eta}_{T_0+1}\ \ \cdots\ \ \hat{\eta}_T\right)'$**.** The number of common factors *C* is not known, but it can be consistently estimated using the information criterion of Bai and Ng (2002). In this procedure

both J and T_1 are required to go to infinity for the estimate of C to be consistent, but no specific relation between J and T_1 must hold. In particular, J may be much larger than T_1.

Hence, the principal components estimators provides consistent estimates $\hat{\Lambda}^C$ and $\hat{M}_0^C$. Let us now, as the second step, estimate μ_1^C, i. e. unit 1's loading coefficients for the common factors. For all $t > T_0$ we have

$$y_{1t}^{Tr} = \alpha_{1t} + z_1'\theta_t + \lambda_t'\mu_1 = \bar{\alpha}_1 + z_1'\theta_t + \lambda_t^{C\prime}\mu_1^C + \underbrace{\lambda_t^{I\prime}\mu_1^I + \left(\alpha_{1t} - \bar{\alpha}_1\right)}_{=:u_{1t}} \tag{20}$$

For fixed C, suppose that T_1 goes to infinity. Since z_1 is known and consistent estimates of θ_t and λ_t^C have been derived, we can run a regression across time with regression coefficients $\bar{\alpha}_1$ and $\mu_1^{C,post}$

$$y_1^{Tr,post} - \hat{\Theta}^{post} z_1 = \bar{\alpha}_1 \iota_{T_1} + \hat{\Lambda}^{C,post}\mu_1^{C,post} + u_1^{post}, \tag{21}$$

where $y_1^{Tr,post} := \left(y_{1T_0+1} \quad ... \quad y_{1T}\right)'$ and $u_1^{post} := \left(u_{1T_0+1} \quad ... \quad u_{1T}\right)'$ collect the respective post-treatment periods.

If z_1 contains a constant term, then the dependent variable is adjusted for the mean of the potential outcomes in the non-treatment case. Therefore, the estimate of $\bar{\alpha}_1$ is a (first) estimate of the average treatment effect over time for unit 1.

By construction, the error term u_1^{post} does not correlate with the regressors $\hat{\Lambda}^{C,post}$, so a simple OLS-estimate of (21) is T_1-consistent for μ_1^C. However, the idiosyncratic shock and the nonconstant component of the treatment effect $\alpha_{1t} - \bar{\alpha}_1$ may well be autocorrelated and therefore we may encounter serial correlation in u_1^{post}.

But this can be dealt with in the usual way. Suppose $u_{1t} = \rho_1 u_{1t-1} + \varepsilon_{1t}$ and ε_{1t} is i.i.d., then multiplying the period $t-1$ equation of (21) by ρ_1 and subtracting the result from the period t equation yields

$$y_{1t}^{Tr} - z_1'\hat{\theta}_t = \left(1-\rho_1\right)\bar{\alpha}_1 + \rho_1\left(y_{1t-1}^{Tr} - z_1'\hat{\theta}_{t-1}\right) + \lambda_t^{C\prime}\mu_1^C - \lambda_t^{C\prime}\rho_1\mu_1^C + \varepsilon_{1t} \tag{22}$$

(22) can be estimated by OLS, restricted least squares or maximum likelihood (or any other suitable, consistent estimator). OLS would have as drawbacks that no direct estimate of $\bar{\alpha}_1$ is possible and that the direct estimates of μ_1^C and ρ_1 will in general not be compatible with the direct estimate of $\rho_1\mu_1^C$. While this property lends itself naturally for specification tests, it would entail the problem that we would not get a unique estimate of μ_1^C. Moreover, OLS of (22) would have to estimate twice as many parameters as OLS applied to (21). Therefore, an

OLS regression of type (22) may suffer from few degrees of freedom if (in the finite samples of applied work) C/T_1 is not much lower than 0.5.

But even if T_1 is sufficiently great, we need a reliable and unique estimate of μ_1^C to solve (16). Therefore, a consistent estimator from the class of restricted least squares or maximum likelihood methods may be preferable. Denote the resulting estimate by $\hat{\mu}_1^C$.

As the third and last step we show that it is possible to solve (16) without knowledge of μ_1^I and M_0^I. For this define

$$\varepsilon_1^{C,post}(w) := \left(y_1^{N,post} - \Lambda^{I,post}\mu_1^I\right) - \left(Y_0^{post} - \Lambda^{I,post}M_0^I\right)w, \;\; w \in \Delta_J$$

where $\Lambda^{I,post}$ and the $J+1 \times J$ matrix M_0^I are the submatrices of Λ^{post} and M_0, respectively, which correspond to the idiosyncratic shocks. Hence,

$$\varepsilon_1^{post}(w) = \varepsilon_1^{C,post}(w) + \Lambda^{I,post}\left(\mu_1^I - M_0^I w\right)$$

and (16) becomes

$$\begin{aligned}
\min_{w\in\Delta_J} MSE_1^{post} &= \frac{1}{T_1}E\left(\left(\varepsilon_1^{C,post}(w) + \Lambda^{I,post}\left(\mu_1^I - M_0^I w\right)\right)'\left(\varepsilon_1^{C,post}(w) + \Lambda^{I,post}\left(\mu_1^I - M_0^I w\right)\right)\middle| I_0\right) \\
&= \frac{1}{T_1}E\left(\varepsilon_1^{C,post}(w)'\varepsilon_1^{C,post}(w) + \left(\mu_1^I - M_0^I w\right)'\underbrace{\Lambda^{I,post}{}'\Lambda^{I,post}}_{=I_J}\left(\mu_1^I - M_0^I w\right)\middle| I_0\right) \\
&\quad + \frac{2}{T_1}E\left(\varepsilon_1^{C,post}(w)'\Lambda^{I,post}\left(\mu_1^I - M_0^I w\right)\middle| I_0\right) \\
&= \frac{1}{T_1}E\left(\varepsilon_1^{C,post}(w)'\varepsilon_1^{C,post}(w) + \mu_1^I{}'\mu_1^I + w'\Omega'\Omega w\middle| I_0\right) \\
&\quad + \frac{2}{T_1}E\left(\varepsilon_1^{C,post}(w)'\Lambda^{I,post}\left(\mu_1^I - M_0^I w\right)\middle| I_0\right)
\end{aligned}$$

since $\mu_1^I{}'M_0^I = 0$ and $M_0^I{}'M_0^I = T_1^{-1}E\left(\Omega'\Omega\middle| I_0\right)$. Note that $\mu_1^I{}'\mu_1^I$ is independent of *w* and can therefore be neglected.

Note further that $\varepsilon_1^{C,post}(w)$ is stochastically independent of the idiosyncratic shocks and hence

$$E\left(\varepsilon_1^{C,post}(w)'\Lambda^{I,post}\left(\mu_1^I - M_0^I w\right)\middle| I_0\right) = E\left(\varepsilon_1^{C,post}(w)\middle| I_0\right)'E\left(\omega_1 - \Omega w\middle| I_0\right) = 0$$

where $\omega_1 := \Lambda^{I,post}\mu_1^I$.

Setting $\hat{\varepsilon}_1^{C,post}(w) := \hat{\Theta}^{post}\left(z_1 - Z_0 w\right) + \hat{\Lambda}^{C,post}\left(\hat{\mu}_1^C - \hat{M}_0^C w\right)$, it follows that solving

$$\min_{w \in \Delta_J} \hat{\varepsilon}_1^{C,post}(w)' \hat{\varepsilon}_1^{C,post}(w) + w' \hat{\Omega}' \hat{\Omega} w \qquad (23)$$

is asymptotically equivalent to solving (16). Note that J and T_1 may approach infinity along some arbitrary path and that $\hat{\Omega}$ is the matrix of residuals from (19).

If w^* denotes the solution to (23), the synthetic control for unit 1 is given by $\hat{y}_1^{N,post} := Y_0^{post} w^*$ and the estimated causal effects of treatment are $\hat{\alpha}_1 := y_1^{Tr,post} - \hat{y}_1^{N,post}$.

V. Conclusions

As innovative as it was at the time of inception, the standard ADH (2010) approach to the construction of synthetic controls is now known to suffer from a number of numerical, statistical and economic issues. We have discussed some known and some less known problems which arise when ADH's method is applied. Against this background we conclude that researchers should try to improve on ADH's method.

One important problem in ADH's approach is the optimality concept. ADH search for synthetic control weights which yield a good fit for the treated unit's pre-treatment outcomes. But what researchers are interested in is a good fit between the synthetic control and the counterfactual potential outcome of the treated unit in the case of non-treatment. While this potential outcome is unobserved, we argue that it can be estimated from post-treatment data of the control group.

This estimate hinges on the assumption that the potential outcomes of all units are generated by a (fairly general) factor model. The treated unit shares the common factors and the coefficients of the covariates with the units in the control group. These unobservables can be estimated from the cross section variation of the control group post treatment. Moreover, the treated units' loading coefficients for the common factors can be estimated from its post-treatment observations (i. e. under treatment). All these estimates are consistent, so that the synthetic control weights can be derived which are optimal in the sense of minimizing the post treatment mean squared synthetic control error.

But clearly, this approach has some limitations. First, there is the appropriateness of the factor model. Simple specification tests and goodness-of-fit-measures for the observable potential outcomes may be used to check its validity. Second, since the treated unit shares the same factor model structure with the control units, the model implies that treatment has (conditional on covariates) been as good as randomly applied. Researchers will need to check thoroughly whether this is, in fact, the case for their treated unit or can at least approximately be true. Third, the method requires a large cross section and a large time dimension. It is unknown so far how large these dimensions should be to justify the the usage of this method.

But little such knowledge exists for other consistent SC-methods, e. g. ADH. A useful extension of this paper will be a Monte Carlo study which assesses the reliability of our method and of competing methods on plausibly calibrated artificial data sets.

And finally, the test of the pudding is in the eating. We have not yet applied this method to real world data sets, although a first such study is to follow soon. For any such attempt, it seems advisable to use and compare different methods and see how well they fare and how plausible the results are when data are not generated in a computer.

References:

Abadie, A., Diamond, A., and Hainmueller, J., (2010): *Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program*, Journal of the American Statistical Association 105, pp. 493–505.

Athey, S., and Imbens, G. W., (2017): *The state of applied econometrics: Causality and policy evaluation*, Journal of Economic Perspectives 31, pp. 3–32.

Bai, J., (2003): *Inferential Theory for Factor Models of Large Dimensions*, Econometrica 71, pp. 135-171.

Bai, J., and Ng, S. (2002): *Determining the Number of Factors in Approximate Factor Models*, Econometrica 70, pp. 191-221.

Becker, M., & Klößner, S., (2017): *Estimating the economic costs of organized crime by synthetic control methods*, Journal of Applied Econometrics 32, pp. 1367–1369.

Becker, M., & Klößner, S., (2018): *Fast and reliable computation of generalized synthetic controls*, Econometrics and Statistics 5, pp. 1–19.

Chernozhukov, V., Wüthrich, K., and Zhu, Y., (2021): *An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls*, arXiv:1712.09089v10 [econ.EM] 20 May 2021.

Ferman, B., and Pinto, C., (2016): *Revisiting the Synthetic Control Estimator*, MPRA Paper No. 73982, https://mpra.ub.uni-muenchen.de/73982/.

Kaul, A., Klößner, S., Pfeifer, G., and Schieler, M., (2015): *Synthetic Control Methods: Never Use All Pre-Intervention Outcomes Together With Covariates*, MPRA Paper 83790.

Kleven, H. J., Landais, C., & Saez, E., (2013): *Taxation and international migration of superstars: Evidence from the European football market*, American Economic Review 103, pp. 1892–1924.

Klößner, S., Kaul, A., Pfeifer, G., & Schieler, M., (2018): *Comparative politics and the synthetic control method revisited: A note on Abadie et al. (2015)*, Swiss Journal of Economics and Statistics 154, p. 11.

Kuosmanen, T., Zhou, X., Eskelinen, J., and Malo, P., (2021): *Design Flaw of the Synthetic Control Method*, MPRA Paper No. 106390, https://mpra.ub.uni-muenchen.de/106390/.

Malo, P., Eskelinen, J., Zhou, X., and Kuosmanen, T., (2020): *Computing Synthetic Controls Using Bilevel Optimization*, MPRA Paper No. 104085, https://mpra.ub.uni-muenchen.de/104085/